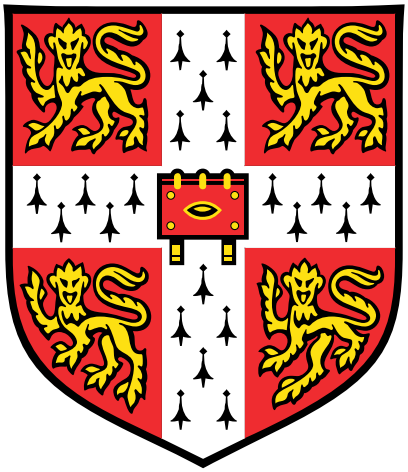
A Systematic Framework for Feature Analysis, Selection, and Property Prediction of Chemical Materials Using Machine Learning



Son Gyo Jung

Department of Physics
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Fitzwilliam College

February 2025

I dedicate this thesis to my loving mother,
whose support and encouragement have been my greatest strength ...

Acknowledgements

I would like to express my profound gratitude to my supervisor, Prof. Jacqueline Cole, for her invaluable guidance throughout my PhD journey. Prof. Jacqueline Cole not only inspired me to embark on this PhD but also provided wise counsel at every stage. Her advice and supervision have been instrumental in the realisation of this thesis.

I owe a profound debt of gratitude to my late grandparents, Sam Chae Jung and Chun Ja Park, and my devoted parents, Kwang Jin Jung and Chae Ah Lee, whose immense sacrifices made it possible for me to pursue my education in the United Kingdom. Their unconditional love, selflessness, and dedication have given me the strength and courage to complete this journey. They have been my steadfast foundation and an unwavering pillar of support and motivation.

I would like to express my thankfulness to all of my colleagues in the Molecular Engineering Group at the Cavendish Laboratory. I am particularly indebted to Guwon Jung with whom I have had the pleasure and honour of collaborating extensively during our tenure in Cambridge. I must also express my deep appreciation to my family and friends, with special thanks to Seung Yeun Yoo and Yedam Jung, for believing in my abilities and providing continuous encouragement throughout this endeavour. It is difficult to fathom navigating this journey without their support. This brief acknowledgment cannot fully express the depth of gratitude I feel for all those who have influenced my journey.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Son Gyo Jung
February 2025

Abstract

This thesis delineates the development and application of a gradient-boosted statistical feature selection workflow within the field of materials informatics. Its utility is demonstrated via the prediction of various material properties across different research domains, including solid-state physics, condensed matter physics, materials science and engineering, as well as spectroscopic characterisation. The overarching objective is to establish a holistic, universally applicable materials discovery workflow that minimises computational overhead, while mitigating potential human biases throughout the model development stage. Such a general-purpose workflow is adept at accommodating both organic and inorganic materials, thereby establishing it as a versatile and adaptable tool for diverse material inputs.

Chapter 1 reviews the progress in the materials discovery process and the pivotal role that machine learning has played in these advancements. It examines the shortcomings that are associated with current prediction methods in materials informatics. This discussion sets the stage for the development of a systematic approach to feature analysis, selection, and modelling of chemical materials aimed at enhancing materials discovery.

Chapter 2 delves into the detailed components of the proposed workflow, demonstrating its effectiveness in predicting material properties. The underlying theoretical concepts and essential background information that support the development and application of the workflow are presented therein.

In **Chapter 3** and **Chapter 4**, the application of the proposed workflow to predict the properties of inorganic materials is presented. A comprehensive study is conducted on predicting the Curie temperature based on the chemical composition of ferromagnetic materials. Additionally, the workflow demonstrates its capability to autonomously construct magnetic phase diagrams by extrapolating relationships learned from a curated materials database. The application of the workflow is further extended across diverse research domains, showcasing its ability to accurately predict material properties when benchmarked against high-throughput calculations and experimental measurements. This is explored in subsequent subsections, encompassing the prediction

of band gaps, predictive modelling of high-entropy alloys and amorphous metallic alloys, as well as the estimation of critical temperatures for superconductors.

In **Chapter 5**, the focus is shifted to the prediction of material properties in organic materials. Building on the details of the workflow presented previously, this chapter highlights its predictive capabilities across a broad range of organic properties, including drug-related attributes and quantum mechanical characteristics. A key demonstration is the integration of the workflow with other machine learning algorithms, such as unsupervised learning methods that generate substructure vector embeddings. Additionally, a study is presented where convolutional neural networks are used in conjunction to predict optical absorption peaks in organic materials, showcasing the workflow's compatibility with advanced machine learning techniques.

Lastly, **Chapter 6** concludes this thesis and suggests future research topics that could complement or further broaden the scope of predictive modelling within the field of materials informatics.

Publications

The research included in the chapters of this thesis has been previously published as follows:

Chapters 1 and 2

Jung, S. G., Jung, G., Cole, J. M., Gradient boosted and statistical feature selection workflow for materials property predictions. *J. Chem. Phys.*, 2023, 159 (19), 194106.

Chapters 3 and 4

Jung, S. G., Jung, G., Cole, J. M., Machine-Learning Prediction of Curie Temperature from Chemical Compositions of Ferromagnetic Materials. *J. Chem. Inf. Model.*, 2024, 64 (16), 6388–6409.

Jung, S. G., Jung, G., Cole, J. M., Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow. *J. Chem. Inf. Model.*, 2024, 64 (4), 1187-1200.

Jung, S. G., Jung, G., Cole, J. M., Predictive Modeling of High-Entropy Alloys and Amorphous Metallic Alloys Using Machine Learning. *J. Chem. Inf. Model.*, 2024, 64 (19), 7313–7336.

Jung, S. G., Jung, G., Cole, J. M., Machine-Learning Predictions of Critical Temperatures from Chemical Compositions of Superconductors. *J. Chem. Inf. Model.*, 2024, 64 (19), 7349–7375.

Chapter 5

Jung, S. G., Jung, G., Cole, J. M., Automatic Prediction of Molecular Properties Using Substructure Vector Embeddings within a Feature Selection Workflow. *J. Chem. Inf. Model.*, 2025, 65 (1), 133-152.

Jung, S. G., Jung, G., Cole, J. M., Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks. *J. Chem. Inf. Model.*, 2024, 64 (5), 1486-1501.

Other Publications

The following publications are not incorporated into this thesis:

Jung, G., **Jung, S. G.**, Cole, J. M. Automatic materials characterization from infrared spectra using convolutional neural networks *Chem. Sci.*, 2023, 14, 3600-3609.

Gartside, J. C., **Jung, S. G.**[1], Yoo, S. Y. et al. Current-controlled nanomagnetic writing for reconfigurable magnonic crystals. *Commun. Phys.*, 2020, 3, 219.

[1] Co-first author

Table of contents

List of figures **xix**

List of tables **xxiii**

1 Introduction **1**
1.1 Advancements in Materials Discovery Processes 1
1.2 Machine Learning in Molecular and Materials Science 3
1.3 Fundamental Concepts in Machine Learning 6
1.3.1 Types of Machine Learning 7
1.3.2 Gradient-Boosted Decision Trees 9
1.3.3 Bias-Variance Tradeoff 10
1.3.4 Hyperparameter Tuning 11
1.3.5 Regularisation 15
1.4 Materials Informatics 17
1.4.1 Molecular and Chemical Descriptors 17
1.4.2 Simplified Molecular Input Line-Entry System 19
1.5 Scope of this Thesis 20

2 Gradient Boosted and Statistical Feature Selection Workflow for Materials Property Predictions **23**
2.1 Introduction 24
2.2 System Overview 25
2.2.1 (A) Data Acquisition & featurisation 27
2.2.2 (B) Feature Selection, Analyses and Engineering 28
2.2.3 (C) Model Optimisation 34
2.2.4 (D) Materials Property Prediction 36

3 Property Prediction of Inorganic Materials: Curie Temperature and Band Gap 37
3.1 Prediction of Curie Temperature of Ferromagnetic Materials 38
3.1.1 Introduction 38
3.1.2 Methods 43
3.1.3 Regression Analysis of Curie Temperatures 47
3.1.4 Blind Test 66
3.1.5 Predictive Capabilities for ML-Model Applications: Two Case Studies 70
3.2 Prediction of Band Gaps of Inorganic Materials 76
3.2.1 Classification of Materials by Metallicity 76
3.2.2 Regression Analysis of Band Gap 79

4 Property Prediction of Inorganic Materials: Alloys and Superconductors 93
4.1 Predictive Modelling of High-Entropy Alloys and Amorphous Metallic Alloys 94
4.1.1 High-Entropy Alloys 94
4.1.2 Amorphous Metallic Alloys 108
4.2 Predictions of Critical Temperatures of Superconductors 120
4.2.1 Classification of Superconductors 120
4.2.2 Regression Analysis of Critical Temperatures 127
4.2.3 Predictive Capabilities for ML-Model Applications 139

5 Property Prediction of Organic Materials 149
5.1 Automatic Prediction of Molecular Properties Using Substructure Vector Embeddings within a Feature Selection Workflow 150
5.1.1 Introduction 150
5.1.2 Methods 153
5.1.3 Results and Discussion 156
5.2 Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks 165
5.2.1 Introduction 165
5.2.2 Methods 166
5.2.3 Results and Discussion 175

6 Conclusions and Future Work 189
6.1 Conclusions 189
6.2 Future Outlook 192
6.2.1 Hybrid Modelling Approaches in Materials Science: Integrating Physics-Based and Data-Driven Models 193
6.2.2 Expanding the Scope of Machine Learning in Materials Science through Semi-Supervised Learning Frameworks 194

References 197

Appendix A Supporting Information: Chapter 3 219
A.1 Class-Probability Distributions 219
A.2 Predictions Against DFT Calculations 220

Appendix B Supporting Information: Chapter 4 221
B.1 DFT-Aware Auxiliary Model Performance 221
B.2 Experimental-Aware ML Model Performance 223
B.3 Error Distribution by Type of Solvents 224

List of figures

2.1 Overview of GBFS Workflow Compartmentalised into Four Distinct Stages 26

3.1 Curie Temperature - Periodic Table Highlighting the Compositional Space 42
3.2 Curie Temperature - Alterative View of the GBFS Workflow 46
3.3 Curie Temperature - Test Set Results of the ML-Based Predictions of T_c Against the Ground Truth 48
3.4 Curie Temperature - Distribution of T_c Values 50
3.5 Curie Temperature - Evaluation of MAE, RMSE, and R^2 51
3.6 Curie Temperature - Gradient Boosting Feature Selection Results . . . 58
3.7 Curie Temperature - Multicollinearity Reduction, Permutation Feature-Importance, and Recursive Feature Elimination Result 62
3.8 Curie Temperature - Bayesian Optimisation Result 63
3.9 Curie Temperature - Feature Relevance Plot 64
3.10 Curie Temperature - SHAP Framework Results 65
3.11 Curie Temperature - ML-based Predictions of T_c For Unseen Rare-Earth-Based Chemical Compound 74
3.12 Curie Temperature - ML-Generated Magnetic Phase Diagrams 75
3.13 Band Gap - Performance for the Classification of Materials by Metallicity 77
3.14 Band Gap - Regression of the ML-Based Predictions of E_g Against Experimental Measurements 81
3.15 Band Gap - Distribution of Band Gap Values of Nonmetals in the Training Set 82
3.16 Band Gap - Multifidelity Model Performance Against Experimental Measurements and DFT-Based Computed Results 87
3.17 Band Gap - Regression Analysis of E_g 89

4.1 High-Entropy Alloys - ML Predictions of Bulk Modulus 97

4.2 High-Entropy Alloys - Feature Relevance and Permutation-Based Feature Importance Plots for the Regression Analysis of K Values in Equimolar Quaternary HEAs . . . 99
4.3 High-Entropy Alloys - SHAP Framework Result . . . 99
4.4 High-Entropy Alloys - Property Predictions . . . 106
4.5 Amorphous Metallic Alloys - Confusion Matrix for the Classification of Alloys by Their Glass-Forming Ability . . . 112
4.6 Amorphous Metallic Alloys - Feature Relevance and Permutation Importance Plots . . . 113
4.7 Amorphous Metallic Alloys - Property Predictions . . . 119
4.8 Critical Temperature - Classification Performance on the Test Set . . . 123
4.9 Critical Temperature - Distribution of Critical Temperature . . . 131
4.10 Critical Temperature - ML-Based Predictions . . . 132
4.11 Critical Temperature - ML-Based Prediction For Distinct Groups of Superconducting Materials . . . 133
4.12 Critical Temperature - Distribution of Errors . . . 133
4.13 Critical Temperature - Electronic Density of States . . . 147
4.14 Critical Temperature - ML-Based T_c Predictions of $(Ca_{1-x}La_x)FeAs_2$ On the La Content x . . . 148
4.15 Critical Temperature - ML-Based Virtual Compositional Analysis of Hg-Pb-Ba-Ca-Cu-O Series of Compounds . . . 148

5.1 Organic Molecules - Workflow of GBFS-Mol2Vec . . . 155
5.2 UV-Vis - Overview of Our Operational Workflow Compartmentalised Into Two Distinct Subworkflows . . . 167
5.3 UV-Vis - Diagram Depicting the Process of Feature Matrix Generation From a Smiles String of a Chemical Molecule . . . 171
5.4 UV-Vis - Schematic of the Deep Residual Convolutional Neural Network Architecture . . . 174
5.5 UV-Vis - Regression Analysis of the ML-Based Prediction of Vertical Excitation Energies Against DFT-Based Calculations . . . 176
5.6 UV-Vis - Regression Analysis of the Multifidelity ML Model Prediction of the Optical Absorption Peak Wavelengths Against Experimental Measurements . . . 178
5.7 UV-Vis - Distributions of the Absolute Errors of the ML-Based Prediction 178
5.8 UV-Vis - Reduction in Total Loss Achieved by the Top 10 Features . . . 183

5.9 UV-Vis - Examples of Input Molecules and the Corresponding Prediction of $\lambda_{\max}$ Values Against Their Experimental Measurements 186

A.1 Band Gap Classification - Probability Distribution of the Final ML Classifier 219
A.2 Band Gap - Bayesian-Optimised Model Performance Against DFT Calculations 220

B.1 UV-Vis - Regression Analysis of the ML-Based Prediction of the Vertical Excitation Energies Against the DFT-Based Calculations 222
B.2 UV-Vis - Regression Analysis of the ML-Based Prediction of the Experimental Optical Absorption Peak of Unseen Chemical Dyes Against Their Experimental Measurements 223
B.3 UV-Vis - Distributions of the Absolute Errors of the ML-Based Prediction of λ_{max} Against the Experimental Measurements, Partitioned by Solvent Types 224

List of tables

3.1 Curie Temperatures - A List of Most Relevant Features 52
3.2 Curie Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of T_c Values (i.e., Model Output), Along With the Percentage Difference From the Mean Experimental T_c Values (Ground Truth)[a]. 67
3.3 Band Gap - Summary of the Performance Metrics for the Final Classifier of Materials by Metallicity 76
3.4 Band Gap - Summary of the Band Gap Predictions Against Experimental Measurements 82
3.5 Band Gap - Examples of Input Chemical Composition and the Corresponding Prediction of Band Gap (E_g), Sorted by the Absolute Percentage Difference From the Mean Experimental Values in Ascending Order. 90

4.1 High-Entropy Alloys - MAE for the prediction of bulk modulus for quaternary HEAs 96
4.2 High-Entropy Alloys - Performance Metrics for the Classification of HEAs by Their Structural Symmetries 105
4.3 High-Entropy Alloys - Prediction of Wigner-Seitz Radius and Bulk Modulus for the $NiCoFeCrAl_x$ quinary system 107
4.4 Amorphous Metallic Alloys - Performance Metrics for the Classification of Alloys by Their Glass-Forming Ability 113
4.5 Amorphous Metallic Alloys - A List of Most Relevant Features 113
4.6 Critical Temperature - Classification of Superconductors by Their T_c Values Relative to Threshold $T_{\text{thres}} = 10$ K 122
4.7 Critical Temperature - Performance of Different ML Methods 132
4.8 Critical Temperature - A List of Most Relevant Features 139

4.9 Critical Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of Critical Temperature (i.e., Model Output), Along With the Percentage Difference From the Experimental T_c Values (i.e., Ground Truth)[a]. 145

5.1 Organic Molecules - Summary of Data Sets 155
5.2 Organic Molecules - QM9 Data Set 155
5.3 Organic Molecules - Performance of Various ML Models on the QM9 Test Set 162
5.4 Organic Molecules - Model Performance on Additional Regression Tasks 162
5.5 Organic Molecules - Model Performance on Photovoltaic Efficiency of Organic Molecules 162
5.6 Organic Molecules - Model Performance on Classification Tasks 164
5.7 UV-Vis - A List of Data Sources and the Corresponding Number of Experimental Measurement of Peak Optical Absorption Wavelengths . 168
5.8 UV-Vis - A List of Features Used to Generate the Feature Matrix for DR-CNN 171
5.9 UV-Vis - An Overview of Model Performance for Predicting Peak Optical Absorption Wavelength 179
5.10 UV-Vis - A List of Features Identified to Have the Most Relevance in the Prediction of λ_{max} Values Against Their Experimental Measurements 179

Chapter 1

Introduction

1.1 Advancements in Materials Discovery Processes

The discovery of novel materials has driven advancements in modern technology, ranging from semiconductor technology and data storage to developing sophisticated microchips, enabling computational capabilities that far surpass those envisioned a decade earlier. Designing new materials involves carefully evaluating tradeoffs between the costs associated with research, development, and manufacturing, and the performance enhancements they provide. Decisions on material selection are made to optimise these factors. Nevertheless, continual advancements in technology and evolving consumer demands perpetually drive the need for materials that exhibit innovative properties, pushing the boundaries of science.

In materials science, traditional methods for discovering new materials often rely on a trial-and-error approach, which poses a significant bottleneck in the design-to-device workflow. A trial-and-error approach is notably resource-intensive, necessitating considerable time investment and a sustained level of specialist human capital to provide the domain expertise to translate theoretical concepts into practical experiments, and vice versa [1, 2]. Experiments are often constrained by the necessity for adequate facilities and the ability to replicate experimental conditions, which can hinder progress in materials science and engineering. Consequently, materials documented in the scientific literature seldom progress to functional applications or practical devices. Additionally, the potential for selection bias in peer-reviewed publications may foster an overly optimistic outlook, as failures are often overlooked or go unreported. This practice may result in the repetitive execution of unsuccessful experiments. Such dynamics underline the pronounced inefficiency within the traditional materials discovery process,

and clearly highlight that empirical methods are not cost-effective for exploring the extensive chemical and material landscape.

Computational methods have been instrumental in accelerating the discovery of novel materials across various research domains, providing profound insights that enhance scientific understanding and technological innovation. Over recent decades, high-throughput computational calculations, primarily based on Density Functional Theory (DFT) [3], have become a preferred method, either as an alternative or a complement to experimental research effects often conducted on a trial-and-error basis. These *ab initio* methods have, to some extent, optimised the process of material characterisation by enabling the computational simulation of materials and their properties. As a result, they facilitate a rapid exploration of vast chemical and property spaces at a rate unattainable with traditional design-to-device workflows. While high-throughput calculations remain computationally intensive, limiting their application to predicting ground state properties, recent advancements in high-performance computing have broadened the usability of DFT calculations. This broadening of scope solidifies such high-throughput methods as one of the key achievements in computational materials science [4].

The integration of computational techniques indeed heralds a transformative approach to materials discovery. These methods can estimate or predict target properties of a material from first principles, facilitating the shortlisting of the most promising candidates for experimental validation. This strategy can reduce the overhead costs associated with the materials discovery process. Nevertheless, efficiently navigating the immense chemical landscape using high-throughput methods remains impractical due to the required computational resources. This limitation has prompted the investigation of alternative approaches that can leverage the abundant data available in both scientific literature and data repositories without incurring substantial computational costs. When enhanced with features based on stoichiometry and/or chemical structural information, such predictive modelling strategies pioneer a new research area. This area is poised to accelerate the discovery of materials through the targeted design of novel materials tailored to specific desired properties.

It is also important to acknowledge that these efforts within computational materials science have culminated in the creation of comprehensive data repositories, such as the Materials Project (MP) [5, 6] and the Open Quantum Materials Database (OQMD) [7]. These repositories contain extensive data sets of chemical structures and material properties. The availability of such aggregated *ab initio* chemical information, combined with the rise of big-data initiatives, has spurred a marked interest in data-driven

methodologies. These approaches are valued for their ability to process and analyse high-dimensional data sets, enabling the discovery of previously unseen patterns and relationships. Specifically, this advancement has propelled the application of machine learning and data mining, thereby broadening our understanding of the underlying chemistry and physics of materials. For instance, these data-driven approaches have proven to be highly accurate in predicting various properties, including band gaps [8–10] and formation energies [11, 12]. This underscores the efficacy of materials screening processes for discovering new materials within highly complex feature spaces across diverse applications [13–17]. Inevitably, these developments have prompted the emergence of a new subfield within materials science, known as materials informatics [18–20].

Despite the surge in data-driven materials informatics, the field remains far from fully leveraging the available data, particularly the unstructured information within the scientific literature. Experimental measurements of material properties, which are the ground truth of prediction as opposed to estimations derived from *ab initio* methods, have historically required manual curation. More recently, computational approaches have been developed to automate the extraction of such information from the extensive corpus of scientific literature. A notable example is ChemDataExtractor, a text-mining application widely utilised in the chemistry domain. This toolkit employs natural language processing techniques alongside advanced machine learning methods to expedite the creation of structured databases from unstructured or disparate sources of information [21, 22]. These efforts have paved the way for predicting material properties against experimental measurements across various research domains.

1.2 Machine Learning in Molecular and Materials Science

Data-driven approaches in materials science employ materials informatics [1, 2] and machine learning [23–26] to enhance predictive accuracy. In a typical supervised learning workflow, chemical information is transformed into a machine-readable format using feature descriptors [27–31]. These descriptors, which capture essential chemical attributes, are used to train predictive models. The utility of these models spans two main applications: (i) predicting the properties of novel chemical materials in regression tasks, and (ii) classifying materials into different categories or classes in classification scenarios.

Furthermore, recent developments in deep learning have emphasised the utility of convolutional neural networks (CNNs) and graph neural networks (GNNs) for automatic feature extraction, with GNNs particularly adept at incorporating the connectivity information of molecular structures directly into their network architecture [32–35]. These techniques provide a universal framework for representing chemical materials effectively. Nevertheless, deploying these techniques hinges on the supervised training of neural networks for molecular property prediction, a process frequently hindered by the limited availability of labelled data. Consequently, there has been a surge in the popularity of unsupervised or self-supervised learning frameworks that leverage a large corpus of unlabelled data sets [36–38].

In the realm of unsupervised or self-supervised learning, transformer-based language models, pretrained on extensive unlabelled data sets, have demonstrated significant progress in various downstream natural language processing tasks. Notably, Ross et al. [37] developed molecular embeddings by training a transformer encoder model with a linear attention mechanism. This model, named Molecular Language transFormer (MoLFormer), was trained using SMILES (Simplified Molecular Input Line-Entry System [39]) sequences from a corpus of 1.1 billion unlabelled molecules. This unsupervised learning approach demonstrates the efficacy of large-scale molecular language models in capturing essential chemical and structural information (i.e., molecular representation learning). These models enable the prediction of diverse molecular properties after being fine-tuned on data sets tailored to specific tasks, showing their competitive performance against other supervised and unsupervised language models and GNNs.

Despite the innovation of large-scale molecular language models and other modelling methodologies that rely on extensive corpora of unlabelled data within unsupervised or self-supervised frameworks, they encounter other challenges, chiefly the need for substantial computational resources. For instance, the aforementioned MoLFormer model [37] is considered relatively efficient and scalable as it employs adaptive batch bucketing, parallelisation, and a linear attention mechanism. Nevertheless, the training of such models presents considerable computational demands. The referenced study used 16 graphics processing units (GPUs) and required 208 hours to complete merely four epochs of pretraining, with a batch size of 1600 molecules per GPU. Without the application of bucketing and linear attention, batch sizes would be confined to fewer than 50 molecules per GPU, necessitating the use of over 1000 GPUs to maintain comparable processing throughput.

Given the constraints of limited computing resources, supervised learning often emerges as the preferable solution for many researchers, provided there is an adequate

amount of labelled data. Nonetheless, data sets characterised by high dimensionality and complex features, and the processes involved in model optimisation and hyperparameter tuning can pose similar challenges. Model training and optimisation typically rely on regularisation techniques to achieve model generalisation, which can substantially escalate computational requirements. The necessity for aggregated computing power becomes increasingly apparent with the growth of data volume (as seen in unsupervised learning), the naive addition of exploratory features (which are often irrelevant to the target property), or the increasing complexity of predictive algorithms or models (such as when employing CNNs). These factors can subsequently limit model scalability and generalisation.

Another significant hurdle that warrants discussion is the issue of model interpretability, as machine learning algorithms are often treated as black boxes. Despite ongoing research efforts aimed to enhance the interpretability of models, this aspect is frequently overlooked in practical modelling scenarios. Many peer-reviewed publications fail to emphasise how the algorithm derives outcomes from input features adequately. This black-box approach often leads to a neglect of understanding the relevance and interactions of features within a model, while a disproportionate focus is placed on performance metrics, which can result in overfitting. Addressing these limitations within a materials-discovery pipeline or workflow will shorten the timeline required to broaden the materials landscape and accelerate the search for novel materials with specific properties.

In this thesis, we present a machine learning-based feature-selection workflow designed to mitigate the aforementioned shortcomings associated with increasing data dimensionality and the black-box treatment of machine learning models. Specifically, we employ a distributed gradient-boosting framework [40–42], integrated with exploratory data and statistical analyses [43–48], along with treatments for multicollinearity, to establish a systematic approach to feature analysis and selection. This approach enables us to identify a subset of features within a highly complex feature space, ensuring that the selected features exhibit minimal redundancy and maximum relevance to the target variable or classes. The efficacy and efficiency of our workflow are detailed in subsequent chapters of this thesis, showcasing materials-property predictions that are either superior or comparable to the benchmarks set by state-of-the-art reports in the scientific literature. We demonstrate that our methodology eliminates the need for standard regularisation techniques such as L1 [49], L2 [49, 50], and dropout [51] during model optimisation, thereby achieving model generalisation at a significantly reduced computational cost. This is particularly beneficial for large-scale, data-driven

materials discovery. Moreover, our proposed workflow provides insights into feature relevance and interaction throughout its operational stages, while also eliminating human selection bias during model development. Our research underscores the value of a systematic approach to feature analysis, selection, and predictive modelling of chemical materials using machine learning.

The following sections of this chapter delineate the core principles of machine learning and statistics. This foundational discussion paves the way for a comprehensive review of key concepts in materials informatics. Additionally, each chapter of this thesis is supplemented by studies related to either inorganic or organic material properties discussed therein. For the scope of this thesis, please refer to Section 1.5.

1.3 Fundamental Concepts in Machine Learning

Machine learning, a subfield of artificial intelligence, involves the development of algorithms and statistical models that allow computers to execute specific tasks by identifying patterns and making inferences from provided input data, rather than following explicitly programmed instructions. Machine learning revolutionises data analysis and interpretation by enabling autonomous systems to derive insights and make predictions or decisions from data. This field spans several types of learning, including: (i) supervised learning, which involves training models on labelled data; (ii) unsupervised learning, aimed at detecting underlying patterns in data without preassigned labels; (iii) semi-supervised learning, a method where labelled data is combined with unlabelled data during the training process; (iv) transfer learning, a technique that leverages features or embeddings learned from models previously trained on extensive data sets to enhance the performance of new models on other tasks often with limited data; and (v) reinforcement learning, a strategy where agents enhance their decision-making skills by learning from the outcomes of their actions, whether through rewards or penalties. These diverse techniques render machine learning indispensable across numerous domains, including image recognition, natural language processing, and sophisticated decision-making frameworks and autonomous systems in sectors such as finance and healthcare.

1.3.1 Types of Machine Learning

Supervised Learning

In supervised learning, algorithms are trained on labelled data, which includes pairs of input features and corresponding output values. These algorithms are designed to learn potential mappings from inputs to correct outputs by recognising patterns in the data. The primary tasks in supervised learning are classification and regression. In classification, the model is trained to identify the most probable category or label for given inputs, while in regression, the model predicts a continuous output based on the input features. In materials science, classification tasks might involve predicting the crystal symmetry of materials or identifying materials that are likely to exhibit specific attributes, such as a superconductivity transition temperature above that of liquid nitrogen. These tasks are typically formulated as either multi-class or binary classification problems, depending on the nature of the scientific question at hand. Conversely, regression tasks in materials science may involve predicting a material's band gap energy or constructing phase diagrams that illustrate variations in Curie temperature as a function of compositional changes. These tasks are typically achieved by either interpolating or extrapolating the learned relationships and patterns within the data to facilitate either in-distribution or out-of-distribution predictions, respectively.

Unsupervised Learning

In unsupervised learning, algorithms are developed to discern patterns and structures within unlabelled data. Common tasks in unsupervised learning include clustering and dimensionality reduction. An illustrative case is that of language models employing attention mechanisms, which are trained on vast corpora of unlabelled data. These models generate embeddings that capture features of the input data, which can then be fine-tuned on specific data sets for predicting properties of materials, as previously mentioned. Additionally, k-means clustering and Principal Component Analysis (PCA) are widely recognised unsupervised algorithms, favoured for their relative simplicity in implementation and low computational demands. Clustering methods are commonly used to explore potential underlying relationships within the feature space, grouping similar data points closely together. Similarly, PCA establishes orthogonal axes that enhance variance between groups and separate data by maximising inter-class variability. These techniques are useful for analysing patterns between certain features in relation to the target variable or class.

Semi-Supervised Learning

Semi-supervised learning combines labelled data and a larger collection of unlabelled data. This method acts as an intermediary between supervised learning, dependent entirely on labelled data sets, and unsupervised learning, which uses no labelled data. Such an approach can enhance performance over purely unsupervised methods without requiring extensive labelled data. It can facilitates better generalisation on unseen data, proving highly efficient for practical scenarios where labelled data is limited, yet unlabelled data is plentiful. The premise is that when unlabelled data is used in conjunction with labelled data, it offers a more comprehensive insight into the properties of a data set.

Semi-supervised learning employs various techniques to enhance model training. For instance, self-training begins by training on a small labelled data set, using the model to predict labels for unlabelled data, and incorporating high-confidence predictions back into the training set as accurate labels. In contrast, co-training involves training two models simultaneously on different subsets of features, where predictions of a model help refine the other. Further methodologies are explored within the realms of graph-based methods and consistency regularisation. Graph-based methods use a graph structure where nodes are samples and edges indicate similarity, allowing label information to spread from labelled to unlabelled nodes. Consistency regularisation ensures that a model produces the same output for an unlabelled sample even after undergoing a slight alteration, promoting the learning of robust features. These methods collectively aim to leverage the strengths of both labelled and unlabelled data to improve learning.

The utility of semi-supervised learning is particularly evident in situations where procuring comprehensively labelled data sets is either cost-prohibitive or time-consuming. This feature makes it well-suited for applications such as natural language processing and materials property prediction, where labelled data may be scarce.

Transfer Learning

Transfer learning is a technique in machine learning that involves reusing a model developed for one specific task as the foundation for another. This approach is particularly prevalent in deep learning, where pre-trained models are used as starting points for a particular task such as in computer vision and natural language processing, capitalising on their ability to save significant computational resources and time required to develop neural network models from scratch. The primary advantage of transfer learning is its efficiency in leveraging pre-existing models trained on extensive data sets,

which have already learned to identify general features. By leveraging these pre-learned features, this capability facilitates the rapid development and enhanced performance of new models, especially in scenarios where data availability is limited.

In practice, transfer learning is implemented through two primary techniques: (i) feature extraction and (ii) fine-tuning. Feature extraction involves using the learned representations from an existing network to derive meaningful features from new samples, wherein the base layers of the model are frozen, and only a new layer is trained for the new task. On the other hand, fine-tuning entails further training of the pre-trained model, typically after an initial feature extraction phase, to better adapt the model to the new task. This approach is especially beneficial when limited data is available for a new task or when the new task closely aligns with the original training task. Transfer learning drastically reduces the computational costs and the time required for model development compared to constructing a new model from scratch with random initialisation.

1.3.2 Gradient-Boosted Decision Trees

The field of machine learning encompasses a diverse array of algorithms, from simpler models like linear and logistic regression to more complex approaches such as gradient-boosted decision trees (GBDTs) and various deep learning techniques, as documented in scientific literature. Our discussion herein will primarily concentrate on GBDTs, which are extensively used as final predictive models for both classification and regression tasks in the later chapters of this work.

Framework of Gradient-Boosted Decision Trees

GBDTs are predictive models based on an ensemble of independent weak learners, or trees. Consider a data set $D = \{(\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_n = (x_{n,m=1}, x_{n,2}, ..., x_{n,M})$ is the feature vector of sample n, among N total samples ($n \in N$), with M total features ($m \in M$) and target variable y_n. GBDTs compute the predicted value $\hat{y}_n$ of y_n via the summation of K independent trees, denoted as f_k, with structure s and leaf weights w:

$$\hat{y}_n = \sum_{k=1}^{K} f_k(\boldsymbol{x}_n), \tag{1.1}$$

where $f_k \in F$ and $F = \{f(\boldsymbol{x}) = w_{s(\boldsymbol{x})}\}(s : \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T)$ is the classification and regression tree space with T number of leaves in the tree. In other words, T and w are codomains, or sets, that constrain the output of the tree structure. The final prediction

of the target variable for a sample is then computed by the summation of the weight of the leaves into which the sample falls according to the decision rules of the constructed trees.

To learn the set of decision rules, or functions, Chen et al. [41] proposed a regularised learning objective:

$$\mathcal{L} = \underbrace{\sum_{n=1}^{N} l(y_n, \hat{y}_n)}_{\text{Convex loss function}} + \underbrace{\sum_{k=1}^{K} \Omega(f_k)}_{\text{regularisation term}} , \tag{1.2}$$

where the regularisation term is given by:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2, \tag{1.3}$$

with γ and λ coefficients corresponding to T and w, respectively. Equation 1.3 constrains the model from learning complex predictive functions and penalises the model from overfitting to the training data. Using Equation 1.2, the learning objective at the t-th iteration can be rewritten as:

$$\mathcal{L}^{(t)} = \sum_{n=1}^{N} l(y_n, \hat{y}_n^{(t-1)} + f_t(\boldsymbol{x}_n)) + \Omega(f_t), \tag{1.4}$$

where $\hat{y}_n^{(t-1)}$ is the prediction of the n-th sample at the $(t-1)$th iteration and f_t is the tree at the t-th iteration. Equation 1.4 represents an additive type of model training, where the training proceeds iteratively through the addition of new trees that best predict the errors, or residuals, of prior trees. The final prediction $\hat{y}_n^{(t)}$ is subsequently made by combining the prediction of the previous tree $\hat{y}_n^{(t-1)}$ with its predicted error from tree f_t, which minimises the learning objective.

1.3.3 Bias-Variance Tradeoff

The bias-variance tradeoff is an important concept in machine learning, and it describes the compromise between errors introduced by the modelling assumptions (i.e., bias) and errors arising from the model sensitivity or response to small variations in the training data (i.e., variance). Bias arises due to oversimplification; that is when simplifying assumptions made by a model prevent it from capturing the underlying relationship between inputs and outputs, often leading to underfitting. Models characterised by high bias, such as linear models, tend to be simplistic and unable to adequately capture complex patterns and data structures. Variance, in contrast, refers to errors that occur when a model is overly complex relative to the data set. Such complexity leads the

model to capture not only the underlying data patterns but also the random noise within the training set. Such elevated sensitivity often results in overfitting, particularly in sophisticated models such as deep neural networks handling high-dimensional data sets, which are inherently susceptible to high variance. This susceptibility can degrade model performance on new, unseen (i.e., out-of-sample) data. Consequently, careful selection of models and their architecture, along with an effective discernment in feature inclusion, are critical in identifying an optimal balance between bias and variance, thereby minimising overall error and enhancing the effectiveness of machine learning models. Regularisation techniques are commonly used with complex models to foster model generalisation. For further details on regularisation, please refer to Section 1.3.5.

1.3.4 Hyperparameter Tuning

Hyperparameter tuning is a pivotal process in machine learning aimed at optimising the parameters that govern the architecture and behaviour of a model, but which are not learned during the training phase. These hyperparameters, set before training commences, can include variables such as the learning rate, the number of hidden layers and neurons in neural networks, and the number of trees and maximum depth in tree-based models. Unlike hyperparameters, model parameters, such as weights in neural networks, are dynamically learned and adjusted throughout the training process. The goal of hyperparameter tuning is to identify the optimal configuration of these predefined parameters that maximises model performance, evaluated through metrics such as accuracy, precision, recall, F1-score, or the area-under-the-curve of the receiver-operating-characteristic curve (AUC-ROC) for classification tasks, and metrics like mean squared error and root mean squared error for regression tasks.

Hyperparameter tuning employs various strategies such as grid search, random search, Bayesian optimisation, and combinations of these methods. Grid search thoroughly evaluates model performance across a defined grid of hyperparameter values, identifying the configuration that delivers optimal results. However, this method becomes significantly more computationally demanding as the number of hyperparameters and the granularity of the grid increase. Conversely, random search adopts a stochastic approach to select combinations of hyperparameters for training, enhancing efficiency over grid search, especially when some hyperparameters have a negligible effect on performance. However, each evaluation step in the random search is independent, with no subsequent steps informed by the outcomes of previous evaluations. Bayesian optimisation conceptualises the performance of a model as a function of its hyperparameters, leveraging historical evaluation data to identify optimal

configurations more effectively. Proper hyperparameter tuning can markedly improve the accuracy of a model and its operational efficiency, which are vital for constructing robust machine-learning models. Yet, despite its significance, hyperparameter tuning frequently receives inadequate attention or is underutilised in computational materials science, where manual selection practices risk introducing human biases. Consequently, adopting a systematic optimisation strategy is essential for accurately determining the architecture of these models, which are devoid of human intervention. In this thesis, Bayesian optimisation has been employed across various tasks. Further details are discussed in the sections that follow.

Bayesian Optimisation

In Bayesian optimisation, a probabilistic surrogate model is built to approximate an objective function f, which is based on a performance metric that is to be maximised (or minimised) under three constraints: (i) the analytical expression of f and its derivatives are unknown (i.e., no closed form), (ii) f is expensive to evaluate, and (iii) the evaluations of f may result in noisy responses. In this work, the surrogate model is a Gaussian Process (GP), the generalisation of a Gaussian distribution to a distribution over functions that is characterised by mean and covariance (or positive definite kernel) functions. It consists of a prior distribution that represents the prior beliefs over all possible f. The optimisation algorithm sequentially refines the surrogate model following a set of observations via Bayesian posterior updating, yielding a posterior mean and variance functions that better approximate f over the space of the objective functions [52–55].

The use of GPs to construct the surrogate model stems from the fact that Gaussian distributions are self-conjugate with respect to a Gaussian likelihood function; that is, a Gaussian prior yields a Gaussian posterior when the likelihood function is also Gaussian, that is:

$$\underbrace{P(f|D_{1:t})}_{\text{GP Posterior}} \propto \overbrace{L(D_t|f)}^{\text{Likelihood}} \underbrace{P(f|D_{1:t-1})}_{\text{GP Prior}}, \tag{1.5}$$

where the observations are of the form $D_{1:T} = \{(\boldsymbol{\theta}_{t=1}, \phi_{t=1}), ..., (\boldsymbol{\theta}_T, \phi_T)\}$, a total of T observations with query points $\boldsymbol{\theta}$ within the hyperparameter space and the output ϕ_T of f. Therefore, with a conjugate prior for a Gaussian likelihood function, the posterior mean and covariance functions can be computed given the set of observations [49, 54–56].

Bayesian optimisation employs an acquisition function, which uses the GP posterior to evaluate the utility of candidate points within the hyperparameter space that may improve to the current best evaluation of $f(\boldsymbol{\theta})$. There exists a tradeoff between exploration and exploitation since an acquisition function wishes to 'explore' regions of high uncertainty in the posterior while 'exploiting' known optimal regions where the posterior mean, and therefore, the objective function is expected to be high. The next query point $\boldsymbol{\theta}_{t+1}$ is realised by maximising the acquisition function with respect to the exploration of a high-variance region and the exploitation of a high-mean region, both of which yield high acquisition values. After evaluating each query point, the surrogate model is then sequentially updated, producing a more informative posterior distribution [53, 57–59]. Three acquisition functions, denoted by α, were employed in this work, each corresponding to one of the following acquisition schemes: (i) Probability of Improvement (PI) [57], (ii) Expected Improvement (EI) [58], and (iii) Upper Confidence Bounds (UCB) [59]. Please see Algorithm 1 for the pseudo-code of Bayesian optimisation implemented herein.

Note that the acquisition strategies described herein assume that the objective function is based on a performance metric to be maximised. However, if the objective function is to be minimised, the next query point is determined by maximising the negative of the acquisition functions or by taking the lower confidence bound.

Algorithm 1: Bayesian optimisation with Gaussian process prior

input: objective function f, hyper-parameter space $\boldsymbol{\theta}$, acquisition functions α, initialisation points T_{init}, maximum number of evaluation T

$y_{best} \leftarrow 0$;

for $t = 1$ **to** T_{init} **do**

  select $\boldsymbol{\theta}_t$ via randomly sampling;

  compute exact objective function $y_t \leftarrow f(\boldsymbol{\theta}_t)$;

  if $y_t > y_{best}$ **then**

    $\boldsymbol{\theta}_{best} \leftarrow \boldsymbol{\theta}_t$;

    $y_{best} \leftarrow y_t$;

  end

end

for $t = T_{init} + 1$ **to** T **do**

  build probabilistic model for f conditioned on previous observations $D_{1:t-1}$;

  compute all possible true functions using Gaussian process regression;

  optimise acquisition functions α independently based on the posterior distribution and propose a candidate point for each acquisition scheme $\boldsymbol{\theta}_{t,s} \leftarrow argmax_{\boldsymbol{\theta}}\ \alpha_s(\boldsymbol{\theta}|D_{1:t-1})$ for $s = \{PI, EI, UCB\}$;

  choose next evaluation point $\boldsymbol{\theta}_t \leftarrow argmax_{\boldsymbol{\theta}}\ \ softmax(\mu(\boldsymbol{\theta}_{t,s}))$;

  compute exact objective function $y_t \leftarrow f(\boldsymbol{\theta}_t)$;

  if $y_t > y_{best}$ **then**

    $\boldsymbol{\theta}_{best} \leftarrow \boldsymbol{\theta}_t$

    $y_{best} \leftarrow y_t$

  end

end

return $\boldsymbol{\theta}_{best}$

1.3.5 Regularisation

Numerous regularisation techniques are available to enhance model generalisation, including L1 (i.e., LASSO regression) [49], L2 (i.e., Ridge regression) [49, 50], and dropout [51]. In this thesis, the discussion will predominantly focus on L1 and L2 regularisation, given their relevance and application in tree-based models, which are central to the subsequent chapters of this work.

L1 and L2 regularisation penalise a complex model through the addition of a regularisation term in the loss function. The L1 regularisation term is equal to the sum of the absolute value of the weights, or coefficients, of features $m \in M$

$$\hat{\mathcal{L}}_{(L1)} = \underbrace{\mathcal{L}(y, \hat{y})}_{\text{Loss function}} + \overbrace{\lambda \sum_{i=1}^{M} ||w_i||_1}^{\text{regularisation term}}, \tag{1.6}$$

where $\mathcal{L}(y, \hat{y})$ is the original loss function of a model, λ is the regularisation parameter, w is the weight of a feature, and $||w||_1$ is the 1-norm, or L1 norm, of w [49, 50, 60]. In contrast, L2 regularisation term is equal to the sum of the squares of the feature weights

$$\hat{\mathcal{L}}_{(L2)} = \mathcal{L}(y, \hat{y}) + \lambda \sum_{i=1}^{M} ||w_i||_2^2, \tag{1.7}$$

where $||w||_2$ is the 2-norm, or L2 norm, of w [49, 60, 61].

The role of λ in Equation 1.6 and Equation 1.7 differs, and this becomes apparent when examining how the weights are updated. Consider updating the weights using an iterative optimisation algorithm, such as the gradient descent algorithm, which requires the first derivative of the loss function. For L1 regularisation, the updated weights are given by:

$$\begin{aligned} w_{new} &= w - \eta \frac{\partial \hat{\mathcal{L}}_{L1}}{\partial w} \\ &= \begin{cases} w - \eta(\mathcal{L}(y, \hat{y}) + \lambda), & \text{if } w > 0 \\ w - \eta(\mathcal{L}(y, \hat{y}) - \lambda), & \text{if } w < 0 \end{cases} \end{aligned} \tag{1.8}$$

where η is the step size. The presence of λ in Equation 1.8 tends to shrink the weights to zero and achieve sparsity, i.e., a model with few non-zero coefficients. In effect, the features deemed less relevant are disregarded from the model by updating their weights to zero and, in turn, reducing model complexity. This type of regularisation essentially functions as a form of feature selection, except that it can incur a high computational

cost during model optimisation with a high-dimensional data set [49, 60, 62]. Our proposed pipeline addresses such an issue by eliminating irrelevant features prior to optimising the predictive model to realise model generalisation.

For L2 regularisation, the updated weights are given by:

$$\begin{aligned} w_{new} &= w - \eta \frac{\partial \hat{\mathcal{L}}_{L2}}{\partial w} \\ &= w - \eta(\mathcal{L}(y, \hat{y}) + 2\lambda w), \end{aligned} \tag{1.9}$$

where updating the weights depends on the regularisation term $2\lambda w$ (cf. Equation 1.8). The regularisation term encourages the weights of the less relevant features to shrink and penalises the model for attempting to make very accurate predictions. The tendency to reduce the weights to zero is relatively low compared to L1, and the weights are reduced more evenly to minimise the contribution from the less relevant features rather than eliminate them. Consequently, L2 regularisation does not yield sparse solutions and feature-selection capabilities do not exist. Therefore, it leads to low model interpretability, although model generalisation can be realised [49, 60, 61].

The applicability of L2 regularisation extends to data sets that suffer from multicollinearity. High degrees of correlations between features can cause large variances in the weights, leading to unstable weights sensitive to small changes in the model. The high variance arises because it becomes increasingly difficult to determine the contribution of each feature to the variation in the target variable as the features tend to change in unison. In general, the ability to realise model generalisation diminishes as the severity of multicollinearity increases [60, 63, 64].

L2 regularisation counteracts such effects by constricting the magnitude of the weights, while allowing each feature to contribute to the prediction of the target variable. This step is equivalent to taking a weighted average of the highly correlated features in computing the variation in the target variable. The constriction of the weights is achieved by the regularisation term in Equation 1.7, a penalty equivalent to the square of the magnitude of the weights. The impact of correlated features is then reduced by minimising the sum of squares of the weights, penalising large weights, which in turn lowers the variances of the weights [49, 60, 63, 64].

Multicollinearity effects are treated differently in L1 regularisation. High correlations between features are reduced by arbitrarily eliminating some of those features by shrinking their weights to zero [49, 60, 62]. In the feature-selection process of our pipeline, the effect of multicollinearity is circumvented through correlation and

hierarchical cluster analyses, coupled with a suite of statistical feature analyses, which assists in determining the features that are to be eliminated.

The description of L1 and L2 regularisation above is directly applicable to models such as artificial neural networks and logistic regression. However, there is a subtle difference in the context of GBDTs. In GBDTs, regularisation aims to control the degree of model complexity by restricting the growth of trees. L1 regularisation is applied to leaf scores rather than directly to the weights of the exploratory features, which effectively restricts the depth of trees and minimises the contribution from irrelevant features. L2 regularisation tends to restrict tree growth by penalising large leaf scores, while moderately punishing those associated with irrelevant features. Both are consistent within the framework of GBDTs since model complexity is determined by the tree complexity, which is defined by several hyperparameters such as the number of leaves, samples per leaf, tree depth and gain criterion for optimal split points [41, 42].

1.4 Materials Informatics

As introduced in earlier sections, materials informatics represents an interdisciplinary domain that merges materials science with informatics and data-driven methodologies to expedite the discovery, design, and development of novel materials. As an emergent field, it harnesses machine learning and statistical analysis to derive valuable insights from extensive data sets encompassing material properties and compositions. The primary aim of materials informatics is to enhance the prediction and optimisation of material properties, thereby facilitating the discovery of new materials more efficiently than is possible with traditional experimental approaches. Here, we will introduce various material representation techniques, specifically featurisation methods relevant to organic and inorganic materials.

1.4.1 Molecular and Chemical Descriptors

A variety of descriptors are used to generate feature vectors that accurately represent materials. Structural descriptors achieve this by encoding structural information through functional mapping from atomistic systems to high-dimensional vector spaces. Concurrently, basic chemical descriptors transform chemical names into a discretised format, typically represented in fractions, indicating the presence and quantity of chemical elements. Selecting and utilising features that comprehensively depict the

chemical space and crystal or molecular structure is essential for achieving accurate material characterisation and prediction.

Many descriptors have been developed over recent years, and they can broadly be classified into two categories: local and global. A local descriptor encodes local spatial regions of an atomic structure. This makes it desirable for predicting localised properties, such as adsorption energies and atomic forces, in addition to properties derived from aggregating localised spatial contributions. Notable examples of local descriptors include the Smooth Overlap of Atomic Positions (SOAP) [27] and Atom-Centered Symmetry Functions (ACSF) [28]. In contrast, global descriptors encapsulate the entire spatial region of an atomic structure, making them apt for predicting properties such as formation energies and band gaps. Examples of such global descriptors include the Ewald Sum Matrix [29], Sine Matrix [29], Coulomb Matrix [30], Bag of Bonds [31], and Many-body Tensor representation [65].

In practice, chemo-structural descriptors are generated by concatenating extensive feature vectors derived from both structural and chemical descriptors. A notable example is the Classical Force-field Inspired Descriptor (CFID), which integrates five distinct sub-methods [66]. These sub-methods encompass pairwise radial distributions, nearest-neighbour interactions, bond-angle variations, dihedral-angle configurations, and core-charge distributions. The rationale is to equip the machine-learning model with the capacity to discern chemical structure-property relationships that far surpass the capability of manual analysis. These comprehensive feature vectors are then used in model training, facilitating the statistical prediction of either properties of unseen materials in regression tasks or the classification of materials into distinct groups or classes in classification problems. In certain instances, zero-padding is necessary to achieve a uniform output length, a common prerequisite for most machine learning algorithms.

Additional features can be generated by calculating various statistical metrics, including the average, range, and standard deviation of different elemental properties among the constituents of a chemical composition. For instance, specific elemental properties such as melting temperature and electronegativity for each chemical element within a composition can be obtained from data sources from Pymatgen [67], Magpie [68], and Deml [69]. These elemental properties are then used to create an array of statistics, providing a detailed feature representation of the elemental characteristics. Moreover, a variety of features can be computed using various Python libraries such as Matminer [70] and Pymatgen [67].

Another noteworthy approach involves using of graph-based methods, which vary in form but fundamentally involve constructing a graph with nodes representing the elemental properties and edges illustrating the bonding characteristics, thus reflecting the material structure under consideration. This graphical representation facilitates the detailed characterisation of node-specific features of the constituent elements, while the edges depict the bonding interactions. These graph vectors undergo convolution over a predefined number of iterations before aggregating into a comprehensive feature representation or embedding [33]. This embedding is then used to train a predictive model. Material embeddings can also be sourced from pre-trained models such as the Materials Graph Network (MEGNet) [34], which is trained on extensive data sets of materials. Leveraging these embeddings from MEGNet enables the application of transfer learning, improving the prediction of various of properties even with constrained data sets.

1.4.2 Simplified Molecular Input Line-Entry System

In organic materials, SMILES (Simplified Molecular Input Line-Entry System) represents chemical molecules [39]. SMILES is a chemical notation system grounded in molecular graph theory principles, enabling the precise depiction of molecular structures through natural grammar. The adoption of learning on SMILES has become widespread in predicting molecular properties due to its compact representation, which clearly delineates chemical elements, various bond types, branches, aromaticity, cyclic structures, and local chirality.

Numerous descriptors derived from the SMILES representation can be used to generate features that capture various characteristics of an organic material, such as functional groups and unique substructures. Examples include Morgan fingerprints, RDKit molecular fingerprints, ElemNet, Maccs keys, element property fingerprints, and atom pair counts [67–75]. These descriptors facilitate the extraction of detailed chemical information, enhancing the analysis and modelling capabilities within materials informatics. Later in Chapter 4 (Section 5.2), we demonstrate that the SMILES representation can generate a feature matrix or image that can be processed using CNNs. This process results in a vector embedding that effectively captures key features of a chemical compound. Such a system of molecular representation also introduces new possibilities within both unsupervised and self-supervised frameworks, enabling the prediction of molecular properties through language models [36].

Nevertheless, awareness of conformational effects is essential in spectroscopy and molecular chemistry [76]. These effects appear as variations in atomic arrangements within a molecule due to differing spatial orientations, known as conformations. Although spectroscopic data can provide insights into these conformational changes by representing a Boltzmann average of all conformers in a solution, SMILES strings cannot account for these variations. When a study does not differentiate between single or multiple conformations in experimental measurements, nor addresses potential alterations to the conformational ensemble caused by environmental interactions (such as solvent effects), it is imperative to consider these variables. This is particularly important when employing methodologies that solely rely on SMILES representation for deriving input features.

1.5 Scope of this Thesis

Thus far, an overview of advancements in materials discovery processes has been presented, emphasising the crucial role of machine learning. Additionally, essential concepts have been introduced for further exploration in subsequent chapters, along with a discussion of the shortcomings in current computational materials science practices. Notably, these shortcomings include the large computational cost incurred during the training of machine learning models, which escalates with an increasing volume of data and the indiscriminate or naive inclusion of extensive input features. While using unsupervised learning with a large volume of unlabelled data presents compelling possibilities, it requires substantial computing resources. Such demand often cannot be justified given the marginal enhancements it offers over alternative state-of-the-art methods reported in the scientific literature. Moreover, current practices frequently overlook the need to thoroughly understand the interactions among features in relation to the target variable or class. In response, we propose a novel, fully systematic modelling approach that minimises human intervention during the model development phase. We have formulated a workflow that essentially serves as a preemptive regularisation step before model training and optimisation. This method efficiently isolates a select subset of features most pertinent to the target variable or class, facilitating a deeper understanding of the feature interactions critical for accurate predictions, while improving model generalisation on out-of-sample data.

To that end, this thesis is dedicated to the development and application of our proposed workflow across a range of research domains, encompassing condensed matter physics (solid-state physics), and various areas within materials science and engineering. The overarching aim is to establish a fully systematic, general-purpose materials discovery workflow that minimises both the requirement for computational resources and the potential for human biases in model development and prediction processes.

In Chapter 2, a comprehensive overview of our proposed workflow is presented, detailing each component and its respective role within the system. It also elaborates on the foundational theoretical concepts and key background information that underpin the development and application of the workflow. The utility of this workflow is subsequently illustrated through its application in predicting various material properties against the results derived from high-throughput calculations and experiments.

In Chapter 3, we employed the workflow to predict material properties within the realm of inorganic chemistry. The chapter delves into the automated prediction of Curie temperature of ferromagnetic materials and the construction of magnetic phase diagrams. We further demonstrate the predictive capabilities of the workflow. For instance, a predictive modelling of high-entropy alloys and amorphous metallic alloys is performed, where the results are validated through two case studies: a regression analysis on the bulk modulus of high-entropy alloys, and a classification analysis on the glass-forming ability of amorphous metallic alloys.

Chapter 3 also showcases the prediction of band gaps in inorganic materials using a multifidelity modelling approach, where the integration of high-throughput calculations with experimental measurements reduced the number of input features that are required for a sustained model performance. We also present the results of the prediction of the critical temperature of superconductors using only their chemical compositions, where we detail the development of a classification model that is designed to identify chemical compositions that are likely to demonstrate a critical temperature above 10 K. Simultaneously, a regression model was developed to estimate critical temperature values. This dual-model approach allows for a comprehensive analysis of how chemical constituents influence the superconducting properties, providing insights into the behaviour of materials at critical thresholds.

Chapter 4 extends the application of our workflow to the domain of organic chemistry, demonstrating its versatility and broad utility. The workflow is evaluated on over 20 complex predictive tasks, including the prediction of quantum mechanical properties and drug-related properties, using SMILES strings as the sole input representation. To enhance predictive accuracy, we incorporated substructure vector embeddings into

a semi-supervised learning framework, showcasing the capability of the workflow to leverage other advanced techniques. The flexibility of the workflow is further highlighted through its integration with existing machine learning methods, such as convolutional neural networks, to predict optical absorption wavelengths in organic molecules. These studies underscore the workflow's extensive adaptability and effectiveness across diverse applications in materials science, establishing it as a powerful tool for advancing predictive modelling in both organic and inorganic chemistry.

Chapter 2

Gradient Boosted and Statistical Feature Selection Workflow for Materials Property Predictions

This chapter introduces a machine learning-based approach for materials discovery that addresses the challenges of feature interpretability and relevance in the context of model development and prediction. It also focuses on reducing the computational burdens inherent in model optimisation, particularly when managing large data sets. A workflow is developed that incorporates a gradient boosting framework with statistical feature analyses, which systematically refines features that are most salient to the target properties of materials. This method minimises feature redundancy and eliminates the need for costly regularisation techniques during model optimisation and training. The efficacy of this approach is validated through various case studies in the subsequent chapters, where we show that the results are either superior or comparable to the state-of-the-art methods.

This chapter is based on **Jung, S. G.**, Jung, G., Cole, J. M., Gradient boosted and statistical feature selection workflow for materials property predictions. *J. Chem. Phys.*, 2023, 159 (19), 194106.

Prof. Jacqueline Cole and I conceived the overarching project and designed the study. I developed the workflow, performed the data acquisition and featurisation, the statistical analyses, the model training and fine-tuning, and analysed the data under the PhD supervision of Prof. Jacqueline Cole. Guwon Jung assisted with the data gathering and the model development. All figures and tables are reprinted with permission from *J. Chem. Phys.* 159, 194106, Copyright 2023 American Institute of Physics.

2.1 Introduction

Computational methods have played a vital role in accelerating the discovery of novel materials in many areas of research, providing profound insights from a purely scientific discovery and developing novel technologies. Over the last few decades, high-throughput computational calculations, typically based on Density Functional Theory (DFT), have become a preferred alternative or complement to experimental research which is primarily conducted on a trial-and-error basis; this often incurs substantial overhead costs. The process of material characterisation, to a certain extent, has been streamlined via *ab initio* methods as they allow for the computational simulation of materials and their properties. This feat has enabled an exploration of the vast chemical and property space within various research domains, such as batteries [77–81] and photo-electrochemistry [82–84], that has proceeded at a pace which cannot be realised via traditional design-to-device workflows [85–88].

These efforts within computational materials science have led to the creation of repositories with extensive data sets of chemical structures and properties, notably the Materials Project (MP) [5, 6] and the Open Quantum Materials Database (OQMD) [7]. With the availability of such aggregated *ab initio* chemical information and the emergence of big-data initiatives, data-driven approaches have gained a surge of interest due to their ability to process and analyse high-dimensional data sets, from which previously unseen patterns and relations can be deduced. In materials science, they have demonstrated accurate predictions of structures and properties, including band gaps [8–10] and formation energies [11, 12]. These studies exemplify the effectiveness of materials screening for the realisation of novel materials within highly complex feature spaces for various applications [13–17].

Data-driven approaches leverage materials informatics [1, 2] and machine learning (ML) [23–26], and a typical workflow involves transforming the *ab initio* chemical information into a machine-readable format using so-called descriptors [27–31]. Chemo-structural descriptors can be created by concatenating lengthy feature vectors from both structural and chemical descriptors [66]. The rationale is to allow the machine-learning (ML) model to infer chemical structure-property relations that are far beyond the capability of manual analysis. The resulting feature vectors are subsequently used for model training, which enables the statistical estimation of: (i) properties of unseen materials in a regression problem, or (ii) the class or group to which they belong in a classification problem.

There are two stumbling blocks associated with the workflow described above. These comprise: (i) the lack of model interpretability since ML algorithms are treated as

black boxes, and (ii) the computational requirement that necessitates high-performance computation through aggregated computing power. Although the use of ML in materials science is still relatively new [25], the black-box treatment means that understanding the feature relevance and interactions within a model is often neglected, while a greater emphasis is put on performance metrics; this leads to overfitting. Moreover, due to the rapid growth of chemical data [5, 7], a high computational cost is incurred during model optimisation. This cost becomes even more apparent with an increase in the number of irrelevant features that are naively considered and the complexity of the algorithm, which limits model scalability and generalisation. Resolving such limitations in a materials-discovery pipeline or workflow will shorten the timeline required to broaden the materials landscape in search of novel materials with specific structures and properties.

In this work, we present an ML-based feature-selection workflow that aims to address these shortcomings that become amplified with an increasing dimensionality of the data set [89]. Specifically, we use a distributed gradient-boosting framework [40–42], in conjunction with exploratory data and statistical analyses [43–48] and multicollinearity treatments, to identify and select a subset of features within a highly complex feature space; this affords minimal feature redundancy and maximal relevance to the target variable or classes. The efficacy and the efficiency of the workflow are demonstrated through materials-property predictions that are comparable to state-of-the-art reports from the scientific literature.

We show that our methodology can be employed in such a way that standard regularisation techniques, such as L1 [49], L2 [49, 50] and dropout [51], are not necessary during model optimisation. Circumventing the need for such techniques, especially during the hyperparameter optimisation stage, achieves model generalisation at a significantly low computational cost, which is advantageous for large-scale data-driven materials discovery.

2.2 System Overview

The workflow is illustrated in Figure 2.1 and it consists of four stages: (i) data acquisition and featurisation, (ii) feature selection, analysis and engineering, (iii) model development and optimisation, and (iv) materials-property prediction. In this section, each stage is discussed chronologically. We discuss how our workflow addresses the flaws of a typical ML-assisted property-prediction workflow, which rely heavily on regularisation techniques.

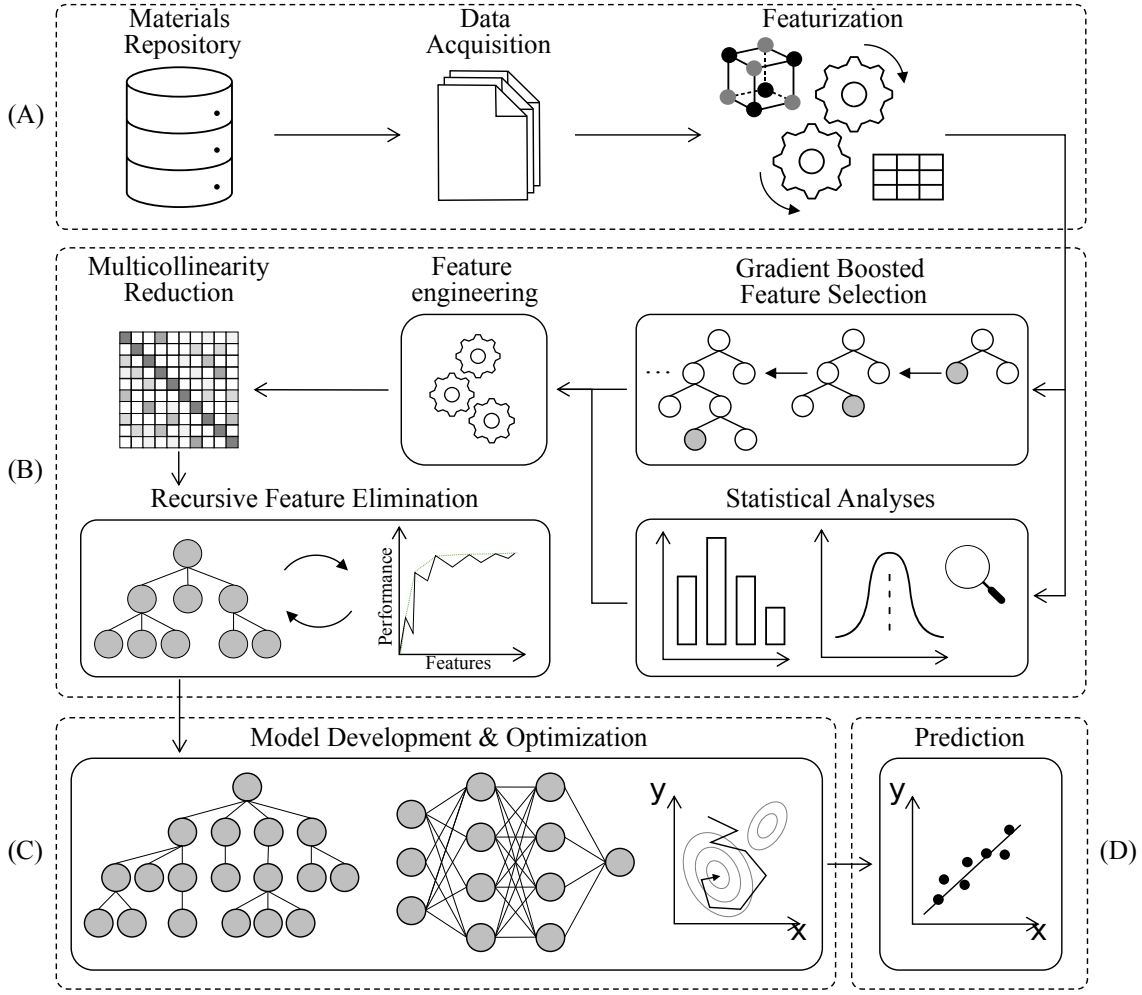


Fig. 2.1 Overview of our operational workflow compartmentalised into four distinct stages: (A) data acquisition and featurisation, (B) feature selection, analyses and engineering, (C) model optimisation, and (D) materials-property prediction. Each stage is encapsulated by the dashed borders.

2.2.1 (A) Data Acquisition & featurisation

Data Acquisition

We utilised a collection of data sets assembled from previous scientific studies, some of which are a part of the benchmark test suite known as the Matbench test suite v0.1 [90]. Using such a benchmark test suite also serves as a point of reference to which we can compare the performance of the proposed workflow. In certain case studies, we compiled customised data sets through rule-based scraping techniques or advanced ML-based approaches like ChemDataExtractor [21, 22]. Our primary focus is on processing aggregated data; therefore, we will not delve into various data acquisition methods.

Featurisation & Data Preprocessing

A high-dimensional feature vector was generated by leveraging an extensive set of chemo-structural descriptors. These included using composition and structure featurisers module from Matminer [70] and the classical force-field inspired descriptors from JARVIS [66]. The latter provides five sub-methods. These include pairwise radial, nearest neighbour, bond and dihedral angles, and core-charge distributions, leading to 1557 features per structure. Additional descriptors were created using Pymatgen [67] to identify functional groups and relevant crystal symmetries, and featurise specific chemical information of interest. Furthermore, we expanded the featurisation stage to allow users to access and generate further features based on data from Pymatgen [67], Magpie [68], Deml [69], and the neural network embeddings of elements generated using the MatErials Graph Network (MEGNet) [34]. When using all of the descriptors mentioned herein, a base feature vector (i.e., pre-feature engineering) of length 4052 is created for each material by concatenating the output of the descriptors. To demonstrate the general purpose of the proposed workflow, we choose to exclude some of the descriptors in the prediction of certain material properties.

For example, in some benchmark data sets (from previously published papers), computationally expensive features are included to improve model performance, leading to a computationally expensive model overall. In our study, when we used such a data set, our sole purpose was to demonstrate that the proposed workflow can generate results that are either better or comparable to the previously published work, with the added benefit of model interpretability.

Before proceeding to feature analysis and selection, a data splitting is performed (e.g., with a split ratio of 1 to 4 for the test and the training sets, respectively). A

stratified data splitting is preferred when conducting a classification analysis using an imbalanced data set, while a random data split is carried out for regression analyses. We then retained 20% of the training set as an out-of-sample validation set for the feature selection and model optimisation stages. The numerical features in the training set were normalised so that there were no biases due to the differences in the value ranges between different features. Meanwhile, categorical features were one-hot encoded to transform them into discrete variables. As for the test set, the normalisation was performed separately using a scaling factor that was defined by the training set. This process ensures a consistent scaling of features, and guarantees that feature analysis and model optimisation are independent of the test set.

Methods for Class Imbalance

An imbalanced data set in classification is one whose distribution is unequal to those of other classes. This imbalance often introduces biases towards minority classes as they are under-represented. In this study, we addressed the class imbalance at the data level, that is, to oversample minority and/or undersample majority classes to achieve an equal distribution of the classes without using a weighted cost function or tuning classification thresholds. Specifically, we examined three variations of the Synthetic Minority Oversampling Technique (SMOTE) [91, 92], Adaptive Synthetic (ADASYN) oversampling [93] and two variations of Random Oversampling (ROS) [94, 95] to synthesise feature vectors that resemble the characteristics of the minority classes within the high-dimensional feature space. We found that the ROS based on a smoothed bootstrap re-sampling technique [95], referred to herein as smoothed-ROS, was the most effective in alleviating potential learning biases. The technique introduces a perturbation to the oversampled feature vectors by adding Gaussian noise that can be regulated by a smoothing matrix. This prevents overtraining of a model on certain values of a feature, which helps to improve the generalisation of the model. Our workflow accommodates all the minority oversampling techniques mentioned herein, as user options.

2.2.2 (B) Feature Selection, Analyses and Engineering

Gradient Boosted Feature Selection

Feature selection is the process of selecting a subset of input variables, or features, when developing a predictive model. However, it is often overlooked in the field of computational materials science. The standard practice is to naively consider all the

input features during model training, as ML algorithms can infer complex patterns between features and the target material property. This often leads to overfitting and hinders model generalisation. Various regularisation techniques are employed to address these problems, which, unfortunately incurs substantial computational cost.

Here, we demonstrate the viability of using a distributed gradient-boosting framework in conjunction with statistical analyses to perform feature selection that can: (i) identify key features by their relevance to the target variable or classes, (ii) provide insights into feature interactions, and (iii) afford scalability. Specifically, we leverage the gradient-boosted decision trees (GBDTs), as presented by Chen et al. [41] and Ke et al. [42], to evaluate the relevance of features to a target variable or class by examining the significance of split points in learning the decision trees (DTs). These two methods implicitly embed regularisation terms into the feature analysis steps of our ML workflow rather than in a subsequent optimisation step which would reflect the conventional approach.

When trees are grown using Chen's implementation [41], candidate split points are evaluated based on the loss reduction observed after a split. This is calculated using:

$$\mathcal{L}_{split} = \frac{1}{2}\Big[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{(H_L+H_R)+\lambda}\Big] - \gamma, \tag{2.1}$$

where $G_{L,R} = \sum_{x_i \in I_{L,R}} g_i$ and $H_{L,R} = \sum_{x_i \in I_{L,R}} h_i$ are the summation of the first and second-order derivatives (i.e., gradients) of the loss function over sample sets I_L and I_R of left and right leaf nodes after the split, respectively, and λ and γ are the regularisation parameters [41]. In Ke's implementation [42], the loss reduction of splitting a leaf node on a feature is referred to as the 'variance gain'. The mathematical formulation of the variance gain of a split, while being analogous to Equation 2.1, does not parameterise the second-order gradient. There is also a partitioning of the first-order gradient by different subsets of samples due to the implementation of Gradient-based One-Side Sampling (GOSS). By defining subset A, which contains the top-a fraction of samples with large gradients, and its complement subset A^c, which contains the remaining $(1-a)$ fraction of the samples with smaller gradients, subset B is defined as b fraction of samples in A^c that are randomly selected. The estimated variance gain of splitting a leaf node, or a split point p, with feature m is then given by:

$$\hat{V}_m(p) = \frac{1}{N}\Big[\frac{(G_{A_L} + \frac{1-a}{b} G_{B_L})^2}{N_L^m(p)} + \frac{(G_{A_R} + \frac{1-a}{b} G_{B_R})^2}{N_R^m(p)}\Big], \tag{2.2}$$

where $G_{A_{L,R}} = \sum_{x_i \in A_{L,R}} g_i$ and $G_{B_{L,R}} = \sum_{x_i \in B_{L,R}} g_i$ are the summation of the first-order gradient over subsets $A_{L,R}$ and $B_{L,R}$, where the subscripts indicate the left and right of the split point p, respectively, $N^m_{L,R}(p)$ are the number of samples to the left and right of the split according to feature m at point p, and $\frac{1-a}{b}$ is the normalisation term [42]. Therefore, optimal tree growth can be realised using the split points that maximise Equation 2.1 or 2.2. The relevance of a feature is then calculated through the summation of the loss reduction or variance gain, which we will refer to as the total loss reduction hereafter. See Section 1.3.2 for more details of GBDTs; two methods for identifying the candidate split points, namely, the pre-sorted [96, 97] and histogram-based [98–100] algorithms; and two novel techniques in Ke's implementation [42], namely, GOSS and Exclusive Feature Bundling (EFB), where both GOSS and EFB are included as user option in the workflow.

Once the feature relevance was determined, a separate model of GBDTs was trained recursively by considering an increasing subset of features, starting with the feature that exhibits the highest relevance score. Performance metrics were then computed for each iteration until the performance metrics of the validation set had converged and exhibited no further improvement. For classification, the metrics considered were the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC), the Average-Precision score (AP), the average Hamming Loss (HL), and the F1-score. For regression, the metrics considered were the coefficient of determination (R^2), mean absolute error (MAE), and root-mean-square error (RMSE). The convergence of these performance metrics was used as the criterion to determine the preliminary subset of features.

Here, we emphasise that the gradient-boosting framework was not only used to determine feature relevance and understand the feature interactions within a model, but we incorporated the results in a recursive manner to eliminate features that are deemed irrelevant or have a weak contribution in the prediction of the target property or class.

Feature Analyses

Having generated the preliminary subset of features using these GBDT models, which we refer to as gradient boosting feature selection (GBFS) hereafter, we performed independent statistical analyses to study the exploratory features. Specifically, we conducted four types of hypothesis-based testing methods in addition to mutual information and discriminant analyses [101]. This enabled us to identify statistically significant features with respect to a target variable, class or other exploratory features,

and measure the mutual dependence between them. The results from the statistical analyses are used to refine the preliminary subset of features that were selected based on the gradient-boosting framework.

The hypothesis-based testing methods employed in this study are bivariate in form, as we are primarily concerned with understanding the causal relationship between an exploratory feature and the target variable or class. These testing methods included: (i) a comparison of means using either the F-test in one-way analysis of variance (ANOVA), the two-sample t-test or a Z-test for a continuous and a categorical feature, depending on the number of categories or classes [102–106], (ii) a categorical association using Pearson's chi-squared (χ^2) test for two categorical features [107–109], and (iii) a correlation analysis using Pearson's correlation coefficient (R) for two continuous features, where the ANOVA approach to regression analysis is taken by converting R into a regression F-statistic [43, 108]. One of these hypothesis-based testing methods was used for performing this statistical inference, the selection of which depended on the type of features being examined. The statistical significance of a given exploratory feature was inferred from the test statistics which resulted from testing the hypotheses that an association exists between two features.

We also explored two additional methods for feature analysis that are alternative to hypothesis-based testing: (i) mutual information (MI) and (ii) discriminant analysis using logistic regression. We used the concept of MI to measure the level of dependency between two features in both classification and regression problems, by quantifying the amount of information, or entropy, realised for a feature as a result of the observation of another. For a pair of features, MI determines the difference between their joint distribution and the product of their marginal distributions, with a higher value of MI indicating a greater dependency between the two features [110]. We adopted the MI estimator presented by Kraskov et al. [46] and Ross [47] to calculate the MI, which is based on entropy estimations from k-nearest neighbour distances, as originally proposed by Kozachenko et al. [48].

A discriminant analysis is generally used in classification to determine features that can discriminate or distinguish between target classes. In this study, we conducted a discriminant analysis based on logistic regression, so-called logistic discrimination, for both binary and multi-class problems. To determine how discriminative a feature is towards the target classes, we calculated the change in either the odds ratio or the log-odds ratio as the value of the feature was increased by one unit. In contrast, all other features were kept constant [44]. The change in the odds ratio indicates how discriminative the feature can become towards the target classes by the variation of

its value. Such a technique was leveraged to identify highly discriminative features, and we used the result to improve the odds or posterior probability of a minority class through feature engineering. For a multi-class problem, there are $C-1$ number of log-odds ratios to consider with respect to a pre-selected target class. This leads to $C-1$ independent binary logistic regression models [49, 111].

Features Engineering & Multicollinearity Reduction

The preliminary subset of exploratory features, selected via the combined efforts of the GBFS and the statistical analyses, provided the basis for new features to be engineered. Our rationale for engineering additional features was two-fold. We sought to: (i) create new features to address biases that may have existed due to the under-representation of minority samples, and (ii) realise unforeseen feature relations that may be of statistical significance, thereby providing new insights within the scientific domain of interest.

A brute-force approach was adopted to generate additional features by considering a number of features, m, deemed statistically relevant to the two objectives mentioned above. We computed two permutations of m features (i.e., $P_2^m = \frac{m!}{(m-2)!}$); that is, the differently ordered arrangements of a two-feature subset of m features. For a given pair of features from the list of permutations, a new feature was then defined as the ratio between the two features, resulting in P_2^m additional features.

A correlation analysis was performed on the data set that had been augmented by incorporating the newly engineered features and the preliminary subset of exploratory features. This analysis identified multicollinearity within the data set by constructing a Pearson's correlation-coefficient matrix and eliminating features from the matrix elements that shared a correlation coefficient of 0.85 or higher. This step helped minimise feature redundancy and remove unnecessary features among engineered ones.

A hierarchical cluster analysis was incorporated to achieve further multicollinearity reduction. We implemented an agglomerative clustering methodology, whereby each feature is considered as its own cluster in the first instance, and each cluster is subsequently paired or grouped to form larger clusters moving up the hierarchy. We chose Spearman rank-order correlation as the metric to measure the dissimilarity between the features. Ward's minimum variance criterion is the linkage criterion, specifying the dissimilarity as a function of the pairwise distance between the features [112]. We aimed to minimise the total intra-cluster variance by identifying the pair of clusters exhibiting the smallest increase in the total intra-cluster variance when merged. Once the hierarchical clusters had been formed, we applied a distance threshold of one unit of Ward's linkage distance, and removed collinear features. That is to say,

clusters with Ward's linkage distance of one or less were deemed sufficiently correlated that similar information could be obtained by retaining just a single feature from each cluster that had been formed using such a distance threshold. This, in turn, further reduced the multicollinearity within the data set.

The threshold applied in the hierarchical cluster analysis can be adjusted to either increase or decrease the stringency of the criterion. Due to the prior treatment of the multicollinearity using the correlation analysis, in addition to the performance-based feature elimination described below, we avoided applying an overly stringent criterion. This also reduced the risk of potential information loss as a result of eliminating features that were moderately correlated.

The final step of the feature-selection process was to identify the optimal (i.e., the smallest) subset of features among the candidates by employing a wrapper method, based on recursive feature elimination (RFE) [113]. RFE follows a greedy-search approach to recursively train an estimator with different combinations of features against a performance metric, until the smallest subset of features is realised. Our implementation used GBDTs as the estimator, and the performance was evaluated using the weighted F1-score, the harmonic mean of precision and recall, in classification, and mean squared error (MSE) in regression with n-fold cross-validation.

For the classification analysis, we consider the accuracy and F1-score, where the latter is defined as the harmonic mean of the precision and recall as follows

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2.3}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{2.4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2.5}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2.6}$$

where TP and TN are true positive and true negative, and FP and FN are false positive and false negative, respectively. For the regression analysis, we consider the mean absolute error (MAE), the mean squared error (MSE), and the coefficient of determination that is defined as the square of the Pearson correlation coefficient, R

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{2.7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{2.8}$$

$$R = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}, \tag{2.9}$$

where y and $\hat{y}$ are the true and predicted values, respectively, over N number of samples; $Covar(x,y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N}$ is the covariance between x and y; $Var(x) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$ and $Var(y) = \frac{\sum_{i=1}^{N}(y_i-\bar{y})^2}{N}$ are the variance of x and y; $\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N}y_i$ are the mean of x and y, respectively. The range of R is [-1, 1] and its value indicates the linear tendency of a quantity to change as the values of another is varied.

2.2.3 (C) Model Optimisation

Mitigating the Need for Regularisation

It is important to include a preface to this section to highlight that regularisation techniques such as L1 (i.e., LASSO regression), L2 (i.e., Ridge regression) and dropout were not implemented in our model optimisation stages. We argue that our proposed feature-selection process, coupled with the multicollinearity treatments, can supersede the role of standard regularisation in controlling the complexity of the model and preventing overfitting during model optimisation. See Section 1.3.5 for the detailed theoretical background of L1 and L2 regularisation.

The regularisation parameter in L1 [49, 50, 60] tends to shrink the weights of the features to zero and achieve data sparsity; that is, a model with few non-zero coefficients. In effect, the features deemed less relevant are disregarded from the model by updating their weights to zero and, in turn, reducing model complexity. In a tree-based model, L1 regularisation is applied to leaf scores rather than directly to the weights of the exploratory features, which effectively restricts the depth of trees and minimises the contribution from irrelevant features. This type of regularisation essentially functions as a form of feature selection, except that it can incur a high computational cost during model optimisation with a high-dimensional data set [49, 60, 62]. Our proposed workflow addresses such an issue by eliminating irrelevant features prior to optimising the predictive model using the aforementioned GBFS and statistical features analyses.

In contrast, L2 regularisation does not achieve data sparsity as the feature weights never go to zero and, therefore, feature-selection capabilities do not exist. However, the applicability of L2 extends to data sets that suffer from multicollinearity. High degrees of correlations between features can cause large variances in the weights, leading to

unstable weights that are sensitive to small changes in the model. The high variance arises because it becomes increasingly difficult to determine the contribution of each feature to the variation in the target variable as the features tend to change in unison. In general, the ability to realise model generalisation diminishes as the severity of multicollinearity increases [60, 63, 64]. L2 regularisation counteracts such effects by constricting the magnitude of the weights, while allowing each feature to contribute to the prediction of the target variable. This is equivalent to taking a weighted average of the highly correlated features in computing the variation in the target variable. The constriction of the weights is achieved by the regularisation term, a penalty equivalent to the square of the magnitude of the weights. The impact of correlated features is then reduced by minimising the sum of squares of the weights, so as to penalise large weights, which in turn lowers the variances of the weights [49, 60, 63, 64]. In our workflow, the effect of multicollinearity is circumvented through the correlation and hierarchical cluster analyses since they assist with feature elimination in a less computationally expensive way while increasing model interpretability.

The hyperparameters of these ML models were optimised using a combination of grid search and Bayesian optimisation using Gaussian processes, i.e., a sequential model-based approach. A preliminary scan of the hyperparameter space was made using a grid search to identify the region in which Bayesian optimisation was to be applied. At this stage, most studies would then apply regularisation techniques; however, for reasons discussed earlier, our workflow does not require them. Bayesian optimisation used in this work is outlined in Section 1.3.4.

It should be noted that the order of the stages within the proposed workflow is chosen to maximise a set of performance metrics while simultaneously minimising the computational cost associated with the feature-selection process and the subsequent model optimisation. For instance, a pair of highly correlated features does not necessarily achieve identical loss reduction or a feature relevance score when growing the optimal tree during the GBFS stage of the workflow. Likewise, the feature rankings based on statistical significance testing will differ such that one is identified as having a greater relative association with the target variable; this will always be true, although the significance of the correlated feature may be of similar magnitude. For this reason, we perform multicollinearity reductions after the GBFS and statistical analyses, since the result could otherwise lead to the removal of a feature more relevant to a target variable; a consequence being the selection of a sub-optimal subset of features for the final ML model for a given correlation threshold of user's choice. Moreover, the GBFS stage eliminates the majority of exploratory features ($\gtrsim 90\%$) from consideration.

This significantly reduces the computational cost associated with the downstream feature-selection or elimination processes. Once the most relevant subset of features has been identified, which also maximises the chosen performance metrics, multicollinearity within the features is reduced to improve model generalisation and prevent overfitting. RFE then becomes the final step in our workflow, as we attempt to identify the smallest subset of features that maintains optimal model performance. During this process, the least relevant feature is systematically removed until the designated number of features is achieved or a performance metric stabilises. Given that RFE follows a greedy-search approach, it is also logical from a computational-cost standpoint to apply such a methodology to the final, and therefore the smallest, subset of features before Bayesian optimisation is applied to determine the model architecture.

2.2.4 (D) Materials Property Prediction

Subsequent chapters demonstrate the efficacy of our workflow through various case studies, whereby we develop an ML-assisted screening of inorganic and organic materials by performing various predictive tasks. We focus our discussion on the results of these case studies to provide greater insight into the roles of each component within our feature-selection workflow. Thence, we demonstrate the general applicability of our methodology by exploring additional regression and classification problems in computational materials science.

Chapter 3

Property Prediction of Inorganic Materials: Curie Temperature and Band Gap

This chapter explores the property prediction of inorganic materials using machine learning. It begins by focusing on the prediction of Curie temperature of ferromagnetic materials, exemplifying how the proposed workflow is leveraged to perform this predictive task. It emphasises the importance of a systematic approach to feature analysis and selection. A gradient-boosted feature selection process, combined with Bayesian optimisation, facilitates the creation of a predictive model that can estimate the Curie temperature of ferromagnetic materials solely based on their chemical compositions. These models are evaluated against experimental measurements, demonstrating the efficacy of our approach in predicting material properties. Subsequently, we examine an additional material property, namely the band gap of inorganic materials (Section 3.2).

This chapter is based on:

1. **Jung, S. G.**, Jung, G., Cole, J. M., Machine-Learning Prediction of Curie Temperature from Chemical Compositions of Ferromagnetic Materials. *J. Chem. Inf. Model.*, 2024, 64 (16), 6388–6409;

2. **Jung, S. G.**, Jung, G., Cole, J. M., Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow. *J. Chem. Inf. Model.*, 2024, 64 (4), 1187-1200;

Prof. Jacqueline Cole and I conceived the overarching project and designed the studies. I developed the workflow, performed the data acquisition and featurisation,

the statistical analyses, the model training and fine-tuning, and analysed the data under the PhD supervision of Prof. Jacqueline Cole. Guwon Jung assisted with data gathering, model development and optimisation, and contributed to the analysis of the results. All figures and tables are reprinted with permission from J. Chem. Inf. Model., Copyright 2024 American Chemical Society.

3.1 Prediction of Curie Temperature of Ferromagnetic Materials

3.1.1 Introduction

Magnetic materials have been pivotal in driving remarkable technological advancements, most notably in data storage, where their intrinsic physical properties have been exploited in conjunction with semiconductor technology. This has fundamentally transformed methods for encoding and retaining data bits, resulting in a technological paradigm shift, an accompanying proliferation of consumer applications in data storage, and the emergence of spin electronics as a novel research field [114–118].

These advances have stimulated efforts to better understand the underlying physics that governs the responses of magnetic materials to various energy terms and external factors such as temperature [114, 119]. Such responses engender intricate structure-property relationships in magnetic materials. These are characterised by the interactions between magnetic spins, or moments, which are contingent upon both the crystal geometry and chemical composition of the materials.

Moreover, a magnetic material experiences a loss of collective magnetic order at a specific temperature. This phase-transition temperature is known as the Curie temperature (T_c) for ferromagnetic materials. Their ordered magnetic properties cease at T_c or above, where only paramagnetic effects are observed; these have limited utility. Therefore, materials with a high T_c values are sought after to attain thermally stable magnetic states or magnetisation for functional applications. To this end, ferromagnetic materials whose T_c value is significantly greater than room temperature are particularly attractive. However, the multitude of potential interactions among magnetic ions and their surrounding environments renders the prediction of magnetic behaviour and properties challenging.

The exploration of magnetic materials has primarily been steered by experimental research efforts that rely on trial-and-error methods. Such research is very time-intensive, incurs significant operational costs and necessitates a substantial and sustained level of

specialist human capital since a rich amount of domain knowledge is critical to research progress. Therefore, the realm of magnetic materials discovery stands to gain from data-driven methodologies that facilitate the targeted design of novel materials based on a desired property, particularly through the application of machine learning (ML).

The accessibility of large volumes of chemical data, coupled with the rise of big data initiatives, has resulted in a growing interest in data-driven materials discovery. The proficiency of data science in processing and analysing large-scale, high-dimensional data sets can realise a design-to-device pipeline for materials discovery at a pace unattainable by conventional experimental processes. Data-driven approaches leverage materials informatics and ML [23, 25]. A typical materials-informatics workflow involves transforming chemical data into a machine-readable format using feature descriptors [27, 66, 120]. The generated features are subsequently used for model training, which facilitates the statistical prediction of: (i) properties of unseen chemical materials via a regression analysis, or (ii) the specific class or category to which materials are associated using a classification algorithm. The rationale is to empower ML models to deduce relationships between chemical compositions, material structures and their properties, which exceed the capability of manual analysis. These techniques have already demonstrated their prowess in accurately predicting chemical structures and properties for a variety of materials applications [25, 26, 121–124]. This includes using multi-fidelity modelling strategies that harness high-throughput computational calculations and experimental measurements in tandem [125, 126]. These examples showcase the effectiveness of materials screening for the realisation of novel materials within highly complex feature spaces.

Various ML techniques have been employed to predict T_c values for a range of ferromagnetic material applications. For instance, Court et al. [127] identified magnetocaloric effects in 2,448 chemical compounds from the Heusler alloy family, whose T_c values were mined from the scientific literature using the natural-language-processing toolkit, ChemDataExtractor [21, 22]. They trained a gradient-boosting regression model using the XGBoost python library [41] with 58 element-level features and the magnetic field features, realising a coefficient of determination (R^2) of 0.71 and a mean absolute error (MAE) of 59.8 K. This model was leveraged by Ucar et al. [128] in their study of magnetic entropy prediction.

Long et al. [129] conducted a study aimed at advancing the design of ferromagnetic materials using ML techniques. They employed random-forest algorithms for classifying materials as ferromagnetic or antiferromagnetic based on their magnetic ground states and for predicting the T_c value of ferromagnets. The study realised a classification

accuracy of 87% and an R^2 of 91% for the regression task. Their random-forest algorithm was trained on a data set of 1749 ferromagnetic inter-metallic compounds, intentionally excluding oxides and compounds that lack Cr, Mn, Fe, Co, and Ni atoms. Additionally, the analysis incorporated 139 chemical and 26 structural descriptor features for each compound.

Dam et al. [130] performed a regression-based feature-selection process to analyse and predict T_c in binary alloy compounds composed of $3d$ transition metals and $4f$ rare-earth elements, using ML techniques. They found that model accuracy was optimised when the first 5 to 10 descriptor features were used, with the concentration of the rare-earth element being the most salient among the 28 descriptors utilised. The study reported an R^2 of 0.96 and an MAE of 41 K, using a limited data set comprised of 108 experimental data. Additionally, they noted a gradual decrease in model performance with the incorporation of additional descriptors, aligning with expectations.

Notably, Nelson et al. [131] conducted an ML-driven study that predicted T_c values of ferromagnetic materials solely from their chemical composition. They trained a random-forest regression model using features based on their chemical compositions with ca. 2,500 chemical compounds. Their final model was trained using 129 features and achieved an R^2 of 0.87 and an MAE of 57 K on the test set, while a cross-validation R^2 of 0.81 was realised. A similar study was carried out by Belot et al. [132] in an attempt to predict T_c values using a larger data set of ferromagnetic materials, while examining a diverse range of methods for material representation. They trained and validated a random-forest model and a k-nearest neighbours algorithm using two sets of data, first with ca. 2,500 and second with ca. 3,000 chemical compounds. They reported that the random-forest model realised the highest accuracy using 85 features based on chemical composition, and the use of complex descriptors and dimensionality reduction did not improve the prediction results. In three-fold cross-validation, an MAE of 73 K with a standard deviation of 3.2 K was achieved using the first data set, while an MAE of 71 K with a standard deviation of 2.3 K was achieved on the combined data sets. Another study using a random-forest regression was conducted by Singh et al. [133], which attained a five-fold cross-validation R^2 of 0.91 with a root-mean-square error (RMSE) of 59 K. Nevertheless, this study used a relatively small data set of 220 ferromagnetic and ferrimagnetic compounds to analyse rare-earth-based materials. Moreover, a linear regression was used by Sanvito et al. [134] in order to accelerate the discovery of new ferromagnets in the Heusler alloy family. The regression was calibrated on experimental measurements of ca. 60 chemical compounds. A typical

error value in the prediction of T_c values was reported to be in the range of 50 K for two chemical classes of the form Co_2YZ and X_2MnZ.

This study presents a new way for predicting T_c values, aiming to assist researchers deliver more accurate and transparent data-driven designs of ferromagnetic materials through a systematic integration of feature engineering, analyses, selection, and optimisation processes [123]. Our proposed workflow, referred to as gradient-boosted statistical feature selection (GBFS) hereafter, integrates a distributed gradient-boosting framework, in conjunction with exploratory data and statistical analyses and multicollinearity treatments, to discern a subset of features that is highly relevant to the target variable or class within a complex feature space; this affords minimal feature redundancy and maximal relevance to the target variable or classes.

Our workflow is generalisable, as has already been demonstrated through its use in successfully predicting material-property relationships in other areas of scientific research using data from DFT calculations [123] and experimental measurements [125, 126]. Here, we apply the GBFS workflow to predict T_c values of ferromagnetic materials using literature values as training data that emanate predominately from experimental measurements. The performance of our models is compared to state-of-the-art ML approaches that predict T_c values. In order to impose a stringent standard on our modelling approach and to ensure a direct comparison with state-of-the-art results, we restrict our descriptor sets exclusively to those derived from chemical compositions, abstaining from complex feature descriptors. This decision recognises that experimental T_c values reported in the literature often lack comprehensive crystallographic information, while chemical compositions are readily available.

We herein collate a database comprising ca. 35,000 T_c values from the scientific literature. This database is divided into two segments. The first segment (Dataset 1) includes T_c values sourced from a variety of publications [135–141], predominantly being experimental values reported by Nelson et al. [131] and Belot et al. [132] Dataset 1 serves as the foundation for conducting a regression analysis of T_c values, showcasing the effectiveness of our GBFS workflow within this domain. Moreover, it facilitates an equitable comparison of our modelling approach with state-of-the-art models developed using a comparable data set. We broaden our analysis to the second data set (Dataset 2) by incorporating T_c values from AtomWork [142] into a blind-test scenario, exploring potential applications of our ML models beyond a conventional benchmarking evaluation against the state-of-the-art models. By ensuring that there is no overlap of chemical composition with Dataset 1, we will show that the blind-test prediction notably identifies 90 chemical compositions that exhibit T_c values which exceed room temperature. This

finding is subsequently corroborated by experimental measurements reported in the scientific literature. Figure 3.1 depicts the periodic table, highlighting the compositional space these data sets cover, which exhibit identical chemical coverage.

Furthermore, this study demonstrates the utility of our ML models through two primary applications: (i) the prediction of T_c values for rare-earth intermetallic compounds, and (ii) the generation of magnetic phase diagrams of various chemical compounds, where we assess how well our ML models perform when given the task to make out-of-distribution predictions. We emphasise that no overlaps exist between the previously aforementioned data sets. Our choice of evaluation data sets ensures that each of our ML modelling analyses maintains its distinctiveness, and our results are juxtaposed with corresponding studies utilising identical or similar data sets for comparative purposes.

Overall, we will show that our GBFS approach highlights the importance of thorough feature analysis and judicious selection over merely complex modelling. Thereby, we will notion that a common off-the-shelf model trained on features selected and engineered by the GBFS workflow can yield results comparable or superior to those reported in the literature. The workflow additionally provides insights into feature interactions and their relevance to the target variable. The collated database of T_c values and their corresponding chemical compositions is made available (https://github.com/Songyosk/CurieML) to serve as a resource for researchers engaged in designing ferromagnetic materials.

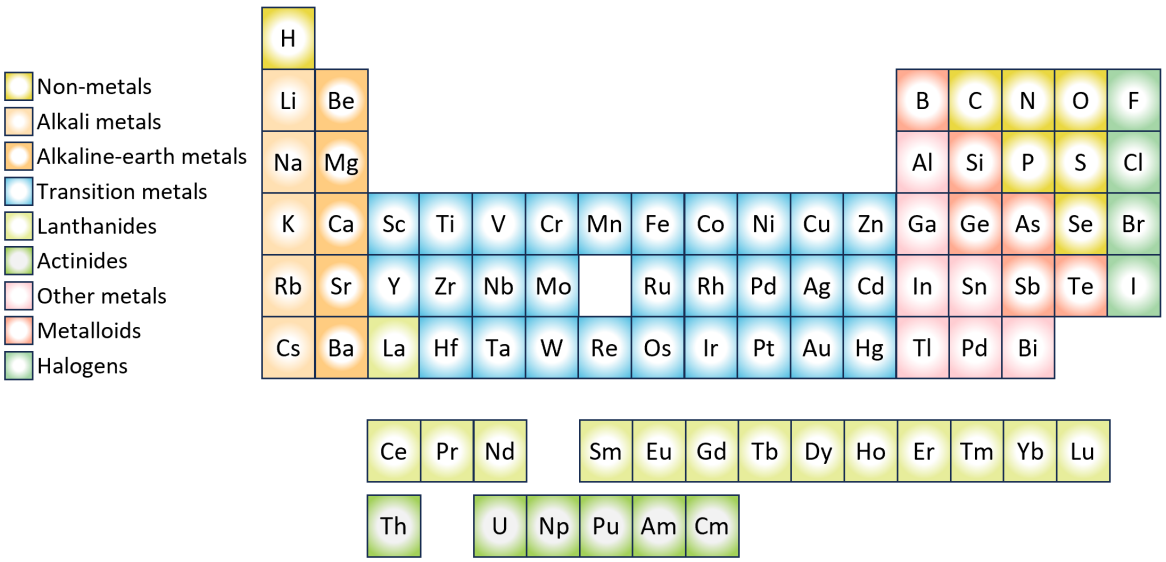


Fig. 3.1 Periodic table highlighting the compositional space represented by both Datasets 1 and 2 that are used in this work.

3.1.2 Methods

Featurisers

A high-dimensional feature vector was constructed using a set of composition-based descriptors, while incorporating the composition featuriser modules from Matminer [70] and Pymatgen [67]. Additional features were generated by computing statistics over elemental attributes that are specific to each chemical composition. These calculations draw from various data sources, including Magpie [68], Pymatgen [67], Deml [69], and neural network embeddings of elements that have been created using the Materials Graph Network (MEGNet) [34].

In materials informatics, these tools are instrumental in providing a range of composition-based features essential for analysing and predicting material properties. Matminer, developed by researchers at Lawrence Berkeley National Laboratory, is a comprehensive Python library designed for materials analysis. It provides an array of composition-based features derived from elemental properties, including atomic radius, electronegativity, and ionisation energy. Matminer enables the computation of various statistical measures related to these properties across constituent chemical elements. Moreover, it offers features related to the ionic character of material bonds, the oxidation states of elements, and electronic structure attributes such as the valence electron configuration of the elements.

Magpie (Materials-Agnostic Platform for Informatics and Exploration) is engineered to generate a diverse array of features from chemical compositions, optimising their use in ML models. The platform offers a wide range of features derived from elemental properties, encompassing both physical properties—such as density, heat capacity, and thermal expansion—and electronic properties—including electronegativity, valence electron concentration, and electron affinity. Magpie computes statistical measures of these properties for the constituent chemical elements.

Pymatgen (Python Materials Genomics) is a materials analysis library offering an extensive suite of tools designed to manage and analyse materials data. In addition to its utility on structural data, Pymatgen also supports the extraction of valuable composition-based features and is integrated with the Materials Project databases [5, 6]. The features generated by Pymatgen include the elemental fraction and atomic fraction of elements within the material, as well as comprehensive statistics on the physical and chemical properties of the constituent elements. These properties encompass melting points, boiling points, thermal conductivity, ionisation energies, and other relevant

parameters, providing a rich data set for materials characterisation and predictive modelling.

In essence, these packages enable the computation of four distinct types of attributes or measures. The first type is stoichiometric attributes, which rely solely on the proportions of elements present within a chemical compound. This category encompasses the number of elements in the compound and various L^p norms of these proportions or fractions. The second type involves elemental property statistics, including the maximum, minimum, mean, range, mode and mean absolute deviation of various elemental properties. In other words, these statistical measures are calculated by considering the properties of each element within a composition. This includes attributes such as the average atomic number, the highest group number on the periodic table, and the standard deviation of Mendeleev numbers among elements in a chemical composition. The third type is electronic structure attributes, which calculate the average fraction of electrons from the s, p, d, and f valence shells across constituent elements. The fourth type pertains to ionic compound attributes, assessing the potential to form an ionic compound by assuming constituent chemical elements are in a single oxidation state. It also includes measures of the ionic characteristics of a compound based on electronegativity.

MEGNet, an acronym for MatErials Graph Network, represents a graph neural network model tailored for materials science applications. It uses a graph-based framework to extract material properties from the structural configurations of molecules and crystals. MEGNet models, once trained, can function effectively as feature extractors for new materials, where the embeddings they generate are used as input features for subsequent predictive models. These embeddings effectively capture the chemical periodicity and inherent trends observable within the periodic table. While the individual interpretation of these embeddings can pose challenges, their utility is well-documented, particularly in transfer learning applications. For instance, embeddings derived from a model trained on a comprehensive data set can enhance the predictive capabilities of other models that are trained on more constrained or limited data sets. In our study, we employed MEGNet embeddings to augment elemental characterisation, facilitating the application of transfer learning from pre-trained models to improve the prediction of T_c values.

Gradient-Boosted Statistical Feature Selection Workflow

Figure 3.2 depicts our overarching GBFS workflow which we applied to predict T_c values solely from chemical composition. Full details of this methodology have been

described previously by Jung et al. [123]. The method as applied to this specific T_c prediction challenge is provided herein.

The GBFS workflow integrates several key components: (i) a gradient-boosting framework to identify a subset of features that maximise relevance to the target variable or class; (ii) statistical analyses of exploratory features to identify those that are statistically significant to the target variable or class; (iii) a feature engineering step for generating additional features; (iv) a two-step multicollinearity reduction process involving correlation and hierarchical cluster analyses to minimise feature redundancy; (v) a recursive feature-elimination process; and (vi) Bayesian optimisation to determine the architecture of the final predictive ML model.

While a comprehensive description of each component of our GBFS workflow has been provided by Jung et al. [123], we herewith provide a summary of its key attributes to ensure clarity and delineate how our approach differs from the aforementioned studies. Our modelling approach offers significant advantages by being highly systematic and minimising human intervention during the feature selection and model development phases. Initially, a comprehensive list of approximately 800 exploratory features is compiled, followed by the computation of the loss reduction attributed to each feature. Concurrently, a suite of statistical tests and analyses based on probability theory and information theory are conducted. These two independent stages effectively identify the most relevant features for the target variable, in this case, the values of T_c. These salient features are then used to generate additional features. A default method employed is a brute force approach, which requires no domain knowledge, although one can opt for manual intervention at this stage to guide feature engineering. The most relevant and statistically significant features, along with the newly engineered features, are then evaluated for multicollinearity.

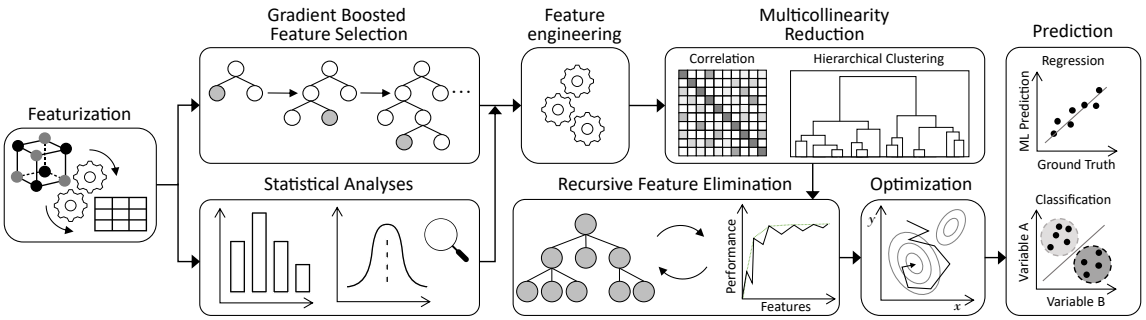


Fig. 3.2 Overview of our operational workflow. Adapted with permission from reference 123. Copyright 2023 the authors. Published by AIP Publishing under a Creative Commons CC BY License.

The initial step in addressing multicollinearity involves eliminating highly correlated features based on a predefined correlation threshold. This step is followed by hierarchical clustering analysis to group similar features. A linkage threshold is set, allowing the algorithm to automatically select one feature from each cluster to represent that cluster. The underlying rationale is that similar information can be derived from a single representative feature within a given cluster group, thereby streamlining the feature space without losing critical information. Subsequently, recursive feature elimination is performed, where a greedy search method recursively prunes features until the desired number of features is reached, or no model deterioration is observed. At this stage, permutation importance analysis is also carried out, which involves randomly shuffling the values of a single feature to observe the impact on performance metrics.

This meticulous process leads to a refined subset of features used to perform Bayesian optimisation. In this study, for instance, 36 features were systematically selected from an initial set of approximately 800 features. The optimisation stage autonomously identifies the most effective model architecture using solely the training set, without the need for human intervention throughout the process. Once the final predictive model is optimised, it is evaluated using the test set — marking the first and only time this data set is used. This rigorous approach ensures that our model is both robust and effective, leveraging systematic methodologies to enhance predictive accuracy and reliability, while eliminating potential human bias.

3.1.3 Regression Analysis of Curie Temperatures

Overall Performance Evaluation

We present the results of the regression analysis of T_c values that employed Dataset 1, which encompasses ca. 11,000 chemical compounds, among which ca. 7,200 represent distinct chemical compositions. The data set was segmented according to a train-to-test split ratio of $4:1$ through random splitting. For duplicate chemical compositions, the median values of T_c were calculated. The decision to adopt the median as the measure of central tendency was motivated by its reduced susceptibility to outliers relative to the mean. The regression was performed by a Bayesian-optimised gradient-boosting algorithm using 36 features, derived exclusively from their chemical compositions. These features were selected from a pool of more than 800 total features through our GBFS workflow.

Figure 3.3 shows the resulting model performance and error distribution on the test set. The blue dot-dash line represents the line of best fit, established through the Ordinary Least Squares (OLS) method; it depicts the relationship between T_c values reported in the literature and our ML-based predictions. Two distinct regression analyses were conducted: the first without the inclusion of MEGNet embeddings (as shown in Figure 3.3 (a)) and the second incorporating these embeddings (depicted in Figure 3.3 (c)). This methodological differentiation arises from the challenges associated with interpreting the physical contributions of the embeddings to the ultimate predictions. Consequently, this strategy enables a comparative examination of how feature interactions influence the prediction of T_c values.

In Figure 3.3 (a), the linear fit exhibits a gradient of 0.9 and a y-intercept of 21.1 K, rounded to one decimal place. The y-intercept suggests a minor systematic bias for lower values of T_c, while the gradient demonstrates a nearly perfect alignment of our predictions to the literature data, with a small underestimation at large temperatures within the range considered herein. The high efficacy of this linear fit is corroborated by an R^2 value of 0.93, which signifies a high correlation between predictions and the ground truth. Moreover, MAE and RMSE values of 38.8 K and 72.2 K were realised, respectively. The higher RMSE is attributable to the increased penalisation of predictions that deviate significantly from their true values. The distribution of absolute errors is depicted in Figure 3.3 (b). A ten-fold cross-validation process was undertaken, affording an R^2 of (0.92 ± 0.01), an MAE of (40.8 ± 1.9) K, and an RMSE of (80.0 ± 5.0) K. Observations indicate comparable performance upon integrating

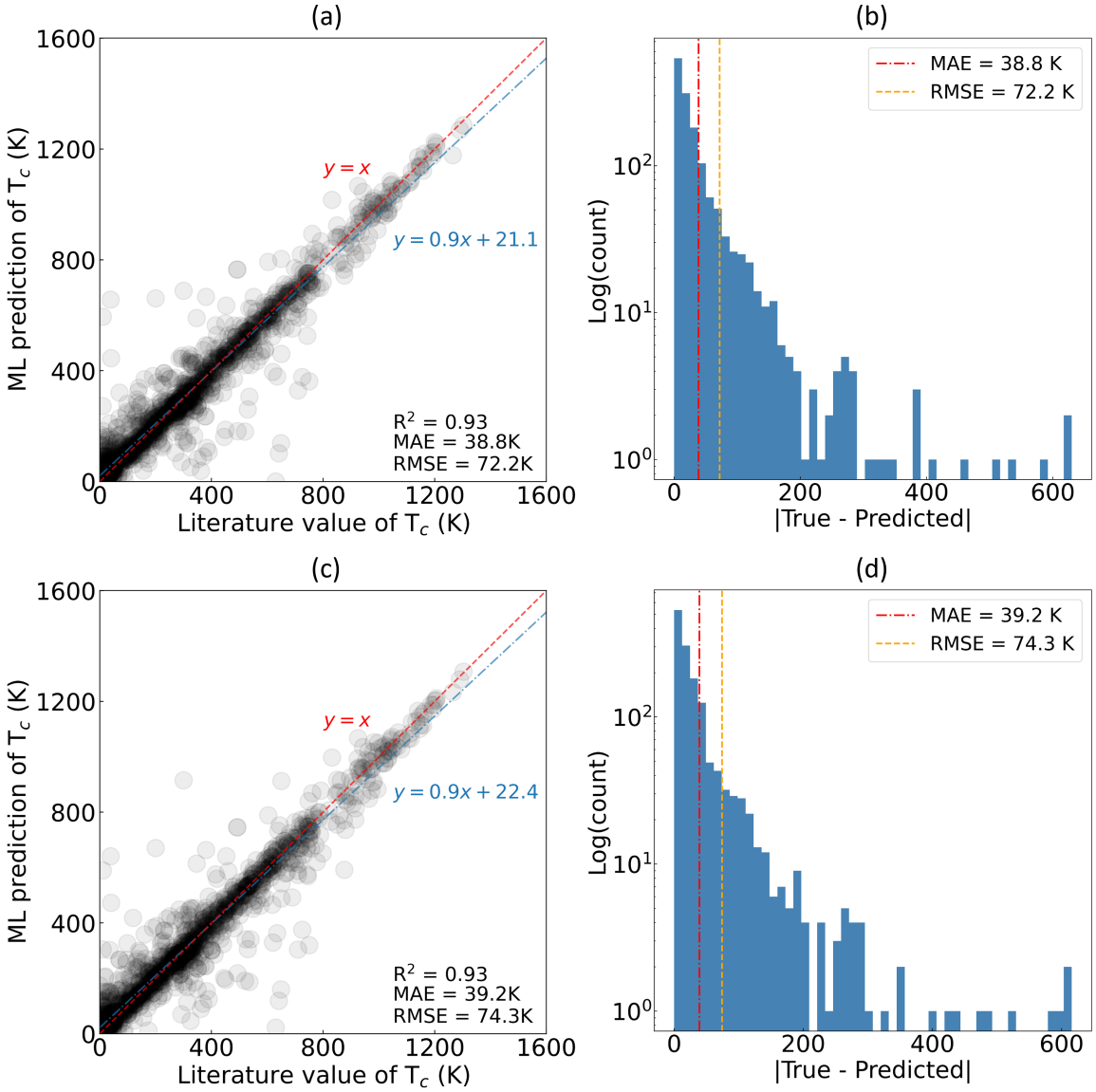


Fig. 3.3 Test set results of the ML-based predictions of T_c against the ground truth for the regression model that has been trained and Bayesian-optimised on the final subset of features selected by our GBFS workflow, where (a) excludes and (c) includes the MEGNet embeddings. The dashed red line represents the hypothetical case, where our ML-based predictions would equal the literature values (ground truth). The blue dot-dash line is a linear fit generated using the OLS method. (b) and (d) show the corresponding distribution of absolute errors for (a) and (c), respectively, where the dashed line (−.) in red indicates the MAE and the dashed line (−−) in orange indicates the RMSE.

MEGNet embeddings, as illustrated in Figure 3.3 (c) and the corresponding error distribution in (d). This underscores their effectiveness in facilitating transfer learning.

Notably, our modelling results outperformed the state-of-the-art report by Nelson et al. [131] (test set R^2 of 0.87 and MAE of 57 K with 129 features, and a cross-validation R^2 of 0.81) and Belot et al. [132] (a cross-validation MAE of 71 K with a standard deviation of 2.3 K with 85 features). These observations can be attributed to two primary factors. Firstly, our study involved the aggregation and analysis of an extensive data set, constituting the largest compilation of T_c data available in the literature. The substantial size of our data set enhances our capacity to discern statistically significant relationships between exploratory features and the target variable. Secondly, our implementation of the GBFS workflow allowed us to minimise feature redundancy and maximise feature relevance to the target variable. The efficiency of our modelling strategy enabled us to surpass the performance of other studies by employing a significantly lower number of input features, and notably, without resorting to regularisation techniques during model optimisation. The results on both the test set and the cross-validation process further affirm the generalisability of our model without over-fitting to the training set.

Our model performs with particularly high efficacy where $T_c \lesssim 500$, which stands to reason given the predominant concentration of data within this temperature range. However, the number of data points diminishes rapidly with increasing temperature, with the highest T_c of 1,388 K being recorded for the elemental metal, cobalt. This diminishing data density at higher temperatures contributes to the observed discrepancies between the ML predictions and the ground truth in this temperature range, as depicted by the lines of best fit (blue) in Figures 3.3 (a) and (c), which sit below the diagonal lines in red. The population distribution plots for both the training and test sets are illustrated in Figure 3.4. Nevertheless, our ML modelling approach has exhibited precise predictions of T_c despite the absence of any 3-D structural information. Moreover, our approach does not involve specific treatments to accommodate different crystal forms of the same chemical compound, a phenomenon known as polymorphism. Instead, it relies on capturing the median values of T_c for each chemical composition. Our approach strictly trains the model to be agnostic to crystalline polymorphs. Turning our attention back to the error distribution plots in Figures 3.3 (b) and (d), which have a logarithmic scale on the y-axis, it is evident that the majority of the predictions exhibit errors below ca. 60 K. Upon closer inspection of the chemical compositions in the test set, ca. 82% have an absolute error below 60 K, ca. 65% have an absolute error below 30 K, and ca. 32% have an absolute error below 10 K.

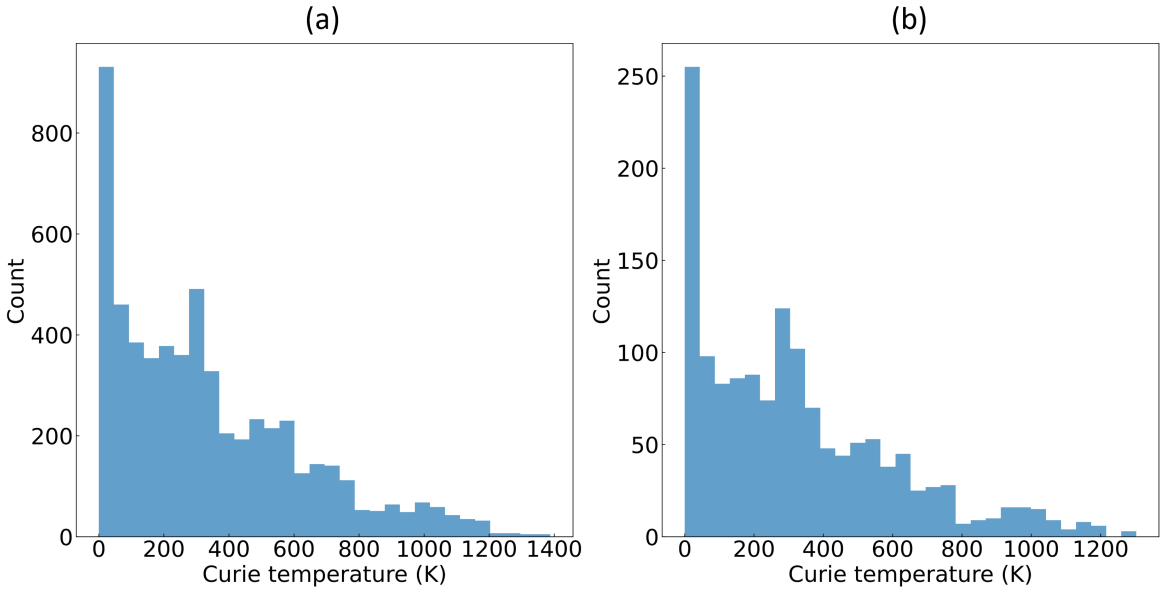


Fig. 3.4 Distribution of Curie temperature (T_c) values in (a) the training set and (b) the test set, sourced from Dataset 1.

We recognise the importance of assessing how model performance varies with data set size. This evaluation helps to ascertain whether the expanded data set used in this study offers any benefit to the prediction accuracy or if the benefits have plateaued which would indicate no significant advantage in using such a data set. Our analysis evaluated the MAE, RMSE, and R^2 values for increasingly larger subsets of the original training set, where the data were randomly sampled. The results shown in Figure 3.5 reveal a continuous improvement as the training set size increases, yet without a definitive convergence in these metrics. Therefore, the data set curated for this study not only enhances model performance but may also hold the potential to offer valuable insights to the wider research community.

It is also essential to clarify that our modelling rationale does not overlook the significant impact that structural information can have in enhancing prediction accuracy. Indeed, integrating such data would undoubtedly improve our ML model by introducing additional variables critical to the predictive outcomes. This is because understanding the arrangement of atoms within a crystal lattice provides insights into the interactions of magnetic moments. For example, body-centred cubic structures, such as those observed in α-iron or ferrite, facilitate specific types of magnetic interactions that are conducive to higher Curie temperatures. Nevertheless, our decision to restrict the feature space exclusively to variables derived from chemical compositions is grounded in two key considerations. Firstly, there is a substantial scarcity of accessible 3-D

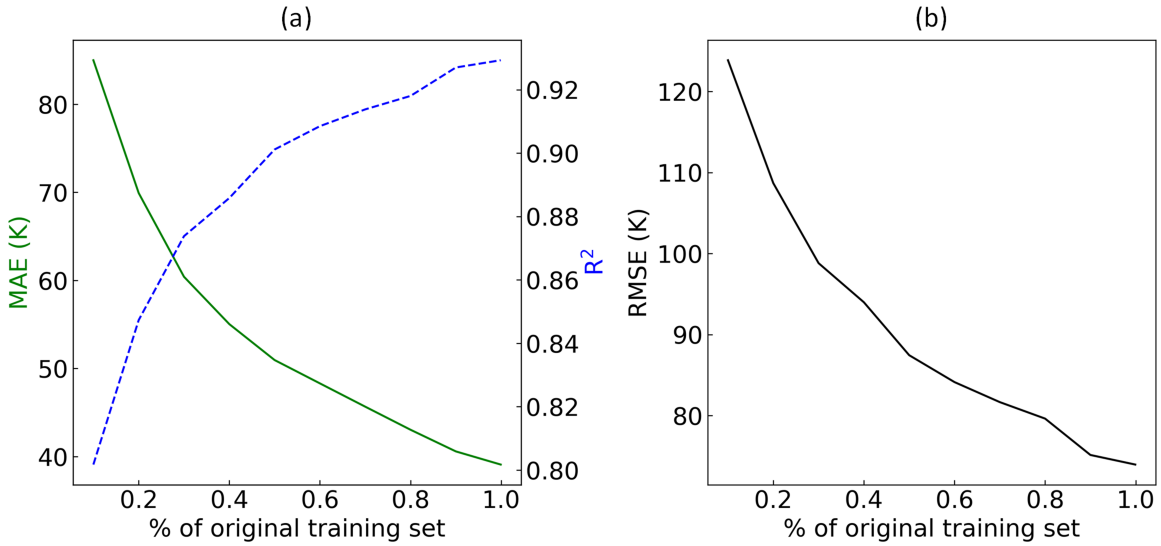


Fig. 3.5 Evaluation of MAE, RMSE, and R^2 conducted by training our ML model on increasingly larger subsets of the original training set, where the data were randomly sampled.

crystallographic information and it is cost-prohibitive to aggregate such information, which impedes the inclusion of structural data. Secondly, constraining the feature set in this manner streamlines the predictive modelling process. The inclusion of structural information would necessitate the exclusion of a substantial portion of the data used in this study due to its unavailability. Additionally, should future predictions be made, the featurisation stage, which involves generating structural features, would be constrained by the limited availability of necessary crystallographic data, thereby limiting the practical utility of the ML model.

Feature Interpretation

Table 3.1 summarises some of the top 10 salient features that contributed to the aforementioned outcomes in our regression analysis. We now seek to rationalise their significant role. As denoted by the realised total loss reduction, the most influential feature pertains to the mean ground-state magnetic moment of elemental solids for atoms within the chemical composition. This feature is categorised under elemental property statistics. The second most influential feature is the presence of cobalt in the chemical composition. Additional features exhibiting significant relevance with the target variable encompass a range of statistical measures pertaining to the number of *d*-valence electrons or vacant *d*-valence orbitals, the volume of the elemental solid, the periodic table's group number, electronegativity, instances where the element with the

Table 3.1 A list of features identified to have the most relevance in the prediction of T_c values.

No.	Feature Description
1	Ground-state magnetic moment of elemental solids for atoms within a given chemical composition
2	Presence of cobalt
3	Number of valence electrons in the *d*-orbitals of elements within a given chemical composition
4	Volume of elemental solids for atoms within a given chemical composition
5	Periodic table column number
6	Electronegativity
7	Number of unfilled electrons in the *d*-orbitals of the elements within a given chemical composition
8	HOMO and LUMO energies and their associated chemical elements
9	Presence of manganese
10	MEGNet embeddings

lowest energy molecular orbital (LUMO) is iron, the inclusion of manganese within the composition, the coefficient of linear thermal expansion, thermal conductivity, atomic radius, and the energy of the highest occupied molecular orbital (HOMO).

The feature selected with the highest relevance was the mean ground-state magnetic moment of elemental solids for atoms within the chemical composition; this was anticipated, considering its direct material association with the measure of magnetic strength or the tendency of a magnetic moment to align with a magnetic field. The selection of the ground-state magnetic moment as a key feature, while seemingly intuitive, is substantiated by both theoretical and empirical rationales. The ground-state magnetic moment is a fundamental property that defines the magnetic behaviour of a material at absolute zero, where the material is in its lowest energy state. Specifically, it quantifies the total magnetic dipole moment in the most stable configuration of the material. This quantification reflects the nature and intensity of the magnetic spins within the material, which are crucial for determining its ferromagnetic characteristics. Thus, the mean ground-state magnetic moment feature is directly linked to T_c, as it embodies the strength of the magnetic interactions within the material. A higher magnetic moment typically indicates stronger magnetic interactions, which are pivotal because they dictate the temperature at which thermal energy disrupts the magnetic ordering, thereby influencing T_c.

Furthermore, the magnitude and presence of the ground-state magnetic moment act as markers of ferromagnetic properties. Ferromagnetic materials, distinguished by a non-zero ground-state magnetic moment, manifest spontaneous magnetic order below T_c due to the alignment of magnetic moments. Therefore, quantifying these moments offers a direct measure of the extent and nature of ferromagnetism in the material, positioning the mean ground-state magnetic moment as a critical metric for T_c.

Empirical research consistently demonstrates a correlation between T_c and the nature and strength of magnetic interactions, as encapsulated by the magnetic moment. This relationship is exploited in predictive modelling, wherein the ground-state magnetic moment serves as a foundational component for estimating T_c. For instance, an empirical study conducted by Fecher et al. [143] has observed that T_c values of Co_2-based Heusler compounds display a discernible linear correlation with the magnetic moment. Fecher et al. extrapolated such a linear pattern to ascertain that T_c can surpass 1,000 K in Co_2-based Heusler compounds characterised by a magnetic moment of $6\mu_B$ and 30 valence electrons per unit cell (e.g., Co_2FeSi). This observation additionally underscores the dependence of T_c on both the number of valence electrons and the presence of elemental cobalt in the chemical composition (cf. the second and third-ranked selected features). More broadly, the relevance of such a feature generally applies to alloys that adhere to the Slater-Pauling curve [144, 145]. It is a well-established fact that the magnetisation of $3d$ transition metal substitutional alloys, as a function of the valence electron number per atom, forms the Slater-Pauling curve. Similarly, T_c values of these alloys display a systematic pattern relative to the number of valence electrons, a phenomenon that has also been substantiated through first-principles calculations [146]. Therefore, it is evident that the magnetic moment is correlated with the Slater-Pauling curve, which is discussed in more detail later in this section.

The prominence of cobalt-based attributes as the second most salient feature in predicting T_c aligns with expectations. As previously mentioned, cobalt as an elemental metal exhibits the highest recorded T_c of 1,388 K. Materials with high T_c values predominantly feature cobalt within their chemical composition. As we will see later, subsequent blind-test analysis further reveals that the majority of chemical compositions with $T_c \gtrsim 600$ K include both cobalt and iron, indicating the significance of cobalt in achieving elevated T_c values.

The third most salient type of feature is associated with d-valence electrons or orbitals, for which a clear rationale exists. Slater [144] and Pauling [145] established that the magnetic moments of $3d$ elements and their binary alloys could be characterised by the mean number of valence electrons per atom, offering a simple explanation of the relationship between the number of valence electrons and magnetic moment in ferromagnetic alloys. Specifically, Co_2-based Heusler compounds adhere to the Slater-Pauling rule, which predicts that the total magnetic moment scales linearly with the number of valence electrons [147, 148, 143]. Co_2-based compounds are situated on the localised part of the Slater-Pauling curve, which is indicative of an increasing magnetic moment with an increasing number of valence electrons [149]. It is also

established that in quaternary half-metallic ferromagnetic materials, incorporating a transition metal with $4d$ electrons in conjunction with iron or manganese can elevate of the T_c value. The variation of T_c as a function of the number of valence electrons can be understood through the inter-atomic exchange interaction parameter [150]. These findings elucidate one of several interactions among diverse features as discerned through the GBFS workflow. It accentuates the predictive significance of valence electron characteristics in ascertaining magnetic properties, thereby underscoring the intricacy of feature interrelations within the predictive model.

The GBFS workflow identified additional chemical composition-based features that are also of significance. For instance, the average deviation in the periodic table column positions among elements within the chemical composition demonstrates a notable correlation of ca. -0.45 with the target variable, T_c. This negative correlation indicates that materials with smaller average deviations in the periodic table column positions among their constituent elements tend to be associated with higher T_c values. This characteristic is prevalent in materials comprising metals such as iron, nickel, cobalt, and rare-earth metals. Likewise, characteristics associated with the periodic table's group number, atomic radius, and Mendeleev number emerge as salient features identified by the GBFS workflow. These attributes are closely tied to an element's specific location within the periodic table, underscoring their relevance in the predictive analysis.

Additionally, the inclusion of manganese within the chemical composition demonstrates significant pertinence. While elemental manganese, a metal, does not exhibit ferromagnetism in its pure form, manganese alloys can exist in the form of a Heusler crystal structure, wherein manganese possesses a magnetic moment. More generally, the marked sensitivity of T_c values observed in Mn-containing magnetic alloys is well-documented and can be explained via the empirical Castelliz-Kanomata curves [151, 152]. These curves, validated across various Heusler alloys [153, 154], highlight that the magnetic interactions in Mn depend on structural parameters, such as the Mn-Mn nearest neighbour distance. Furthermore, empirical evidence demonstrates that pressure, and hence the volume, significantly influences T_c values. These findings suggest that identifying certain features in our ML model can assist in distinguishing materials with high T_c values or materials that deviate from the elementary framework of ferromagnetism.

Another noteworthy feature is the mean volume of the elemental solid among the elements in the chemical composition. The exchange energy (E_{ex}) is a continuum

description of the quantum mechanical exchange interaction and is given by:

$$E_{ex} = \int_V A \nabla \left(\frac{\mathbf{M}}{\mathrm{M_s}} \right)^2 dV, \tag{3.1}$$

where A is the material-dependent exchange stiffness, $\mathbf{M}$ is the magnetisation, M_s is the saturation magnetisation and the integral is over the volume of the sample [114]. Equation 3.1 shows that the exchange energy is explicitly linked to the integral over the volume of the sample. This equation provides a lucid and simple illustration of how volume and shape may affect the exchange energy and coupling and, consequently, influence the prediction of a T_c value.

The final category of features to be addressed involves band-gap-related attributes, such as the energies of the HOMO and LUMO or their associated chemical elements, which necessitates further explanation in the context of predicting T_c. For instance, the HOMO signifies the highest energy level that contains electrons, analogous to the valence band in Band theory, and it is directly associated with band-gap energy. Several possible theories could explain the relationship between these features and T_c. For example, the influence of the band gap on T_c has been extensively studied by Coey et al. [155] and Pan et al. [156]. In particular, the T_c values of diluted magnetic oxides (DMOs) are quantitatively explained, based on a bound magnetic polaron mechanism, via the equation:

$$T_c = [(S+1)s^2 x\delta/3]^{1/2} J_{sd} f_O \left(\frac{r_c^{eff}}{r_O} \right)^3 / k_B, \tag{3.2}$$

where S is the localised core spin, s is the donor electron spin, x and δ are the concentrations of magnetic cations and donors, J_{sd} is the s-d exchange parameter (which is related to band gap), f_O is the oxygen packing fraction for the oxide, and r_c^{eff} is the effective cation radius, r_O is the radius for the oxide, and k_B is Bolzmann's constant [155, 156]. The donors form bound magnetic polarons, leading to coupling among the $3d$ moments of the ions within their orbits. When the radius of its orbital is sufficiently large, the ferromagnetic exchange coupling is established from the overlap between a hydrogenic electron and the cations within its orbit. This interaction is influenced by the parameter J_{sd}. Equation 3.2 illustrates the dependency of T_c on the band gap via the J_{sd} parameter, in addition to other crucial factors such as doping and donor concentrations. Indeed, the results of Pan et al. [156] demonstrated that one can enhance T_c values in DMO materials by manipulating their band gap. This exemplification helps to rationalise why the average number of unfilled d orbitals among elements in the chemical composition is also an important feature. Furthermore, it

reinforces the identification of d-valence electrons as among the most pertinent features for predicting T_c, thereby offering additional validation for their relevance.

It is essential to acknowledge that, while the band-gap-related features have a relatively modest level of significance compared to other features discussed in this section, their presence is not negligible. Chemical compositions containing oxygen account for approximately 13% of the data set used to generate the results shown in Figure 3.3. Although they represent a minority, these features appear to help distinguish between metal oxides and purely metal-based compositions. The ability to identify specific features that differentiate minority classes is of importance. The recognition of such attributes enhances the capability of the model to distinguish between different classes of chemical materials. Moreover, the impact of these features on reducing the loss function during model training further supports their relevance.

The HOMO energy also exhibits a degree of correlation with other selected features, such as the average deviation of electronegativity among elements in the chemical composition, and various features derived from p-valence electrons or p-orbitals. Such correlations are particularly pronounced for chemical elements that belong to groups 13-18 of the periodic table, notably within the p-block. This region of the table is characterised by elements that possess high electronegativity values. The pairing of a metal (characterised by low electronegativity) with a non-metal element leads to a large difference in orbital energy, a phenomenon which is accentuated as the electronegativity gap between the paired elements widens. Therefore, it is not surprising that numerous ionic compounds have been identified as being ferromagnetic, exhibiting conductivity levels that are typical of semiconductors. This category of materials encompasses chalcogenides, and halides, or combinations of these groups in the periodic table. In such materials, ions like chromium and europium contribute to the formation of permanent dipole moments. Notably, many rare-earth metals within the lanthanide series exhibit spontaneous magnetisation below specific temperatures. Moreover, ferromagnetic ordering is often observed in ionic compounds with spinel crystal structures, typified by metal oxides with the general composition AB_2O_4.

The final feature to be discussed refers to the MEGNet element embeddings. Upon integrating the MEGNet embeddings, it was noted that 6 of the top 20 most salient features were substituted by these MEGNet embeddings. These learned element embeddings on graph-neutral-network models encode chemical trends in the periodic table. While interpreting individual embeddings can be challenging, previous studies have demonstrated their utility in transfer learning [34, 123, 125]. Specifically, these embeddings can be transfer-learned from a material-property model trained on a larger

data set to enhance property models with smaller data sets. In this study, we leveraged the acquired embeddings to enhance the predictions of T_c values for ferromagnetic materials.

Gradient Boosted Feature Selection

Our GBFS workflow selected a final subset of 36 features from over 800 exploratory features. Here, we present the results associated with this selection process. Initially, gradient boosting decision trees (GBDTs) were recursively trained on an increasing subset of features until R^2, MAE, and RMSE converged, whereby the features were originally ranked based on the total loss reduction achieved during the training process. The evolution of these performance metrics over the course of this selection process is illustrated in Figure 3.6. Evaluation results are shown for both the training set and the validation set. The performance metrics for both sets reached a plateau before ca. 40 features had been included. As expected, a comparatively lower performance on the out-of-sample validation set was observed. Additionally, we noticed a sudden decline in the model performance on the validation set beyond the inclusion of the fifth feature, followed by a subsequent recovery in performance. This phenomenon is not unusual, given that we did not account for multicollinearity at this stage of the workflow, even though it may be present among the exploratory features. In the presence of multicollinearity effects, the total loss reduction is distributed evenly across correlated features, thereby masking the true relevance of individual features to the target variable.

Feature Analyses & Feature Engineering

We sought to understand the causal relationship between an exploratory feature and the target variable by concurrently employing hypothesis-based testing methods of a bivariate form. For instance, a comparison of means was conducted using the F-test in a one-way ANOVA. This process involves a correlation analysis between two continuous features, where the correlation coefficient (R) was translated into a regression F-statistic. These hypothesis-driven testing methods were used to draw statistical inferences. The statistical significance of an exploratory feature was determined by evaluating the test statistics, which were generated to assess the hypothesis that an association exists between two features. Additionally, mutual information (MI) analysis was performed. The concept of MI was employed to quantify the level of dependency between two features, measuring the amount of information, or entropy, gained for a feature through the observation of another. For a pair of features, MI assesses the disparity between

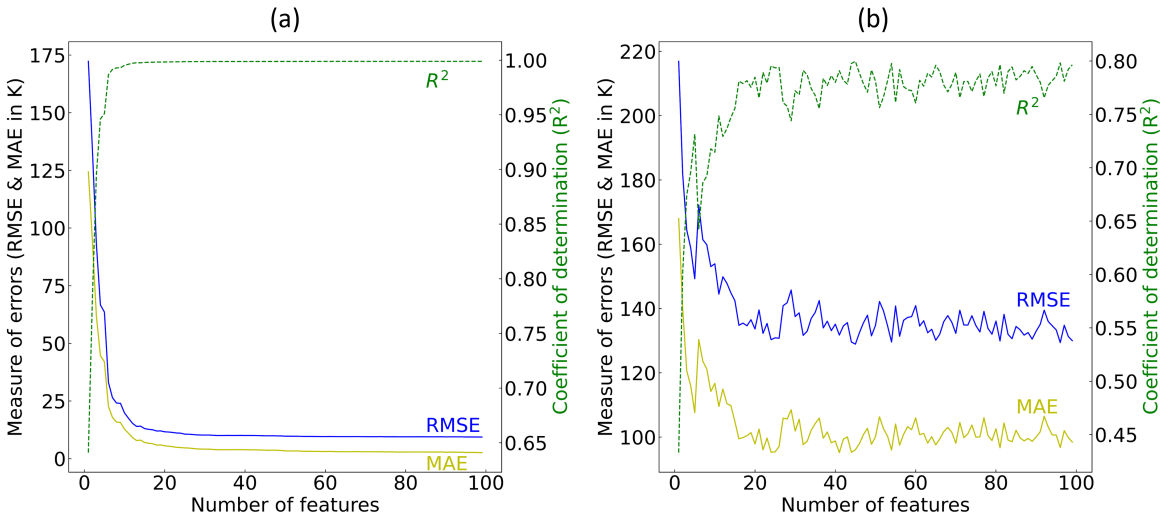


Fig. 3.6 Gradient boosting feature selection results of the regression analysis of T_c value prediction. Model performance of GBDTs on (a) the training set and (b) the validation set, where regression models were trained recursively with an increasing subset of features, beginning from the most relevant feature based on the realised total loss reduction.

their joint distribution and the product of their marginal distributions, with a higher MI value indicating a greater dependency between the two features. We adopted an MI estimator based on entropy estimations that are derived from k-nearest neighbour distances.

When assessing the linear association of each continuous exploratory feature with the target variable through a normalised F-statistic for relative comparison, we identified that the feature demonstrating the highest linear association with the target variable was the mean ground-state magnetic moment of elemental solids for atoms within a given chemical composition (estimated using Magpie data). The second most strongly associated feature was the fraction of transition metals, followed by the mode of the magnetic moment of elemental solids for atoms within a given chemical composition (also estimated using Magpie data). Their normalised F-statistics were 0.75 and 0.68, respectively. Other features exhibited normalised F-statistics below ca. 0.6. Nevertheless, these included some with notable loss reduction, such as the presence or fractional abundance of cobalt in a chemical composition, the mean volume of elemental solids, the periodic table column or group number of the constituent elements, and the number of *d*-valence electrons.

In parallel, the MI analysis indicated that the largest amount of entropy gain was realised when considering the mean ground-state magnetic moment of elemental

solids for atoms within a given chemical composition. Other notable features with a normalised MI score above 0.8 include statistical measures associated with the fraction of transition metals, the periodic table column or group number of the constituent elements, the number of *d*-valence electrons or unfilled *d*-orbitals, electronegativity, and the Mendeleev number. Since the MI analysis incorporates the k-nearest neighbours method, these results essentially suggest that more accurate predictions of T_c can be achieved by considering statistical measures that pertain to the ground-state magnetic moment of elemental solids and the transition metals (i.e., *d*-block elements). It is noteworthy that the estimation of MI involves assessing the probability-density distribution and marginal distributions of the two variables of interest. However, estimations of these distributions become increasingly challenging in higher-dimensional data, given the limited number of samples with respect to the number of dimensions. This limitation often leads to substantial variations in probability; as a result, the estimated information gain in MI analysis may suffer from the high-dimensionality nature of the data set or an inadequate sample density with respect to the dimension of the feature space. The features identified through the GBFS workflow and statistical analyses were used to engineer new features via the brute-force method. This process resulted in an additional 56 features, leading to a total number of 116 features that formed the preliminary subset of features for the regression analysis.

Multicollinearity Reduction, Permutation Analysis & Recursive Feature Elimination

In the next phase of the GBFS workflow, we address multicollinearity reduction within the data set, assess the permutation importance of the selected features, and conduct recursive feature elimination to ascertain the final subset of features that will go forward for Bayesian optimisation of the final predictive ML model.

To mitigate the effects of multicollinearity in the data set, features with a correlation coefficient of 0.8 or higher were systematically eliminated. This process reduced the data set to a subset containing 67 features. The next remediation of multicollinearity effects involved employing a hierarchical cluster analysis, using Spearman rank-order correlation with Ward's linkage distance threshold of 1.5 units. This step led to the retention of 38 features, as only one feature from each cluster was chosen. The optimal distance threshold was determined using the Elbow method with a step size of 0.5 units. The corresponding dendrogram in Figure 3.7 (a) depicts the hierarchical agglomerative clustering of features with respect to Ward's linkage distance, whereby clusters form

as one ascends the dendrogram while the results of the ten-fold permutation feature-importance analysis are shown in Figure 3.7 (b).

Permutation feature importance is quantified as the reduction in model performance when a single feature used in the construction of the model is randomly shuffled. This process disrupts the association between the feature and the target, making any reduction in model performance an indicator of how reliant the model is on that specific feature. The ten-fold feature permutation analysis suggests that the most important feature is the mean ground-state magnetic moment of elemental solids for atoms within a given chemical composition, as estimated using Magpie data, followed by the presence of cobalt in the chemical composition and the mean *d*-valence electrons. These results are consistent with the scientific findings of the statistical analyses conducted independently.

The optimal subset of the remaining features was determined by eliminating further features through ten-fold recursive feature elimination, employing negative RMSE values as the performance metric; see Figure 3.7 (c) for the results. This process led to the identification of the final subset of 36 features. It is helpful to remember that these short-listed features were chosen from an initial pool of ca. 800 original features as well as 56 engineered features, which demonstrates their highest relevance to the target variable without any prior knowledge of the scientific domain.

Model Optimisation & SHAP Analysis

A two-step optimisation process was followed to determine the architecture of the final classifier. The hyperparameters of the model were optimised using a combination of grid search and Bayesian optimisation using Gaussian processes. An initial hyperparameter tuning process was performed by scanning the hyperparameter space using the grid-search method. This procedure subsequently identified the region in which Bayesian optimisation was to be applied. Such an optimisation strategy proves to be particularly effective for an objective function: (i) that has no closed form; (ii) expensive to evaluate; and (iii) whose evaluations yield noisy responses. The partial dependence and evaluation plots derived from the Bayesian optimisation results are presented in Figure 3.8, while the total loss reduction (i.e., the feature-relevance ranking) realised by the top 20 features in the Bayesian-optimised model is illustrated in Figure 3.9.

An independent feature analysis was conducted using the SHapley Additive exPlanations (SHAP) framework [157], a game theoretic approach explaining the output of an ML model. Figure 3.10 (a) displays the plot of average contributions (i.e., the mean absolute SHAP value) of the ten features identified as having the most significant

contributions to the model output. The accompanying beeswarm plot in Figure 3.10 (b) illustrates the impact of these features on the model output by plotting each instance as a single data point together with the SHAP value on the x-axis. These findings align with the features identified by the GBFS workflow (see Figure 3.9), providing additional validation for the effectiveness of our modelling approach. Once again, we observe that the mean ground-state magnetic moment of elemental solids for atoms within a given chemical composition, the presence of cobalt, the mean d-valence electrons, and the mean volume of elemental solids, are among those identified as having the most significant contributions to the model prediction of T_c values. A substantial overlap is evident between the two sets of results.

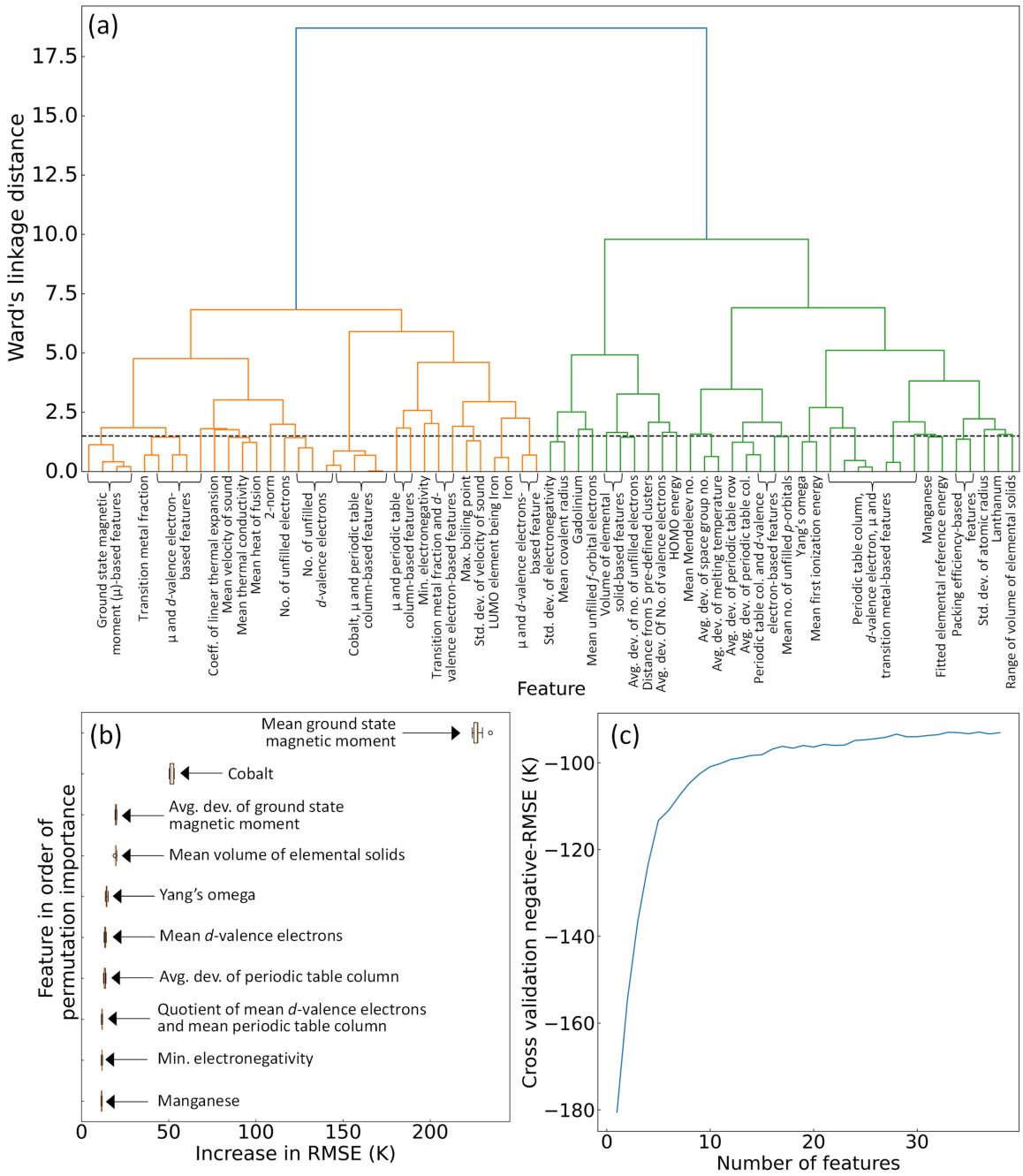


Fig. 3.7 (a) Multicollinearity reduction - the dendrogram of the hierarchical agglomerative clustering using the remaining 67 features after performing the correlation analysis. The dashed horizontal line in black represents the distance threshold of 1.5 units of Ward's linkage distance. (b) The permutation feature-importance box plot for the regression of predicted T_c values. (c) The ten-fold recursive feature elimination result using the negative RMSE as the performance metric.

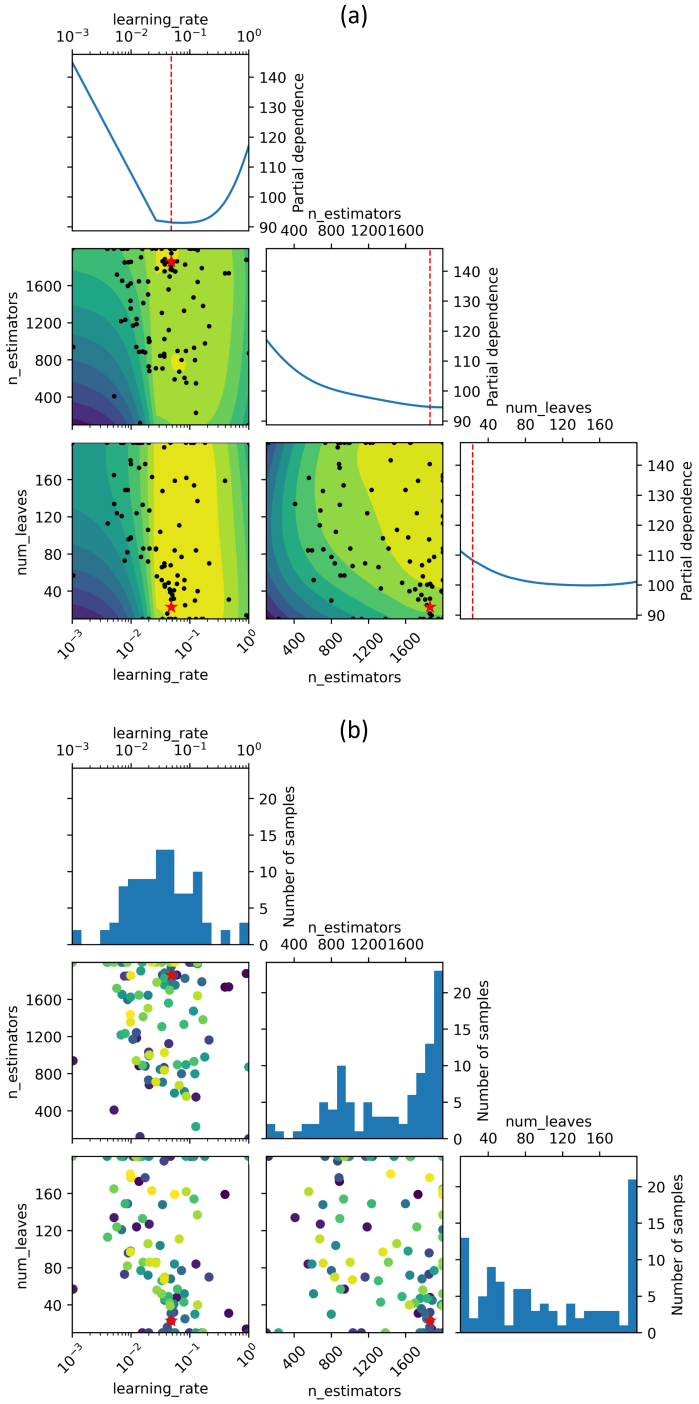


Fig. 3.8 Bayesian optimisation result of the final regression model using the training data, where (a) is the partial dependence plot and (b) is the evaluation plot. The red stars indicate the values of the hyperparameters that achieved the lowest value of the objective function. The approximate position of the objective minimum is indicated by the dashed vertical lines in red.

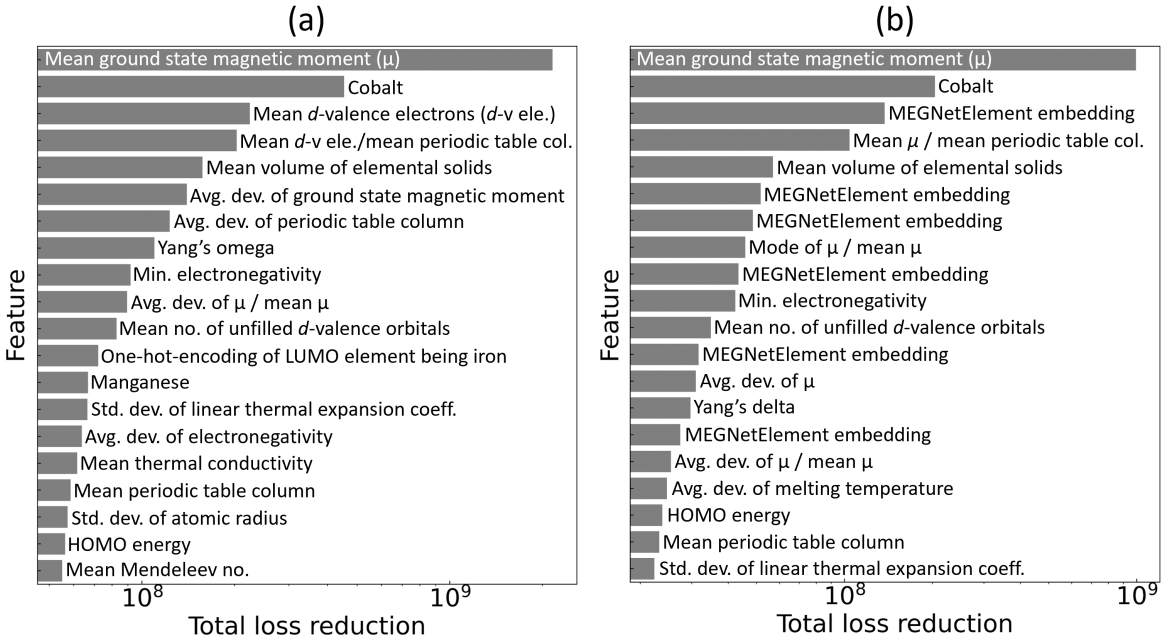


Fig. 3.9 Feature relevance plot - top 20 features selected for the regression analysis of T_c, where (a) is without and (b) is with MEGNet element embeddings, along with the realised total loss reduction (i.e., the relevance score).

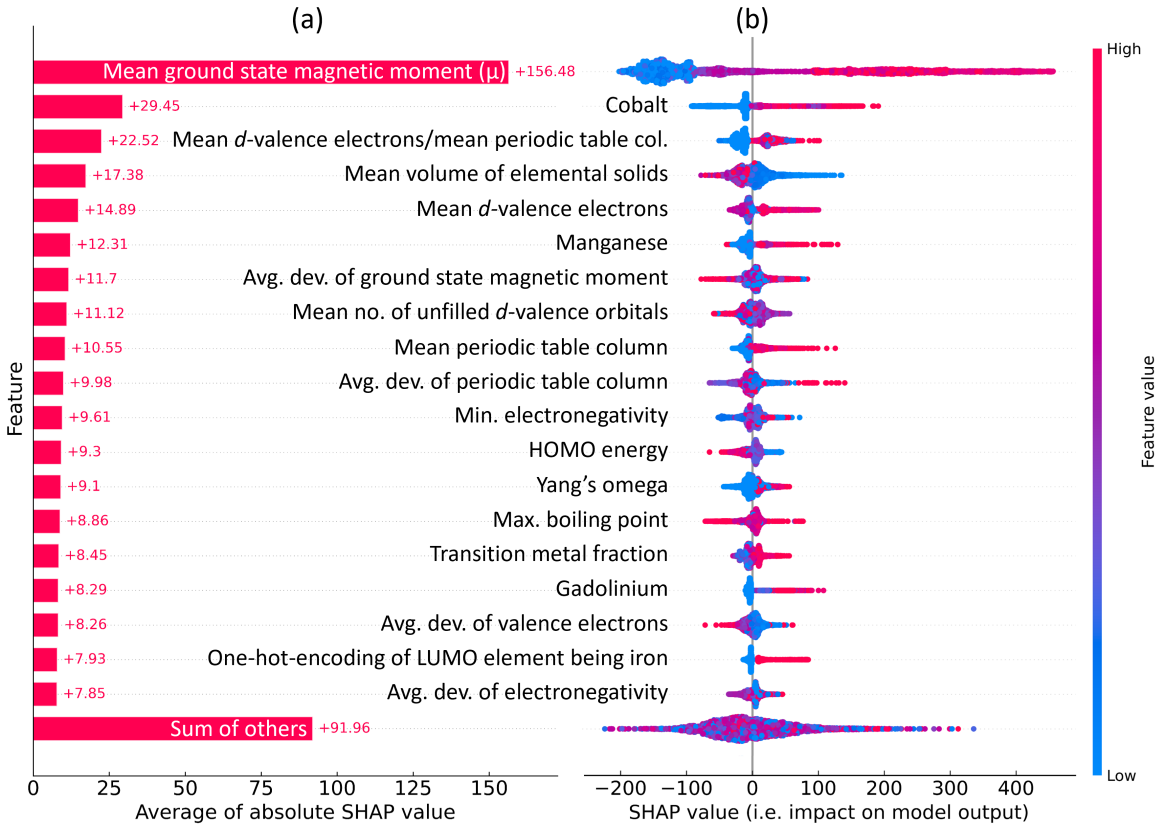


Fig. 3.10 SHAP-based result - (a) the average contribution (i.e., the mean absolute SHAP value) of the ten features identified as having the greatest contributions to the model output. A positive SHAP value indicates a positive contribution to the regression of T_c. The colour scheme corresponds to the original feature value and the broadening shows the density of instances (cf. density plot). (b) The beeswarm plot illustrates the impact of these features on the model output by plotting each instance as a single data point together with the SHAP value on the x-axis, where the y-axis is consistent with (a).

3.1.4 Blind Test

The results discussed thus far present predictions of T_c values against literature values (cf. Figure 3.3) with promising statistical figures-of-merit. Nonetheless, it is important to validate these results by considering how these predictions fare across a diverse range of chemical materials rather than simply demonstrating their collective statistical quality in an anonymised form. Therefore, a comparative analysis was conducted against Dataset 2, i.e., experimental measurements from the inorganic materials database, AtomWork, which contains ca. 16,000 chemical compositions with T_c values, of which ca. 6,600 are unique compositions. We applied our ML model to chemical compositions previously unseen by the model. The predictions of T_c values were subsequently compared against the experimental measurements, with 90 examples summarised in Table 3.2. These examples were randomly chosen from instances where the predicted T_c value exceeded room temperature; more specifically, when $T_c \gtrsim 600$ K. This selection criterion is important, as materials exhibiting such T_c values can maintain thermally stable magnetic states or magnetisation for functional applications.

The aforementioned tendency of our model to underestimate experimental T_c values is apparent, with 64 out of 90 chemical compositions exhibiting a negative percentage difference from the mean T_c values. This underestimation becomes more evident at higher T_c values, as depicted by the best-fit line shown in Figure 3.3 (a). We attribute this anomalous pattern to the scarcity of data related to ferromagnetic materials with high T_c values within the training set, making it challenging for ML models to learn effectively within this temperature range. We anticipate that the availability of more data points in these temperature ranges would enhance the efficacy of our ML models.

Notwithstanding this modest negative bias in our ML model, it predicts T_c values well on an absolute scale. Thereby, the average absolute percentage difference from the mean is ca. 6.7%. Some of the lowest absolute percentage differences were observed for chemical compounds such as $MnCo_2Ga_{0.5}Sn_{0.5}$, $Nd_{1.9}Tb_{0.1}Re_{0.1}Fe_{11.9}Co_2B$, and $Nd_2Fe_{17}N_3$. Meanwhile, some of the highest absolute percentage differences were noted for chemical compounds such as $La_2Mn_3Co_{14}$, MnFeCoGe, $Li_{0.33}Cd_{0.28}Fe_{2.25}Si_{0.089}O_4$, and $Sm_2Cr_{1.2}Co_{12.8}$, whose absolute percentage differences from the mean exceeds 20%. The absolute difference of each predicted T_c value from its experimental measurement is colour-coded according to the classifications green ($0 - 10$ %), amber ($10 - 20$ %) or red (> 20 %). We observe that materials rich in cobalt and iron exhibit the highest T_c values. Among the 90 examples summarised in Table 3.2, iron and cobalt are present in the composition over 60 times. Other noteworthy chemical compositions include the elements, neodymium, samarium and manganese. The majority of the chemical

compositions involve a combination of these chemical elements, as well as their oxide forms.

Table 3.2: Curie Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of T_c Values (i.e., Model Output), Along With the Percentage Difference From the Mean Experimental T_c Values (Ground Truth)[a].

Chemical composition	Experimental T_c (K)				Pred. T_c (K)	% diff. from mean
	n	Min	Max	Mean		
$PrSmCo_{17}$	1	1200	1200	1200	1132	-5.6
$Fe_{0.5}CoAl_{0.5}$	1	1166	1166	1166	944	-19.0
Gd_2ZrCo_{16}	1	1165	1165	1165	1066	-8.5
Pr_2Co_{17}	2	1157	1158	1158	1178	1.7
$YSmZr_{0.6}Co_{16.4}$	1	1133	1133	1133	1084	-4.3
$Y_2Zr_{0.6}Co_{16.4}$	1	1118	1118	1118	1101	-1.5
$Er_2Fe_4Co_{13}$	3	1000	1200	1115	1088	-2.4
$FeCo_2Al_{0.5}Si_{0.5}$	2	1110	1114	1112	1092	-1.8
$CeSmZr_{0.6}Co_{16.4}$	1	1108	1108	1108	1087	-1.9
$Sm2Cr_{1.2}Co_{12.8}$	1	1093	1093	1093	854	-21.9
$YTiCo_{11}$	3	750	1500	1060	994	-6.2
Y_2ZrCo_{16}	1	1050	1050	1050	1017	-3.1
$LaMn_{0.7}Co_{12.3}$	1	1050	1050	1050	1094	4.2
$La_2Ta_{0.5}Co_{16.5}$	1	1032	1032	1032	1013	-1.8
$Sm_2Co_{14}B$	3	1025	1029	1027	1001	-2.5
$La_2Mn_3Co_{14}$	1	1010	1010	1010	751	-25.6
$Pr_2Co_{14}B$	9	955	994	984	1025	4.2
Fe_2CoGa	1	980	980	980	1121	14.4
$Fe_{0.8}Ga_{0.2}$	5	920	1013	961	770	-19.8
$La_2Co_{14}B$	2	955	960	958	979	2.2
$Li_{0.33}Cd_{0.28}Fe_{2.25}Si_{0.089}O_4$	1	943	943	943	735	-22.0
$Ce_2Co_{16}Ga$	2	930	950	940	967	2.9
$Cu_{0.94}Pr_2ZrCo_{15.1}$	1	934.7	934.7	934.7	874	-6.5
$Y_{0.31}Nd_{0.47}Hf_{0.11}Co_{5.33}$	2	786.3	1059	922.7	981	6.3
$Pr_{0.8}Zr_{0.2}Co_5$	1	915	915	915	891	-2.6
La_2TiCo_{16}	2	914	914	914	906	-0.8
$DyFe_{8.5}Ge_{3.5}$	1	910	910	910	865	-5.0
$Gd_2Cr_{1.76}Co_{15.24}$	1	905	905	905	929	2.6
$Cu_{0.2}Mn_{0.02}Fe_{1.9}Ni_{0.8}O_4$	1	900	900	900	724	-19.5
$YMo_{0.5}Fe_{9.2}Co_{2.3}N$	1	879	879	879	739	-15.9
$NdCo_5$	7	620	915	868	871	0.4

Continued on next page

Table 3.2: Curie Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of T_c Values (i.e., Model Output), Along With the Percentage Difference From the Mean Experimental T_c Values (Ground Truth)[a]. (Continued)

$Nd_{1.2}Gd_{0.8}Fe_{10}Co_4B$	1	852	852	852	822	-3.5
$PrGdFe_{10}Co_4B$	1	848	848	848	822	-3.1
$HoTiFe_5Co_6$	2	810	883	847	842	-0.5
$Y_{1.5}Nd_{1.5}Ti_{1.3}Fe_{21.7}Co_6$	1	840	840	840	666	-20.7
Fe_3Si	6	803	860	835	809	-3.1
$PrTi_{0.62}V_{1.14}Fe_{6.82}Co_{3.41}$	1	829	829	829	762	-8.0
$Sm_{0.78}Co_{4.44}Ga$	1	827.4	827.4	827.4	721	-12.9
$Sm_{0.5}Er_{0.5}TiFe_8Co_3$	1	823	823	823	772	-6.2
$MnCo_2Sn$	10	767	860	818	727	-11.1
$Ba_{0.9}Ce_{0.1}Fe_{12}O_{19}$	1	811	811	811	684	-15.6
$Ce_{0.8}Zr_{0.2}Co_5$	1	809	809	809	849	4.9
$GdFe_2$	15	770	895	804	766	-4.7
$GdV_2Fe_{10}N$	1	795	795	795	737	-7.3
MnFeCoGe	2	711	850	781	595	-23.8
$Y_2Fe_8Co_6B$	10	428	923	779	880	13.0
$Cu_{0.5}Fe_2Ni_{0.5}O_4$	3	770	787	778	688	-11.6
$La_2Fe_7Co_7B$	1	775	775	775	783	1.0
$GdNb_{0.65}Fe_{11.35}N$	3	773	773	773	737	-4.7
$Fe_{0.33}Ru_{0.33}Si_{0.34}$	1	770	770	770	623	-19.1
$NdGdFe_{17}N_{2.8}$	1	768	768	768	734	-4.5
$YV_2Fe_{10}N$	1	767	767	767	692	-9.8
$Nd_3Ti_{1.3}Fe_{16.62}Co_{11.08}$	3	699	876	758	879	16.0
$Pr_3Ti_{1.5}Nb_{2.75}Fe_{16.5}Co_{8.25}$	1	753	753	753	726	-3.6
$Sm_{0.7}NbFe_{4.6}N$	1	753	753	753	692	-8.1
Fe_3Ge	3	740	755	750	695	-7.4
$Nd_2Fe_7Co_7B$	27	443	973	749	858	14.5
$Sm_2VFe_{16}N_{2.2}$	1	745	745	745	713	-4.3
$Sm_2Fe_{13}Co_2Si_2N_{2.3}$	1	742	742	742	648	-12.6
$GdTiFe_{11}N$	2	733	745	739	727	-1.6
$Nd_2Fe_{17}N_3$	6	725	743	737	738	0.1
$MnCo_2Ga_{0.5}Sn_{0.5}$	2	700	770	735	736	0.1
$Pr_3Ti_{1.5}Fe_{13.75}Co_{11}Ga_{2.75}$	1	733	733	733	722	-1.5
$Cr_{0.25}Fe_{0.25}CoAl_{0.5}$	2	624	840	732	717	-2.0
$Pr_2Fe_{11.6}Co_2Al_{0.4}B$	2	680	780	730	700	-4.1
$Sm_2Cr_2Fe_{11}Co_4C_2$	2	678	781	730	678	-7.2
$Ce_{0.8}Sm_{1.2}Fe_{17}N_{2.7}$	1	726	726	726	782	7.7

Continued on next page

Table 3.2: Curie Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of T_c Values (i.e., Model Output), Along With the Percentage Difference From the Mean Experimental T_c Values (Ground Truth)[a]. (Continued)

$GdFe_3$	1	725	725	725	662	-8.7
$Nd_2Nb_{0.5}Fe_{11.5}Co_2B$	1	724	724	724	680	-6.0
$Ba_{0.95}Mg_{0.5}Zn_{0.5}Gd_{0.05}Fe_{16}CoO_{27}$	1	722	722	722	733	1.6
$Ce_{0.1}Pr_{0.3}Nd_{1.6}Fe_{11.88}Co_2Si_{0.12}B$	1	718	718	718	720	0.3
$Gd_3Mo_{1.5}Fe_{27.5}N_4$	1	718	718	718	709	-1.3
$Ba_{0.95}Mg_{0.5}Zn_{0.5}Ce_{0.05}Fe_{16}CoO_{27}$	1	718	718	718	678	-5.5
$Ce_{0.1}Pr_{0.3}Nd_{1.6}Fe_{12}Co_2B$	3	715	715	715	725	1.4
$ErNb_{0.65}Fe_{11.35}N$	2	715	715	715	733	2.6
$YRe_{1.2}Fe_{5.4}Co_{5.4}$	1	715	715	715	780	9.1
$TmTa_{0.7}Fe_{11.3}C$	1	714	714	714	694	-2.8
$TbTiFe_{11}C_{0.9}$	1	714	714	714	683	-4.4
$TbNb_{0.65}Fe_{11.35}C$	2	713	713	713	701	-1.7
$Nd_{1.9}Tb_{0.1}Re_{0.1}Fe_{11.9}Co_2B$	1	712	712	712	713	0.1
$Nd_{2.55}V_{0.26}Fe_{10.34}Co_{2.55}B_{1.30}$	1	712	712	712	638	-10.5
$Ho_2Fe_{17}N_3$	3	709	710	710	708	-0.3
$Ce_{0.1}Pr_{0.3}Nd_{1.6}Fe_{11.85}Co_2Al_{0.15}B$	1	710	710	710	706	-0.5
$Ce_{0.1}Pr_{0.3}Nd_{1.6}W_{0.12}Fe_{11.88}Co_2B$	1	707	707	707	715	1.1
$LuNb_{0.65}Fe_{11.35}N$	2	705	705	705	679	-3.6
$Sm_3Mo_{1.5}Fe_{27.5}N_4$	1	704	704	704	717	1.9
$Y_2Fe_{17}N_{2.5}$	7	690	740	704	662	-6.0
$Nd_2Re_{0.2}Fe_{11.8}Co_2B$	3	698	705	703	669	-4.8
$Tb_3Ti_{1.3}Fe_{22.16}Co_{5.54}$	1	703	703	703	660	-6.2
$Li_{0.25}Y_{0.1}Fe_{2.15}Co_{0.5}O_4$	1	700	700	700	763	9.0

[a]The results are sorted by the magnitude of mean T_c values in descending order, where n is the number of experimental values sampled for each chemical composition; min and max represent their range; mean is their corresponding descriptive statistic. The absolute difference between the predicted values and the experimental measurements is colour-coded in green ($0-10$ %), amber ($10-20$ %) or red (>20 %).

3.1.5 Predictive Capabilities for ML-Model Applications: Two Case Studies

We now explore potential applications of our ML models beyond benchmarking against state-of-the-art reports. Specifically, we aim to: (i) predict Curie temperatures of rare-earth intermetallic compounds, and (ii) generate magnetic phase diagrams of chemical compounds, particularly those in a binary system, where T_c values are computed as a function of elemental composition. We assess how well the models perform when given the task of making out-of-distribution predictions. Here, out-of-distribution predictions are defined as model predictions made within a chemical space that is either under-represented or not represented by the data in the training set. Therefore, we are challenging our ML model to explore unfamiliar material spaces by extrapolating from patterns which pertain to chemical-property relationships that they have learned during the model training stage. This exploration will provide a clearer understanding of the predictive capabilities and suitability of our model for more practical applications.

Figure 3.11 depicts the predicted T_c values (as black diamonds) of rare-earth intermetallic compounds against experimental measurements (i.e., the ground truth which is given as grey circles). Figure 3.11 (a) to (c) refers to unseen RX_2 compounds (X = Fe, Co, Ni), where R represents a rare-earth element while Figure 3.11 (d) to (f) corresponds to more intricate R-Fe-based intermetallic compounds. The x-axis in each graph illustrates the chemical composition and the specific rare-earth elements considered in the prediction.

Figure 3.11 shows that the majority of the T_c predictions made by our ML model for these unseen chemical compounds align well with the experimental measurements. However, a few notable deviations are evident. For instance, the predicted magnetic ordering temperature for $NdFe_2$ is ca. 568 K, while the median experimental measurement is 453 K, as evidenced by the large discrepancy shown in Figure 3.11 (a). A closer inspection of the data set reveals that there are three independently measured experimental T_c values for this particular compound, which are quite disparate from each other: (i) 328 K, (ii) 453 K, and (iii) 578 K. Considering this, our prediction appears reasonable.

Furthermore, Figure 3.11 displays the results in order of ascending rare-earth atomic number. This revealed a discernible trend in these rare-earth intermetallic compounds: the incorporation of Gadolinium (Gd) into the chemical composition yields the highest T_c values in all cases. This pattern is evidenced by the prediction of our model, which aligns with experimental data, demonstrating a clear peak when R is Gd for the chemical material classes considered herein. This trend stands to reason, given that

Gd lies in the middle of the lanthanide series, whose elements are mostly stable in their R^{3+} electronic configuration. Thereby, the electronic configuration of Gd^{3+} ions is [Xe]$4f^7$,$5d^0$,$6s^0$, i.e., all its f orbitals carry an unpaired electron that will maximise its magnetic moment. The peak T_c value for Gd will fall off fairly symmetrically as a function of an increasing or decreasing atomic number for R from that of Gd, since the number of unpaired electrons will decrease as the rare earths extend to the lower and upper ends of the lanthanide series.

Figure 3.12 illustrates the ML-generated phase diagrams for (a) Manganese-Cobalt (Mn-Co), (b) Platinum-Nickel (Pt-Ni), (c) Cobalt-Iron (Co-Fe), and (d) Nickel-Iron (Ni-Fe) binary systems; i.e., the prediction of T_c values as a function of an elemental composition for a given class of materials. The black line represents the ML predictions. Experimental measurements, unseen during the training of our ML model, are depicted by red circles and orange triangles, while blue diamonds represent those seen by the models. Overall, we observe a strong correspondence between the predictions and the experimental measurements. The models demonstrate their ability to generate out-of-distribution predictions with high accuracy. Figure 3.12 (d) further showcases the proficiency of the model in generating accurate out-of-distribution predictions, even in the presence of a transition between two phases in the Ni-Fe system, as indicated by the unseen experimental measurements marked in circles and triangles.

We now turn our attention to specific details within certain magnetic phase diagrams. The Mn-Co system has undergone extensive review by Ishida et al. [158] and Men'shikov et al. [159]. Those studies elucidated the formation of disordered alloys and their crystal structures across a range of chemical compositions. The magnetic phase diagram of the Mn-Co binary system referred to in Figure 3.12 (a) encompasses regions of ferromagnetic and antiferromagnetic long-range order, along with super-paramagnetic and super-antiferromagnetic states. The latter states involve mixtures of ferromagnetic and antiferromagnetic clusters with the paramagnetic phase, respectively. Detailed magnetic phase diagrams can be referred to in the cited studies. Our discussion primarily focuses on the ferromagnetic state (from 0 to ca. 25-30 at % Mn) and neglects the antiferromagnetic state (above ca. 35 at % Mn) and the mixture states (in between the ferromagnetic and antiferromagnetic states), since the scope of this study is on ferromagnetic materials; besides, the experimental points for the mixture states are acknowledged to be less accurate, despite their established existence.

The profile of the magnetic ordering temperature of the Mn-Co system is known to monotonically decrease as a function of the Mn content from 0 to 30 at % Mn, before monotonically increasing above ca. 35 at % Mn. Our ML-generated magnetic phase

diagram effectively reproduces these experimentally observed patterns, even though our model was trained exclusively on chemical composition-based features, with only elemental Co being seen by the model among the experimental data that pertains to this binary system. Surprisingly, the inversion point at the critical temperature at ca. 35 at % Mn is captured, despite corresponding to the antiferromagnetic region. The mean magnetic moment ($\bar{\mu}$) within the ferromagnetic states decreases with an increase in Mn concentration relative to Co concentration (i.e., $\frac{d\bar{\mu}}{dc_{Mn}} < 0$). This signifies that in the Mn-Co alloy system, $\bar{\mu}$ diminishes due to the antiferromagnetic positioning of Mn atoms in relation to Co atoms. Specifically, the local magnetic moment (μ) of Mn atoms ranges from ca. $25\mu_B$ to $29\mu_B$ within the concentration range of 0 to ca. 20 at % Mn. The value of $\bar{\mu}$ continues to decrease beyond ca. 20 at % Mn due to a reduction in μ of Co atoms, whose values are influenced by Mn atoms through exchange interactions. The proportion of ferromagnetically interacting Co-Co pairs decreases until it reaches zero at a concentration of ca. 27 at % Mn. In contrast, the proportion of antiferromagnetically interacting Mn-Co and Mn-Mn pairs increases with a higher concentration of Mn. Therefore, the Mn-Co alloy system can be characterised as a typical Ising magnet, with antiferromagnetic interactions prevailing above a concentration of about 27 at % Mn. This observation aligns with the results obtained from our GBFS workflow, where the most important feature is identified as the mean magnetic moment of elemental solids for atoms within a given chemical composition, as one might expect.

Another magnetic phase diagram that warrants detailed discussion is the Ni-Fe alloy system (cf. Figure 3.12 (d)). In contrast to the Mn-Co alloy, the Ni-Fe alloy exhibits ferromagnetic states throughout its entire compositional range. The equilibrium phases of the Ni-Fe system encompass the: (i) liquid phase, (ii) body-centred cubic, high-temperature (δFe) solid solution, (iii) face-centred cubic (γFe, Ni) solid solution, (iv) body-centred cubic, low-temperature (αFe) solid solution, and (v) intermetallic compound of the form $FeNi_3$ [160–163]. Our primary focus is on examining the T_c profile of the Ni-Fe alloy system associated with the Fe-rich (αFe) solid solution and (γFe, Ni) solid solution, as well as exploring the change in the T_c profile that corresponds to the two intermetallic phases.

Determining the phase boundaries in magnetic materials is challenging and their identification generally necessitates the use of advanced experimental materials characterisation methods such as powder X-ray diffraction and scanning transmission electron microscopy. Our ML-based magnetic phase diagram in Figure 3.12 (d) depicts a decreasing trend in T_c values as the Ni concentration increases from 0 to ca. 35 at % Ni.

Beyond this point, an increasing trend is observed until a peak T_c of ca. 860 K is reached at ca. 73 at % Ni. Further increases in Ni concentration lead to a subsequent decline in the T_c value until a concentration of 100 at % Ni has been reached. It should be noted that an additional ferromagnetic phase involving the $FeNi_3$ intermetallic compound is reported in the literature within the Ni concentration range of approximately 60 to 85 at % Ni. However, this analysis does not consider the magnetic ordering temperature of $FeNi_3$, so we would not expect our ML model to identify such a phase. Our T_c predictions otherwise agree well with the experimentally determined magnetic phase diagram for the Ni-Fe binary system produced by Swartzendruber et al. [161].

We note that our model for the Ni-Fe binary system successfully interpolates the chemical relationships derived from the training to make accurate out-of-distribution predictions, despite having encountered only two experimental measurements among those in this binary system during the training process; albeit, the training process is helped by the fact that these two experimental values lie at either end of the limiting compositional range of this binary system such that they give the model anchoring points with high statistical leverage. Our model identifies both the minimum and maximum T_c values, in addition to mirroring the experimental delineation of the T_c profiles corresponding to (αFe) and (γFe, Ni). Such predictions were not anticipated without structural information, especially considering the scarcity of experimental data points seen by the model across this compositional range. However, as discussed earlier, the mean magnetic moment of elemental solids appears to have played a crucial role in facilitating this prediction. In the ferromagnetic Ni-Fe system, $\bar{\mu}$ measured in the (αFe) phase ranges from $2.00\mu_B$ to $2.29\mu_B$, while in the (γFe, Ni) phase, it ranges from $0.61\mu_B$ to $1.93\mu_B$ [161]. This indicates a clear distinction in the range of $\bar{\mu}$ values between the two phases, from which the ML model appears to have extracted the necessary information to distinguish between the two intermetallic T_c phases. The utilisation of magnetic measurements for the study of two-phase Fe-Ni alloys is described by Sucksmith et al. [164, 165]. Lastly, it is crucial to acknowledge uncertainties associated with the examination of (αFe)/(γFe, Ni) phase boundaries. The presence of material impurities can influence the transformation between these two phases. Additionally, environmental conditions during the varying experimental measurement processes will play an important role. For instance, applied pressure has been shown to lower the transformation temperature between the two phases (i.e., (γFe, Ni) $\rightarrow$ (αFe)) [161, 166]. Therefore, discrepancies are expected among the experimental measurements reported in the literature.

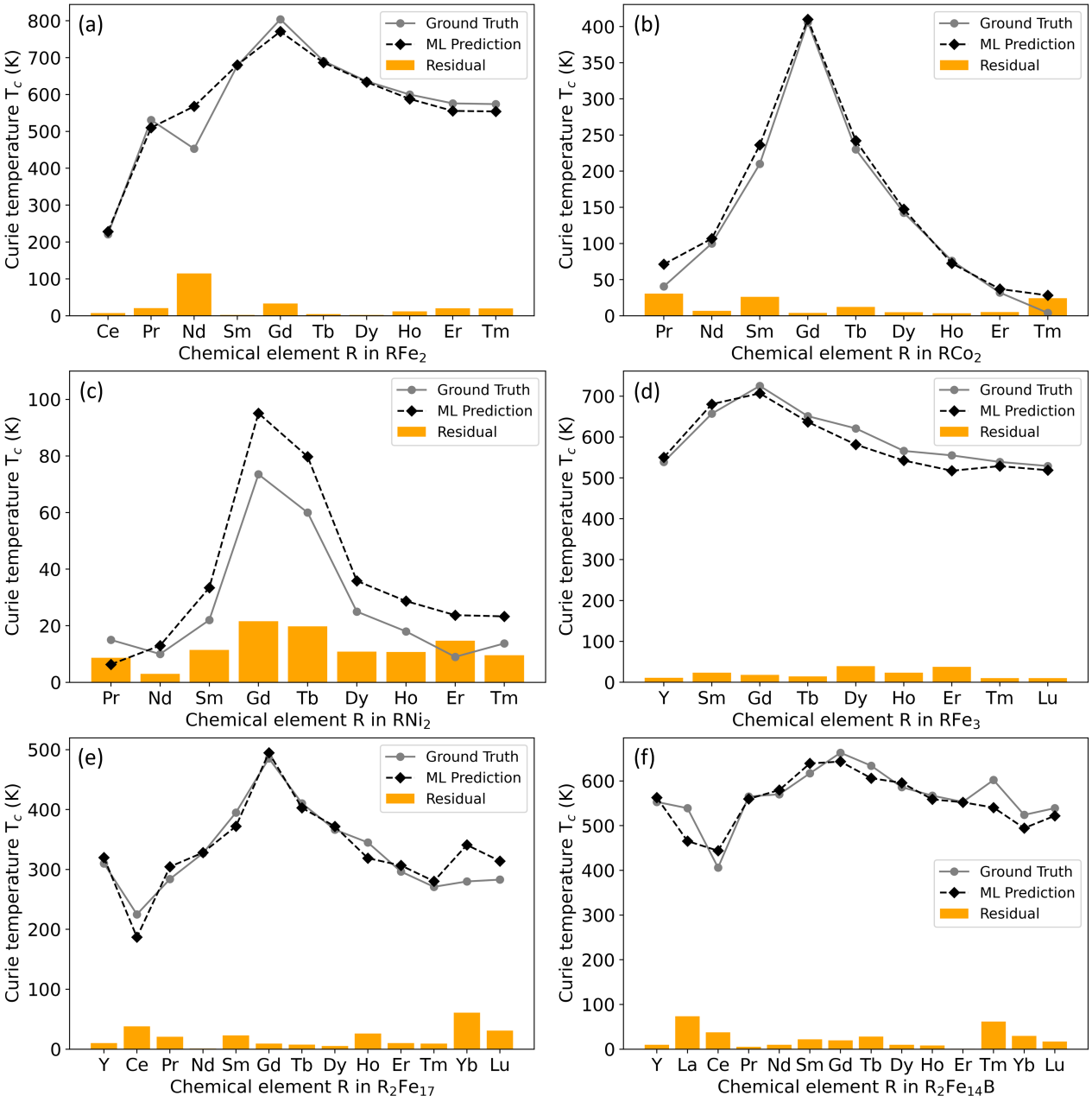


Fig. 3.11 ML-based predictions of T_c values for unseen (a)-(c) RX_2 and (d)-(f) R-Fe-based chemical compounds (black diamonds) against median experimental measurements obtained from T_c values in the literature (grey circles) [114, 140]. The bar chart illustrates the residuals (orange bars), defined as the absolute discrepancies between the predicted values and the experimental measurements.

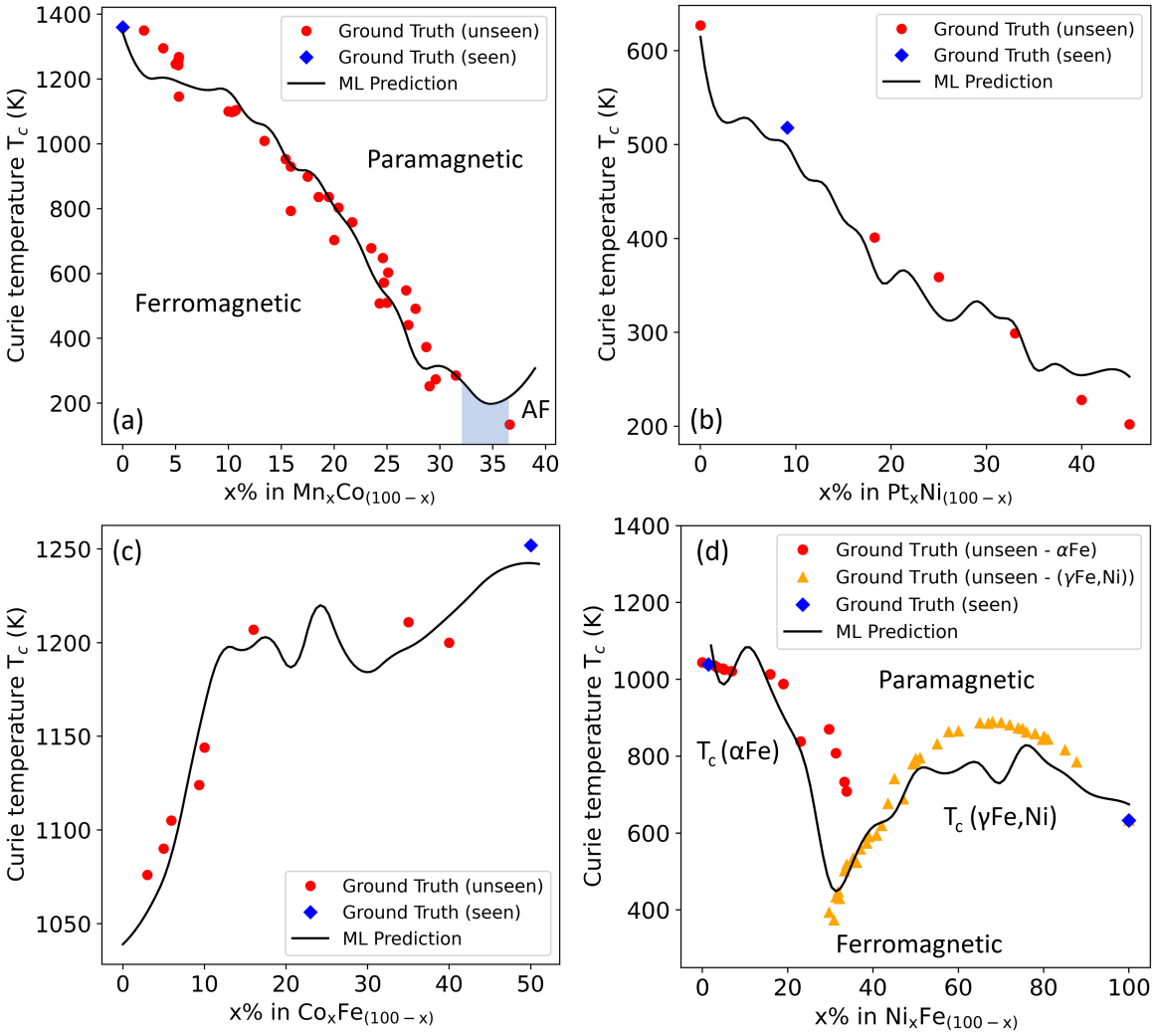


Fig. 3.12 ML-generated magnetic phase diagrams - prediction of T_c as a function of elemental composition for (a) Mn-Co [158, 159], (b) Pt-Ni [135], and (c) Co-Fe [132], and (d) Ni-Fe [160–163] binary systems. The results are compared against seen (diamonds) and unseen (circles and triangles) experimental measurements of T_c values from the literature. The shaded blue region in (a) depicts the approximate phase transition between the ferromagnetic and antiferromagnetic (AF) phases. The two types of unseen experimental measurements in (d) represent the two different Curie temperature profiles in the Ni-Fe binary system.

3.2 Prediction of Band Gaps of Inorganic Materials

Having demonstrated the effectiveness of our workflow in predicting the Curie temperature of ferromagnetic materials against experimental measurements, we now present the results for additional properties of inorganic materials.

3.2.1 Classification of Materials by Metallicity

Performance Results

The classification of materials by their metallicity was performed by using a gradient-boosting algorithm using 27 (out of 827) features selected via the GBFS workflow. The model performance is illustrated in Figure 3.13, along with the performance metrics summarised in Table 3.3. An accuracy score of 0.943 and a balanced accuracy score of 0.938 were achieved. The ROC curve illustrates the diagnostic ability of the model toward the target classes, as the classification threshold is varied by depicting the variation of the true positive rate against the false positive rate (FPR). A macro-averaged AUC-ROC of 0.985 was realised. This indicates that the final classifier is highly discriminative toward the two target classes; this is consistent with the output class-probability distribution (in Appendix A.1), which is illustrative of an almost binary outcome. In comparison, Zhou et al. achieved an accuracy of 0.92 and an AUC-ROC of 0.97.

Table 3.3 Summary of the Performance Metrics for the Final Classifier When Applied to the Test Set.

	Precision	Recall	F1-score
Metal	0.947	0.959	0.953
Non-metal	0.936	0.917	0.927
Macro average	0.942	0.938	0.940
Weighted average	0.943	0.943	0.943

The detection error trade-off (DET) shows the change in the FPR relative to the false negative rate (FNR) as the classification threshold is varied. The point at which the FPR and FNR cross the diagonal, known as the equal error rate (EER), is approximately 0.06. This result suggests a low error rate can be obtained in the model predictions using the selected features. The fact that the plot profiles are fairly symmetric about their respective diagonals indicates that the class imbalance has been well corrected. Moreover, the precision–recall (PR) curve was plotted to assess the

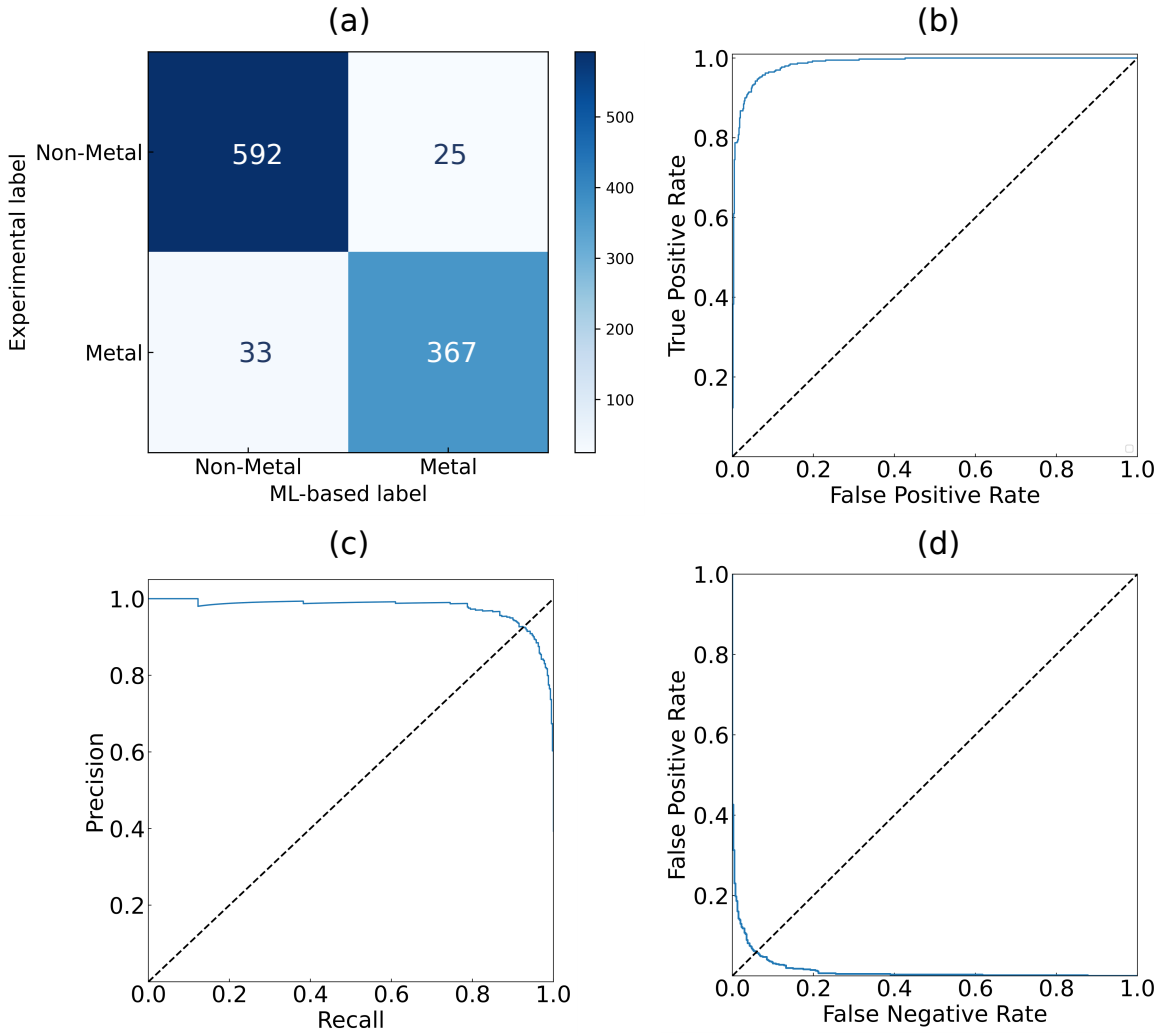


Fig. 3.13 Classification performance on the test set. (a) Confusion matrix, (b) ROC, (c) PR, and (d) DET curves. The macro-averaged AUC-ROC of 0.985, AP of 0.977, and EER of 6% were achieved.

effectiveness of the model, owing to the presence of an imbalanced data set. The PR curve demonstrates the quality of the model performance by illustrating the trade-off between precision and recall as the classification threshold is varied. This is important because imbalanced classes can lead to the FPR becoming less informative as a large number of negatives (i.e., false positives and true negatives) yields a low FPR, while still allowing poor precision. An average precision (AP) of 0.977 was achieved; this is the average of the precision at each classification threshold weighted by the change in the recall from the prior threshold. The result indicates that both high precision and recall are realised, where a low FPR and a low FNR are attained. This implies that the model is capable of returning a high percentage of positives (i.e., true positives and false negatives) that are mostly classified correctly.

The overall model performance can be evaluated via the F1-score. The macro- and weighted-averaged F1-scores were 0.940 and 0.943, respectively. Zhou et al. did not state their F1-score. The benchmark scores from the Matbench test suite v0.1 show the highest F1-score and balanced accuracy score as 0.920 and 0.921, respectively. This further supports that our classifier is highly discriminative toward the two target classes.

Upon closer examination of the predictions, clear trends are discernible in the characteristics of misclassified chemical compounds. An analysis of the two most influential features, based on the total loss reduction, reveals distributions that differ from those observed in the training set. In the scaled distribution of the one-hot-encoding of the highest occupied molecular orbital (HOMO) character corresponding to the *p*-orbital, we observed that in the training set, the average feature values for nonmetals and metals are ca. 0.82 and 0.26, respectively. This finding indicates a stronger association of nonmetals with HOMO with *p*-orbital characteristics. A similar feature distribution is noted among the correctly classified compounds in the test set. However, among misclassified compounds in the test set, the average feature values deviate to 0.40 for nonmetals and 0.67 for metals, which goes against the general trend.

A comparable observation is made regarding the mean number of filled valence *p*-orbitals among elements in the chemical composition. The scaled distributions between correctly classified chemical compounds and those in the training set are consistent, hovering around 0.50 and 0.26 for nonmetals and metals, respectively. This suggests that nonmetals tend to have a higher mean number of filled valence *p*-orbitals. In contrast, misclassified compounds exhibit feature values of ca. 0.46 for nonmetals and ca. 0.42 for metals. Once more, this observation contradicts the overarching pattern. Notably, compounds misclassified as metals manifest a distribution closer to

nonmetals in both cases. It seems that the model tends to classify chemical compounds as nonmetals when relatively higher feature values are observed for these two features, which are associated with p-orbitals.

These observations align with the fundamental principles of chemistry, as the majority of nonmetal and metalloid elements are situated in the p-block, encompassing groups 13–18 of the periodic table. The p-block comprises chemical elements in which np orbitals are filled, resulting in distinctive chemical properties that distinguish them from those of elements in other blocks of the periodic table. Consequently, understanding the statistical metrics related to the number of filled p-valence orbitals offers insights into chemical elements that are highly discriminatory toward the target classes. This leads to specific chemical elements being associated with having a greater likelihood of being a nonmetal, particularly for elements such as Si ([Ne]$3s^2 3p^2$), Ge ([Ar]$4s^2 3d^{10} 4p^2$), and As ([Ar]$4s^2 3d^{10} 4p^3$), which are common in semiconductors and belong to the p-block in periods 3 and 4 and groups 14 and 15.

Oversampling

The benchmark data set used to achieve the results above consists of 6354 compounds, which partition into 2,458 metal ($E_\mathrm{g} > 0$ eV) and 3,896 nonmetal ($E_\mathrm{g} = 0$ eV) materials, and a train-to-test split ratio of 4:1. A total of 827 composition-based features were computed. The imbalanced data set was treated by applying the smoothed random oversampling (smoothed-ROS) method with a shrinkage of 0.35, which is an extension of ROS with the introduction of Gaussian noise. The noise is stochastic. Therefore, it prevented an overtraining of the model on particular values of a feature due to the increased intraclass sample variability, and it appeared to improve model generalisation. We found the smoothed-ROS method to be the most effective in alleviating potential learning biases among other oversampling techniques.

3.2.2 Regression Analysis of Band Gap

Performance Results

The regression analysis of E_g was performed using a gradient-boosting algorithm with 46 features selected via the GBFS workflow. The model performance and error distributions are listed in Figure 3.14. The solid blue line is the line of best fit between the experimental measurements and ML-based predictions generated using the OLS method. The linear fit has a gradient of 0.93 and a y-intercept of 0.16 eV. The y-intercept may indicate a small systematic bias for small values of E_g. Nevertheless, the

linear fit illustrates the close correspondence between the experimental measurements and ML-based predictions. This is further supported by the R^2 value of 0.937, which indicates a high correlation between the prediction and the ground truth. Moreover, MAE and RMSE values of 0.246 and 0.402 eV were realised, respectively, where a greater degree of error in the latter is due to the greater penalisation of predictions with larger deviations from their true values. The distributions of absolute errors and root of squared errors are shown in Figure 3.14b,c, respectively. In comparison, Zhou et al. achieved an R^2 of 0.90 and RMSE of 0.45 eV, which also demonstrates the robustness of their SVR model. Our model performance is comparable or superior to those achieved with alternative methods.

A closer examination of the training set reveals that the chemical composition with the smallest E_g of 0.02 eV corresponds to $Pb_{0.87}Sn_{0.13}Se$ and the largest E_g of 11.1 eV corresponds to MgF_2. The E_g value is less than 3 eV for the majority of the chemical compositions of materials in the training set, as illustrated by the population distribution plot in Figure 3.15. This explains the relatively larger deviations in the prediction for semiconductors with larger E_g values. Specifically, we observe an underestimation of E_g where its value is large, as depicted by the line of best fit (in blue) in Figure 3.14a, which sits below the diagonal line in red. We attribute such an observation to the lack of wide or ultrawide band gap semiconductors in the training data. Nevertheless, the ML modelling approach has demonstrated an accurate prediction of E_g in the absence of structural information. This is despite our methodology not involving specific adjustments or treatments to accommodate different crystal forms of the same chemical compound, a phenomenon known as polymorphism. Our approach strictly trains the model to be agnostic to polymorphs, aligning with the methodology employed by Zhou et al. This ensures a like-for-like comparison in our study. This is to say that the models demonstrate good performance, even though they are unaware of the different crystal forms that the sample compound may adopt, owing to the various ways in which chemical elements can arrange within unit cells of the crystalline lattice of each crystal form. See Section 3.2.2 for the additional regression analysis where we evaluate the model predictions against the median and mean experimental E_g values. Taking our attention back to the error distribution plots in Figure 3.14, which has a log scale on the y-axis, we see that the majority of predictions have an error below ca. 1 eV. At a closer examination, out of 780 chemical compositions in the test set, 670 predictions (ca. 86%) have an absolute error below 0.5 eV and 545 predictions (ca. 70%) have an absolute error below 0.25 eV.

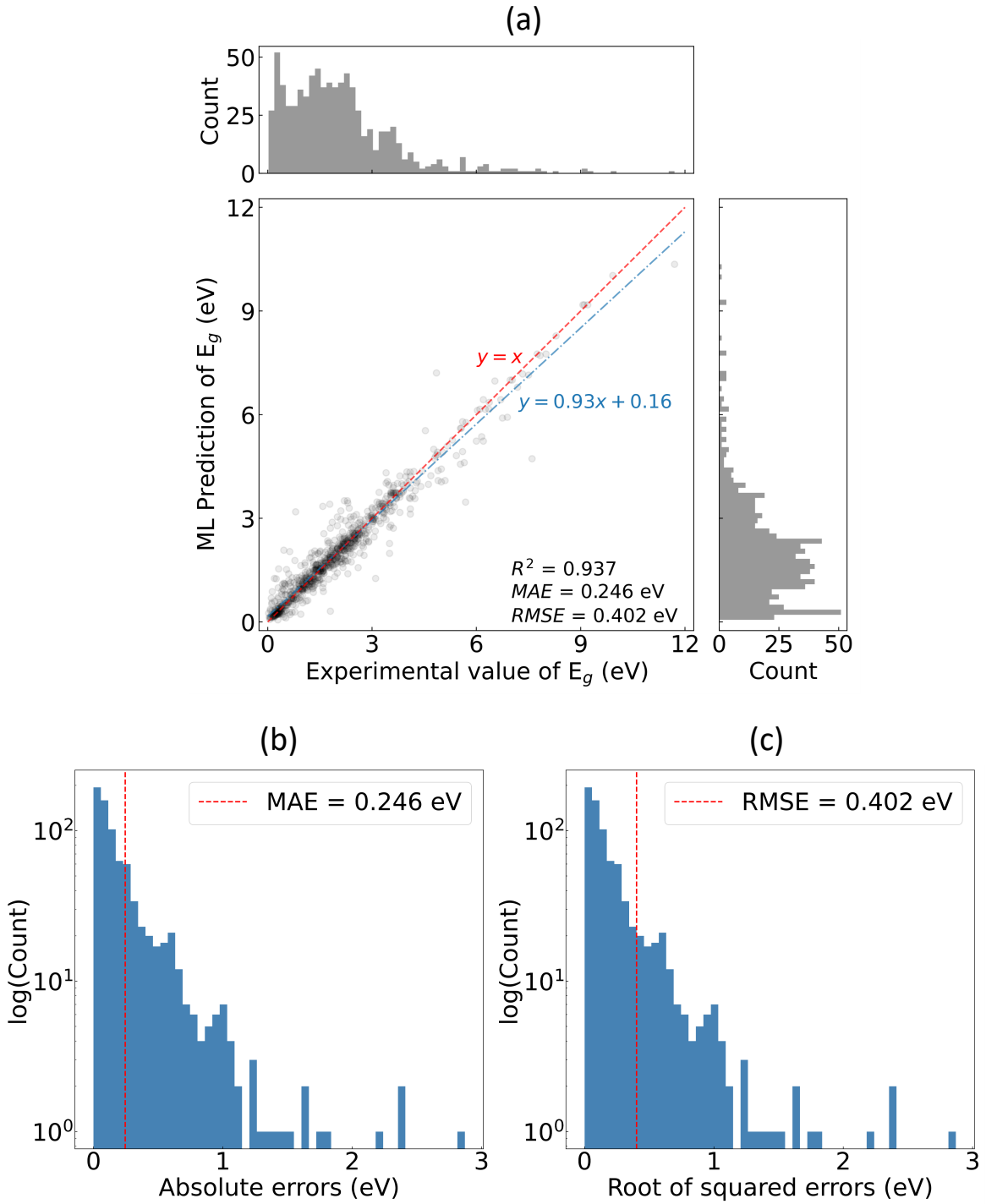


Fig. 3.14 (a) Regression of the ML-based predictions of E_g against experimental measurements, where the regression models are trained on the final subset of features selected by the GBFS workflow. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the experimental measurement. The solid blue line is a linear fit generated using the OLS method. Distribution of the (b) absolute errors and (c) root of squared errors of the ML-based prediction of E_g are shown, where the dashed red line indicates the MAE and RMSE, respectively.

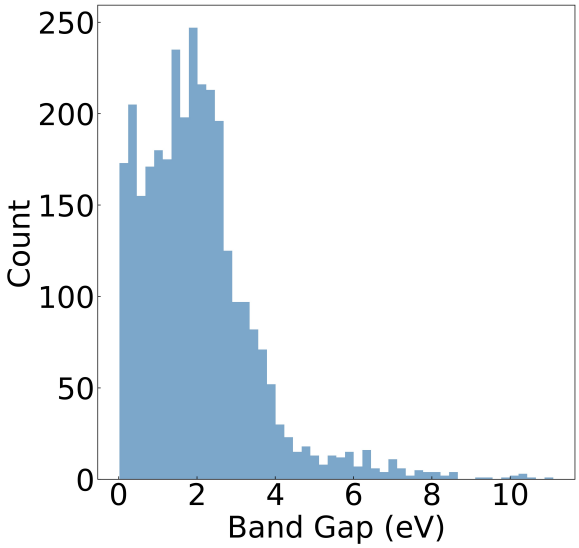


Fig. 3.15 Distribution of band gap (E_g) values of nonmetals in the training set.

Table 3.4 Summary of the Band Gap Predictions (E_g in eV) against Experimental Measurements

Composition	Experimental E_g	GBFS E_g	SVR E_g [10]	PBE E_g	GW E_g	HSE E_g
PbTe	0.19 [167]	0.215 (13%)	0.2 (5%)	0 (-100%) [167]	0.26 (36%) [167]	0.19 (0%) [168]
$CuSbS_2$	1.38 [169]	1.40 (1%)	1.39 (1%)	0.9 (-35%) [169]	1.1 (-20%) [169]	1.69 (22%) [169]
GaN	3.2 [167]	3.00 (-6%)	4.45 (39%)	1.62 (-49%) [167]	3.32 (4%) [167]	3.14 (-2%) [170]
TiO_2	3.42 [171]	3.25 (-5%)	3.99 (16%)	2.13 (-37%) [171]	3.73 (9%) [171]	3.67 (7%) [171]
ZnS	3.91 [167]	3.41 (-13%)	3.12 (-20%)	2.07 (-47%) [167]	4.15 (6%) [167]	3.49 (-11%) [170]
LiF	14.2 [167]	10.53 (-26%)	9.87 (-30%)	9.2 (-35%) [167]	15.1 (6%) [167]	11.47 (-19%) [172]
MAE		0.76	1.16	1.73	0.32	0.63
RMSE		2.29	3.54	5.47	0.18	1.30

The results discussed thus far present predictions of E_g values against experimental measurements with promising statistical figures-of-merit. Nonetheless, it is important to validate these results by considering how these predictions fare against well-studied inorganic compounds rather than simply demonstrating their collective statistical quality in an anonymised form. A comparative analysis is, therefore, conducted against six unseen compounds that are extensively researched both experimentally and at various theoretical levels. The results are summarised in Table 3.4.

The aforementioned underestimation of DFT-based E_g values that have been computed using a PBE functional is apparent among the results, with a negative percentage difference being shown across all compounds. Our model (labelled GBFS E_g) shows a negative percentage difference for E_g values exceeding approximately 3 eV. This aligns with the findings from our regression analysis, where we observed a systematic positive bias of ca. 0.16 eV at lower E_g values and an underestimation at higher E_g values. The gradient between the ML predictions and the ground truth was observed to be 0.93, indicating that the underestimation becomes increasingly pronounced at higher values of E_g. This explains the largest deviations that are observed, for example, in LiF, which predicts an E_g value of 14.2 eV. As previously discussed, we attribute this anomaly to the lack of wide or ultrawide band gap semiconductors within the training set. Nevertheless, ML-based predictions (both GBFS and SVR E_g) yielded lower MAE and RMSE values when compared with high-throughput calculations that incorporated a PBE functional. The result pertaining to GW-type calculations realised the lowest errors among the methods. However, such an approach incurs the greatest computational cost and cannot provide an efficient and automated approach to the E_g prediction. The second lowest errors originated from the high-throughput calculations that employed hybrid functionals (HSE); this is another computationally expensive approach. The MAE achieved using HSE is comparable to the value obtained by our method, while the RMSE is ca. 1 eV lower for the former. Again, this difference stems from the error associated with wide and ultrawide band gap semiconductors. We anticipate a significant improvement of ML models pending the availability of a significantly greater number of ultrawide band gap semiconductors.

We now delve into a more detailed examination of the feature interactions that contributed to the aforementioned results in the regression analysis. The highest relevance, as indicated by the realised total loss reduction, is associated with the HOMO energy (feature number $m = 593$). This is succeeded by the standard deviation of the periodic group among elements in the chemical composition ($m = 258$), the fraction of p-valence electrons ($m = 586$), the average Mendeleev number among elements in

the chemical composition ($m = 115$), the average deviation of electronegativity among elements in the chemical composition ($m = 158$), and the average deviation of the number of filled valence p-orbitals among elements in the composition ($m = 170$).

The selection of HOMO energy by the GBFS workflow as the feature with the highest relevance was anticipated given that it is a parameter directly involved in E_g estimation. The correlation between the HOMO energy and E_g is ca. -0.58 in the training data, indicating that lower HOMO energies correspond to larger E_g values. This correlation is logical as the HOMO represents the highest energy molecular orbital that contains electrons, akin to the valence band in Band theory. The energy of the lowest unoccupied molecular orbital represents the minimum energy level into which an electron can be excited. While its inclusion was expected, its correlation with the E_g values in the training data is not as pronounced as that of the HOMO energy, registering a correlation magnitude of only 0.17.

The subsequent set of features that are deemed significant in predicting E_g is associated with statistical values based on the period group number and Mendeleev order number among elements in the chemical composition. These features are directly linked to the chemical compositions of the compounds. The standard deviation of the periodic group among elements in the chemical composition exhibits a correlation of ca. 0.38 with the target variable. This suggests a discernible trend between the two, indicating that higher deviations in the periodic group number among elements are associated with elevated values of E_g. This aligns with the observation that some of the largest E_g values in the training data are found in chemical compositions such as MgF_2, NaF, RbF, and BeO. These compositions involve s-block metals in groups 1 and 2 paired with p-block nonmetals and halogens in groups 16 and 17 (e.g., oxygen and fluorine), generating some of the largest standard deviations of the periodic group among elements in the chemical composition. Moreover, the Mendeleev number, distinct from the atomic numbering system, is an ordering assigned to each chemical element in the periodic system. Its purpose is to arrange elements so that those with similar behaviours are placed consecutively. Similar to the previous feature, a clear trend is apparent, wherein a lower mean Mendeleev number among elements in the composition correlates with higher values of E_g, with a correlation of ca. -0.53. In the training data, chemical compounds with some of the highest E_g values exhibit a mean Mendeleev number of ca. 10.

Another important set of features to be acknowledged involves p-valence electrons. Two relevant features are identified by the GBFS workflow, which are (i) the fraction of p-valence electrons and (ii) the average deviation of the number of filled valence p-

orbitals among elements in the chemical composition. As discussed in the classification problem, these observations align with fundamental principles of chemistry, given that the majority of nonmetal and metalloid elements reside in the *p*-block, spanning groups 13–18 of the periodic table. The *p*-block encompasses chemical elements in which *np* orbitals are filled, resulting in distinctive chemical properties that set them apart from elements in other blocks of the periodic table. Notably, elements such as Si, Ge, and As, which are well-known in semiconductor applications, fall within this category. A correlation of 0.61 is observed against E_{g} values in the training data for the fraction of *p*-valence electrons.

The final among the most pertinent features to be discussed is the average deviation of electronegativity among elements in the chemical composition. This feature demonstrated the highest correlation with the E_{g} values in the training set, registering a value of ca. 0.68. The elements with the highest electronegativity are typically found in groups 16 and 17 (e.g., fluorine, oxygen, and chlorine), while those with the lowest electronegativity are situated in groups 1 and 2 (e.g., sodium, lithium, potassium, and magnesium). Consequently, the highest average deviation of electronegativity is generally observed between these two regions of the periodic table. The rationale behind selecting this feature lies in the fact that pairing a metal with a nonmetal element corresponds to a large difference in orbital energy. This phenomenon becomes more pronounced as the disparity in electronegativity between the pair of elements increases. In other words, a substantial difference in the electronegativity of two elements in a compound leads to an increase in its ionic properties, which in turn reduces the overlap of orbitals and elevates E_{g} of a material. Consequently, the probability of a material exhibiting a larger E_{g} increases when elements with higher electronegativity, concentrated in the *p*-block, are incorporated into the material composition. This is intuitive, as the strength of ionic bonding or electrostatic interaction is directly determined by the difference in electronegativity between neighbouring ions. The result implies that nonmetal elements, such as oxygen and fluorine, are important attributes to consider when distinguishing metallic bonding from other types of bonding. This stands to reason since metals can exist in the form of oxides, while ionic and metallic characteristics are distinct; an increase in the former contributes to an increase in E_{g} as orbital overlap decreases.

Multifidelity Modelling

Another direction explored in this research was to develop a multifidelity model using 154,718 DFT results from the MP, with 105,584 unique chemical compositions. An auxiliary model was trained on these chemical compositions of materials for the classification process and on 69,219 chemical compositions with $E_g > 0.35$ eV for the regression process, both using the GBFS workflow and a train-to-test split ratio of 4 : 1 (see Appendix A.2 for the performance of the auxiliary models). An energy cutoff of 0.35 eV was applied to the DFT data set to mitigate inherent errors in DFT calculations of E_g. We anticipated that the multifidelity approach could account for the lack of wide band gap semiconductors. However, we observed a comparable performance to that of the models constructed without the multifidelity strategy. The results are listed in Figure 3.16. A closer examination of the feature interactions showed that incorporating DFT-based auxiliary models led to a smaller final subset of features selected via the GBFS workflow: 22 features for the classification and 25 features for the regression. These correspond to a reduction of 5 features in the former and 21 features for the latter. Moreover, DFT-aware features are among the top three in the feature relevance ranking, which is based on the loss reduction realised when training the predictive models. This demonstrates the benefit of such a modelling approach, which has gained popularity within computational material science.

It is crucial to distinguish our work from the multifidelity ML modelling approach conducted by Pilania et al. [9] Their methodology can be classified as multifidelity because it incorporates calculations on 600 chemical compounds using different functionals for each compound, which have varying levels of exchange correlation within DFT. This entails the amalgamation of low- and high-fidelity DFT calculations, enabling cost-effective predictions at a higher fidelity level. Given the multifidelity modelling approach in our work arises from the fact that we have used computational and experimental data sources as input, a comparison of our work with theirs presents challenges, as their exclusive focus on computational data in the context of multifidelity ML models differs from our approach. Nevertheless, the use of a multifidelity approach in their work yielded improvements, while our study demonstrated a singular effect - a substantial reduction in the feature space. Despite being the sole observed outcome, this reduction remains a crucial and noteworthy result, as it can lead to a reduction in the feature space by up to ca. 46%; this renders many of the essential features identified in Section 3.2.2 no longer necessary.

It is unsurprising to encounter this difference, considering the limited data set of 600 chemical compounds in the study by Pilania et al. [9] In such cases, the amalgamation

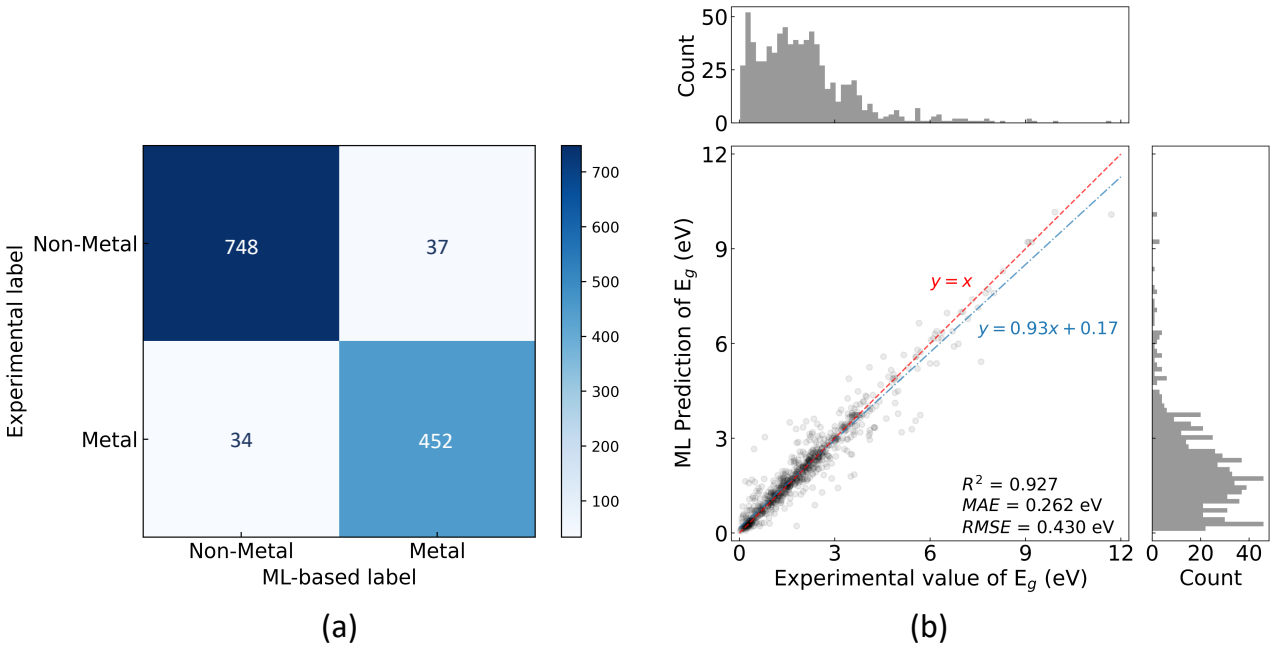


Fig. 3.16 Multifidelity model performance against experimental measurements and DFT-based computed results for the (a) classification of materials by metallicity and (b) regression analysis of E_{g} in the test set, where the classification model was trained with 22 features and the regression model was trained on 25 features, both selected via the GBFS workflow. For the classification process, a macro-weighted AUC-ROC of 0.987, AP of 0.982, F1-score of 0.941, and a balanced accuracy of 0.941 were achieved.

of multiple data sources in the training process would prove advantageous. This is because the augmented training data for the multifidelity model can either facilitate the exploration of a larger chemical space, often mitigating the necessity for extreme extrapolation of learned relationships into unseen territories, or strengthen existing relationships with greater statistical certainty.

To demonstrate this, we employed a pseudo-comparison multifidelity modelling approach utilising only 600 randomly selected experimental measurements, in conjunction with ca. 150,000 DFT calculations. More specifically, we conducted two separate regression analyses: the first involved 600 experimental measurements, and the second utilised the same 600 experimental measurements alongside ca. 150,000 DFT calculations, employing a multifidelity modelling strategy. In the former scenario, we attained an R^2 of 0.80, MAE of 0.50 eV, and RMSE of 0.77 eV. Conversely, in the latter scenario, we achieved an R^2 of 0.86, MAE of 0.44 eV, and RMSE of 0.65 eV. As anticipated, the implementation of the multifidelity modelling strategy led to improvements in the quantified statistical figures-of-merit compared to a standard regression model; this is despite the recognised errors inherent in DFT calculations. Our analysis suggests a trade-off between the broader chemical coverage achieved by incorporating DFT calculations and their inherent uncertainties. Notably, the latter serves as a limiting factor in this study.

Another Experimental Data Set

We extended our regression analysis to another set of experimental measurements, namely those reported by Kiselyova et al. [173] Their data set consists of 7,588 chemical compositions of materials, with 3,233 unique chemical compositions. The result with and without implementing a multifidelity modelling strategy is shown in Figure 3.17. In common with the findings noted above, we find that the model performances are comparable to each other. The use of the multifidelity approach resulted in a reduction of 26 features, with a predominant portion of the total loss reduction being attributed to the DFT-based prediction of E_{g} values. We infer that while DFT calculations facilitate the creation of less complex models in terms of the number of input features, they do not necessarily improve the model accuracy, possibly because of the inherent limitations associated with these calculations.

Moreover, this experimental data set is used to assess the predictions made on the chemical compositions within Pearson's Crystal Structure Database (94,095) and the MP (105,583). A sample of 60 chemical compositions, previously unseen by the model, was randomly chosen, and the predictions were compared against the experimental

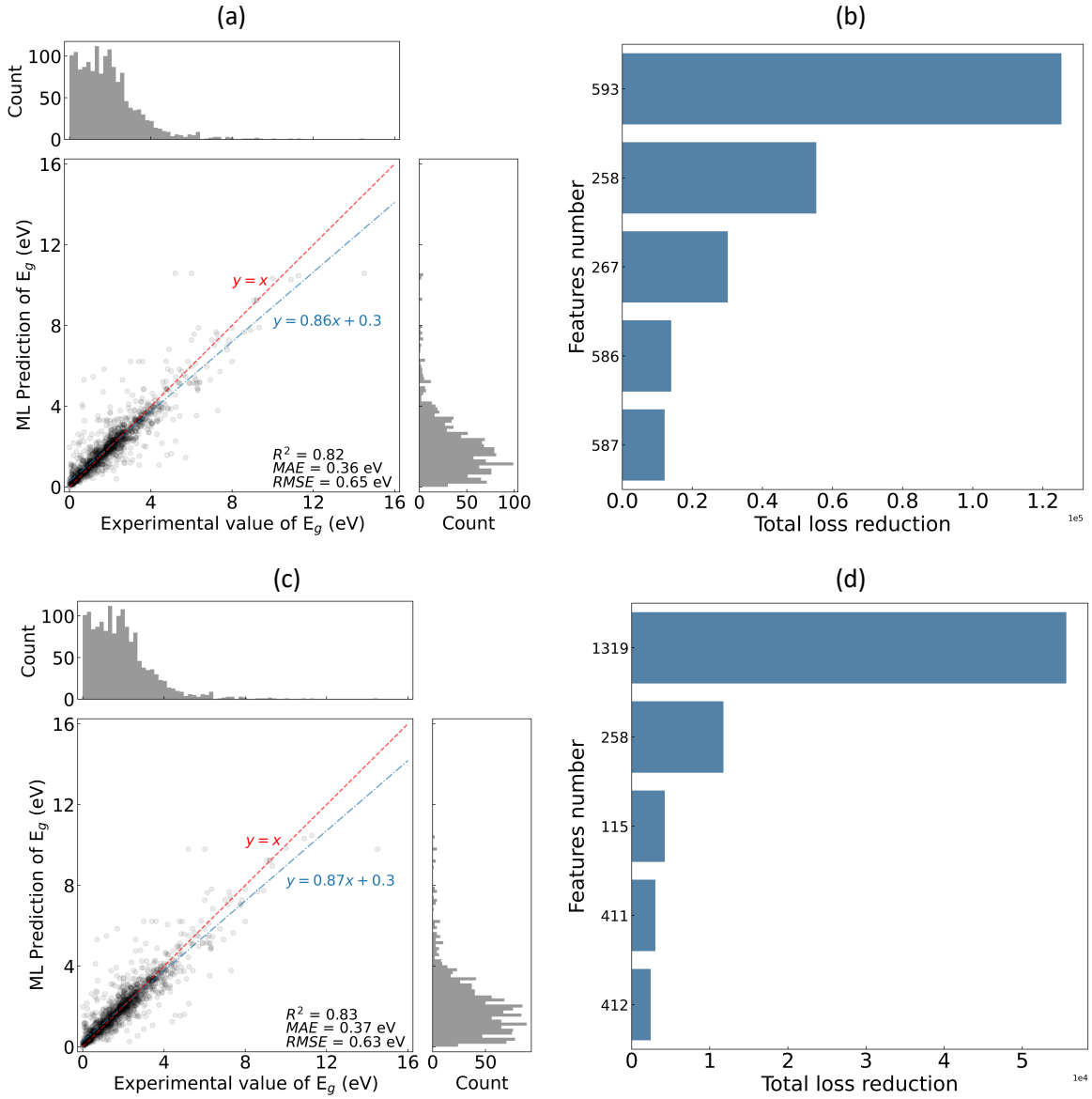


Fig. 3.17 Regression analysis of E_g in the test set against (a) experimental measurements with 42 features, with the 5 most relevant features shown in (b), and against (c) experimental measurements and DFT calculations (i.e., using the multifidelity modelling approach) with 16 features, with the five most relevant features shown in (d). The features were selected via the GBFS workflow. Feature number 1,319 corresponds to the DFT-based prediction of E_g values, which realised the largest total loss reduction.

measurements. The results are summarised in Table 3.5. As anticipated, the model has a tendency to underestimate the experimental E_g values, with 37 and 39 (out of 60) chemical compositions exhibiting a negative percentage difference from the mean and the median E_g values, respectively. The average absolute percentage differences from the mean and from the median are ca. 15 and 13%, respectively. Notably, a lower absolute percentage difference is observed when considering the median E_g value for specific compounds such as $BaCu_2GeS_4$, PPdS, $KP(HO_2)_2$, FeS_2 and As. This suggests that a lower deviation can be realised when considering the median as the measure of central tendency in the presence of multiple experimental measurements, as outliers exert a relatively minimal effect on the median of a given data set compared to the mean.

Table 3.5: Band Gap - Examples of Input Chemical Composition and the Corresponding Prediction of Band Gap (E_g), Sorted by the Absolute Percentage Difference From the Mean Experimental Values in Ascending Order.

Composition	Experimental E_g (eV)					Pred. E_g (eV)	% diff. from mean	% diff. from median
	n	Min	Max	Mean	Median			
$Ga_2(TeO_3)_3$	2	4.14	4.15	4.15	4.15	4.11	-1.0	-1.0
$Hg_3(SI)_2$	1	2.25	2.25	2.25	2.25	2.22	-1.3	-1.3
$BaGeS_3$	1	2.46	2.46	2.46	2.46	2.49	1.2	1.2
HgCl	3	2.84	3.80	3.38	3.50	3.43	1.5	-2.0
$K_3Th_2Cu_3S_7$	1	2.49	2.49	2.49	2.49	2.45	-1.6	-1.6
Bi_2SO_2	2	1.12	1.50	1.31	1.31	1.33	1.5	1.5
$K_2Hg_3(GeS_4)_2$	2	2.64	2.70	2.67	2.67	2.71	1.5	1.5
$La_3Mo_4O_{16}F$	1	3.70	3.70	3.70	3.70	3.64	-1.6	-1.6
$Ba_3Cd(SnS_4)_2$	1	2.75	2.75	2.75	2.75	2.80	1.8	1.8
$Cs_2Ba_3(P_2O_7)_2$	2	5.06	6.31	5.69	5.69	5.79	1.8	1.8
Ga_2PbS_4	4	2.38	2.55	2.46	2.46	2.41	-2.0	-2.0
$La_4Fe(SbS_5)_2$	1	1.00	1.00	1.00	1.00	1.03	3.0	3.0
$Zn(InTe_2)_2$	7	1.13	1.95	1.74	1.87	1.68	-3.4	-10.2
$BaCu_2GeS_4$	8	2.29	2.49	2.43	2.47	2.53	4.1	2.4
Li_2TeMoO_6	1	3.50	3.50	3.50	3.50	3.34	-4.6	-4.6
$PHPb_2N_2O_9$	1	3.81	3.81	3.81	3.81	3.62	-5.0	-5.0
$NbIn(TeO_4)_2$	1	3.50	3.50	3.50	3.50	3.32	-5.1	-5.1
In_5AgS_8	3	1.60	1.79	1.72	1.76	1.63	-5.2	-7.4
$Ba_2DyGaSe_5$	2	2.35	2.35	2.35	2.35	2.22	-5.5	-5.5
$Na_4Mg(SiSe_3)_2$	1	2.85	2.85	2.85	2.85	2.69	-5.6	-5.6
$Na_2Cd(GeS_3)_2$	3	2.57	3.21	2.79	2.60	2.96	6.1	13.8
$Ba_2ErGaSe_5$	2	1.95	1.95	1.95	1.95	2.07	6.2	6.2

Continued on next page

Table 3.5: Band Gap - Examples of Input Chemical Composition and the Corresponding Prediction of Band Gap (E_g), Sorted by the Absolute Percentage Difference From the Mean Experimental Values in Ascending Order. (Continued)

$InHg_7S_6Cl_5$	1	2.54	2.54	2.54	2.54	2.37	-6.7	-6.7
$BaAl_4S_7$	1	3.74	3.74	3.74	3.74	3.48	-7.0	-7.0
$Rb_2Cd_3(B_4O_7)_4$	1	4.76	4.76	4.76	4.76	4.42	-7.1	-7.1
$Cd(GaSe_2)_2$	18	2.10	2.75	2.39	2.41	2.56	7.1	6.2
$KThCuS_3$	1	2.95	2.95	2.95	2.95	2.72	-7.8	-7.8
$Sb_8I_2O_{11}$	5	2.72	2.72	2.72	2.72	2.50	-8.1	-8.1
KCu_2BiS_3	1	1.29	1.29	1.29	1.29	1.40	8.5	8.5
$K_2Mn(SnSe_3)_2$	1	2.00	2.00	2.00	2.00	1.82	-9.0	-9.0
Rb_2VAgS_4	1	1.83	1.83	1.83	1.83	2.00	9.3	9.3
$PrMoO_4F$	1	3.64	3.64	3.64	3.64	3.30	-9.3	-9.3
$ZnSi(AgS_2)_2$	1	3.28	3.28	3.28	3.28	2.96	-9.8	-9.8
$K_2Sn_2Hg_3S_8$	2	2.40	2.50	2.45	2.45	2.19	-10.6	-10.6
PPdS	12	0.70	1.40	1.16	1.38	1.30	12.1	-5.8
$RbH_2(IO_3)_3$	2	4.07	5.08	4.58	4.58	4.00	-12.7	-12.7
$In_4Bi_3S_{10}$	1	1.42	1.42	1.42	1.42	1.24	-12.7	-12.7
$ZnCu_2SiS_4$	3	3.00	3.25	3.17	3.25	2.75	-13.2	-15.4
Te_3As_2	28	0.48	1.88	1.22	1.35	1.40	14.8	3.7
$NbInBi_2O_7$	1	2.70	2.70	2.70	2.70	2.30	-14.8	-14.8
$InCuGeSe_4$	1	1.30	1.30	1.30	1.30	1.50	15.4	15.4
$BaBiBS_4$	1	2.34	2.34	2.34	2.34	1.98	-15.4	-15.4
$NaSc(SeO_3)_2$	1	5.50	5.50	5.50	5.50	4.56	-17.1	-17.1
$RbBaPS_4$	1	3.30	3.30	3.30	3.30	2.70	-18.2	-18.2
$Ba_3Er_2(PS_4)_4$	1	3.30	3.30	3.30	3.30	2.69	-18.5	-18.5
$SbPbIO_2$	1	2.48	2.48	2.48	2.48	1.98	-20.2	-20.2
$Zn_3(PS_4)_2$	2	3.07	3.19	3.13	3.13	2.49	-20.4	-20.4
$AgBi(PSe_3)_2$	1	1.40	1.40	1.40	1.40	1.11	-20.7	-20.7
$NaB_5(H_2O_5)_2$	2	5.61	6.13	5.87	5.87	7.13	21.5	21.5
$Cs_2Mg_2(WO_4)_3$	1	4.53	4.53	4.53	4.53	3.46	-23.6	-23.6
$KP(HO_2)_2$	3	3.20	7.00	5.72	6.95	7.09	24.0	2.0
$Ti(Bi_3O_5)_4$	1	3.09	3.09	3.09	3.09	2.33	-24.6	-24.6
$LiZnPS_4$	1	3.44	3.44	3.44	3.44	2.57	-25.3	-25.3
$NaYb(PS_3)_2$	1	1.85	1.85	1.85	1.85	2.37	28.1	28.1
FeS_2	9	0.02	1.25	0.83	0.92	1.17	41.0	27.2
$Bi_4Pb_7Se_{13}$	4	0.23	0.29	0.26	0.26	0.14	-46.2	-46.2
$Hg_8Bi_3As_4Cl_{13}$	1	4.30	4.30	4.30	4.30	2.04	-52.6	-52.6
$SrBe_2(BO_3)_2$	1	4.69	4.69	4.69	4.69	7.32	56.1	56.1

Continued on next page

Table 3.5: Band Gap - Examples of Input Chemical Composition and the Corresponding Prediction of Band Gap (E_g), Sorted by the Absolute Percentage Difference From the Mean Experimental Values in Ascending Order. (Continued)

$Ba_8U_2PdSe_{16}$	2	0.18	1.60	0.89	0.89	1.43	60.7	60.7
As	10	0.17	1.25	0.80	1.14	1.38	72.5	21.1

[a]Data from ref [173]. Here, n is the number of experimental values sampled; min and max represents their range; mean and median are their corresponding descriptive statistics[a].

Chapter 4

Property Prediction of Inorganic Materials: Alloys and Superconductors

This chapter expands the scope of machine learning-based property prediction for inorganic materials. Building on the demonstrated success of our workflow in predicting the Curie temperature of ferromagnetic materials and the band gap of inorganic compounds, we now apply it to more complex material systems. Specifically, we present predictive results for key characteristics of high-entropy alloys and amorphous metallic alloys (Section 4.1), as well as the critical temperatures of superconductors (Section 4.2). These extensions highlight the versatility and broader applicability of our approach across diverse classes of inorganic materials.

This chapter is based on:

1. **Jung, S. G.**, Jung, G., Cole, J. M., Predictive Modeling of High-Entropy Alloys and Amorphous Metallic Alloys Using Machine Learning. *J. Chem. Inf. Model.*, 2024, 64 (19), 7313–7336;

2. **Jung, S. G.**, Jung, G., Cole, J. M., Machine-Learning Predictions of Critical Temperatures from Chemical Compositions of Superconductors. *J. Chem. Inf. Model.*, 2024, 64 (19), 7349–7375.

Prof. Jacqueline Cole and I conceived the overarching project and designed the studies. I developed the workflow, performed the data acquisition and featurisation, the statistical analyses, the model training and fine-tuning, and analysed the data under the PhD supervision of Prof. Jacqueline Cole. Guwon Jung assisted with data

gathering, model development and optimisation, and contributed to the analysis of the results. All figures and tables are reprinted with permission from J. Chem. Inf. Model., Copyright 2024 American Chemical Society.

4.1 Predictive Modelling of High-Entropy Alloys and Amorphous Metallic Alloys

4.1.1 High-Entropy Alloys

Case Study One: Bulk Modulus Prediction

We begin by evaluating the efficacy of our GBFS workflow for HEAs in comparison to a variety of ML strategies that employed the aforementioned data set. Specifically, the comparative analysis focuses on the prediction of the bulk modulus (K) for HEAs, categorising the alloys into equimolar and non-equimolar groups.

Table 4.1 compares the performance of various methods in predicting K for both the equimolar and non-equimolar compositions. We compared our GBFS results against those afforded using the following ML methods: Deep Sets [174], gradient boosting decision trees (GBDTs) [40], k-nearest neighbour (KNN) [175], linear regression (LR) [176], random forest (RF) [177], and support vector machine (SVM) [178, 179]. We observe lower MAEs in predicting K for equimolar quaternary HEAs in contrast to their non-equimolar counterparts, suggesting that property predictions pose greater challenges for the latter group. Among the evaluated predictive methodologies, LR and KNN recorded the highest MAEs, aligning with expectations that linear or simpler models would yield suboptimal performance. Conversely, SVM, RF, GBFS and Deep Sets methods demonstrated enhanced predictive accuracy.

Notably, GBFS and Deep Sets methods significantly surpassed other ML-based approaches, with our GBFS workflow achieving the lowest MAE during 10-fold cross-validation for equimolar quaternary HEAs and the second lowest MAE for non-equimolar systems. More specifically, our GBFS approach demonstrated remarkable predictive accuracy for equimolar systems, recording an MAE of 4.196 ± 0.432 GPa, an RMSE of 6.152 ± 0.865 GPa, and an R^2 of 0.958 ± 0.012 in a ten-fold cross-validation. Similarly, for non-equimolar systems, our GBFS approach achieved an MAE of 6.415 ± 1.071 GPa, an RMSE of 9.781 ± 2.575 GPa, and an R^2 of 0.914 ± 0.050. The apparent increase in the standard error in the latter case accounts for approximately 68% of the predictions in non-equimolar systems using the GBFS approach; this is attributable to

the inclusion of 132 additional quaternary non-equimolar HEA compositions in the cross-validation process. These distinct compositions were not included in the original training and validation sets, which accounts for the reduced standard errors observed in the alternative models. This exclusion highlights the impact of data set diversity on model accuracy and error metrics.

The results obtained from the regression analyses on the out-of-sample test set are depicted in Figure 4.1. The GBFS-based model demonstrated effective generalisation across unseen chemical compositions within both equimolar and non-equimolar systems. In the case of equimolar compositions, the GBFS-based model attained an MAE of 3.958 GPa, RMSE of 5.411 GPa, and an R^2 of 0.969. A linear regression, conducted via the Ordinary Least Squares (OLS) method, yielded a gradient of 1.0 and a y-intercept of 8.0, revealing a slight positive bias in the predictions for lower values of K. Nonetheless, this systemic bias is considered minimal within the observed range of K values for HEAs. The distribution of absolute errors further indicates that ca. 63% of predictions exhibit an error under 4 GPa, ca. 35% display an error under 2 GPa, and ca. 19% maintain an error below 1 GPa.

For the non-equimolar compositions, the predictive model achieved an MAE of 5.034 GPa, an RMSE of 7.446 GPa, and an R^2 of 0.946 on the out-of-sample test set. A linear regression analysis resulted in a gradient of 0.9 and a y-intercept of 9.8, indicating a mild positive bias for predictions at lower K values and a slight underestimation for higher K values. Similar to the equimolar scenarios, this systemic bias is deemed relatively small within the range of K values considered. The distribution of absolute errors reveals that ca. 42% of predictions have an error below 4 GPa, ca. 25% are under 2 GPa, and ca. 10% fall below 1 GPa. Reflecting expectations, the non-equimolar compositions exhibit greater uncertainty, with a smaller proportion of predictions achieving errors that are below these thresholds in comparison to their equimolar counterparts.

Extending this comparative analysis to encompass ML models reported in other studies presents challenges owing to the variation in the training sets. For instance, the AFLOW-ML and JARVIS-ML models reported MAEs of 8.68 GPa and 10.5 GPa, respectively, in their predictions of the bulk modulus for inorganic structures [66, 180]. Moreover, our prior investigation, which focused on predicting the Voigt–Reuss–Hill approximation for the bulk modulus of isotropic polycrystalline materials, achieved an MAE of 1.167 GPa [123]. This notably lower MAE is ascribed to the larger data set size and the integration of structural features in that study, while in the current analysis, our feature selection is strictly confined to attributes that stem from chemical

composition, in addition to exclusively focusing on compositions that are relevant to the domain of HEAs.

Table 4.1 MAE for the prediction of bulk modulus (K) for quaternary HEAs.

Method	MAE (GPa)	
	Equimolar quaternary HEAs	Non-equimolar quaternary HEAs
GBFS	4.196 ± 0.432	6.415 ± 1.071
Deep Sets	4.596 ± 0.639	6.025 ± 0.415
GBDTs	6.033 ± 0.349	8.622 ± 0.485
KNN	14.953 ± 0.620	12.508 ± 0.424
LR	7.891 ± 0.461	11.179 ± 0.414
RF	7.488 ± 0.574	9.607 ± 0.531
SVM	5.829 ± 0.536	7.636 ± 0.506

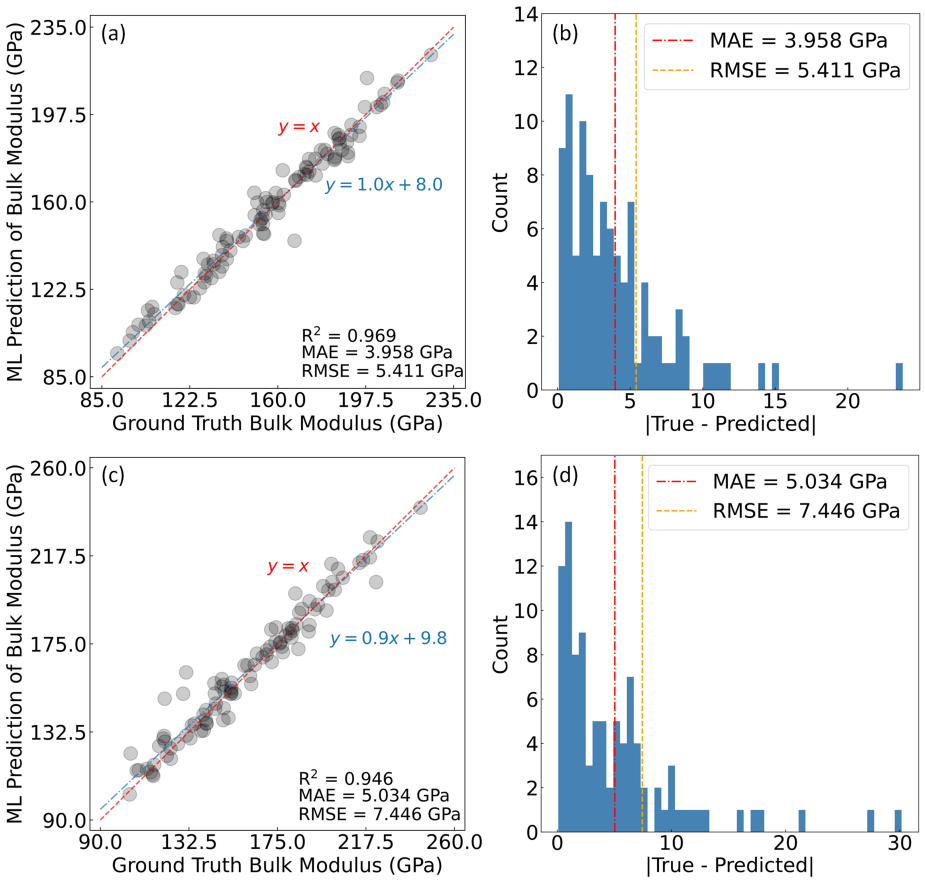


Fig. 4.1 Regression of the ML-based predictions of HEA bulk modulus (K) against literature values in the out-of-sample test set and the corresponding distribution of absolute errors. The subfigures (a) and (b) are associated with equimolar quaternary HEAs, whereas (c) and (d) are associated with non-equimolar quaternary HEAs. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the literature values. The blue dot-dash line is a linear fit generated using the OLS method. Within the error distribution plots, the red dotted line symbolises the MAE, and the orange dashed line denotes the RMSE.

Feature Interpretation

Figure 4.2 (b) details the five most influential features in our regression analysis of the bulk modulus. We now seek to rationalise their significant role. The most salient feature, as denoted by the realised total loss reduction, pertains to the mean bulk modulus estimate derived from the constituent chemical elements of an alloy. The identification of this feature having the highest relevance was anticipated, considering its direct material association with the measure of intrinsic elasticity or the resistance of a material to bulk compression. Insights from the beeswarm plot in Figure 4.3 (b) reveal that higher values of this feature generally align with more positive SHAP values, corroborating our expectations regarding its predictive significance.

The prominence of the mean FERE correction across the constituent elements of an alloy as the second salient feature requires further explanation. The FERE model is used in computational materials science to modify or adjust the reference state energies of chemical elements to better align with experimental values. This correction helps improve the accuracy of thermodynamic models by providing more realistic assessments of formation energies in alloys [181]. For instance, in the context of predicting properties such as the bulk modulus, the accuracy of elemental energies is essential. Incorporating the mean FERE correction as a feature ensures that the underlying thermodynamic calculations for predicting mechanical properties are anchored to well-adjusted, realistic energy states. This leads to more accurate predictions of the bulk modulus. The use of FERE-corrected energy values, therefore, enhances the reliability of these assessments, ensuring that the materials will behave as expected under mechanical stresses.

The range of electrical resistivity among constituent elements was identified as the third most relevant feature. Though not immediately apparent, the connection between electrical resistivity and the bulk modulus holds considerable importance for several reasons. Electrical resistivity offers insights into the electronic structure and bonding characteristics of a material. Materials characterised by strong metallic bonding typically exhibit lower resistivity owing to the unimpeded flow of electrons. This type of bonding directly impacts mechanical properties, such as the bulk modulus, since stronger bonds usually correlate with a stiffer material, thereby enhancing the bulk modulus. Therefore, variations in resistivity can serve as indirect indicators of changes in the bulk modulus. Moreover, high resistivity in metals and alloys is often linked to increased defects or impurities, which can adversely affect mechanical properties. These defects and impurities disrupt the regular atomic arrangement of a material structure and compromise its ability to compress uniformly under stress, which can result in a reduced bulk modulus.

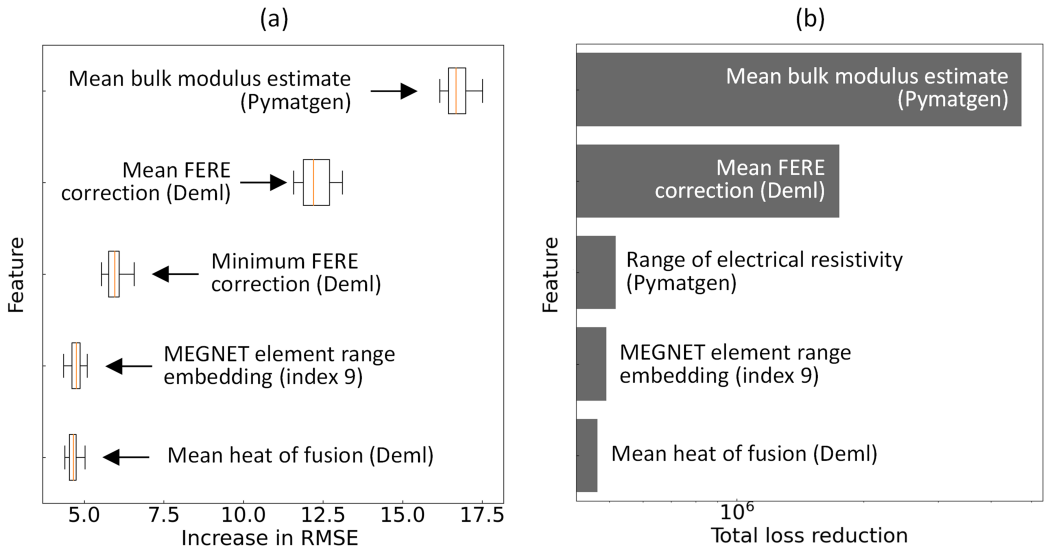


Fig. 4.2 (a) Permutation-based feature importance and (b) feature relevance plots for the regression analysis of K in equimolar quaternary HEAs, displaying the five most significant features.

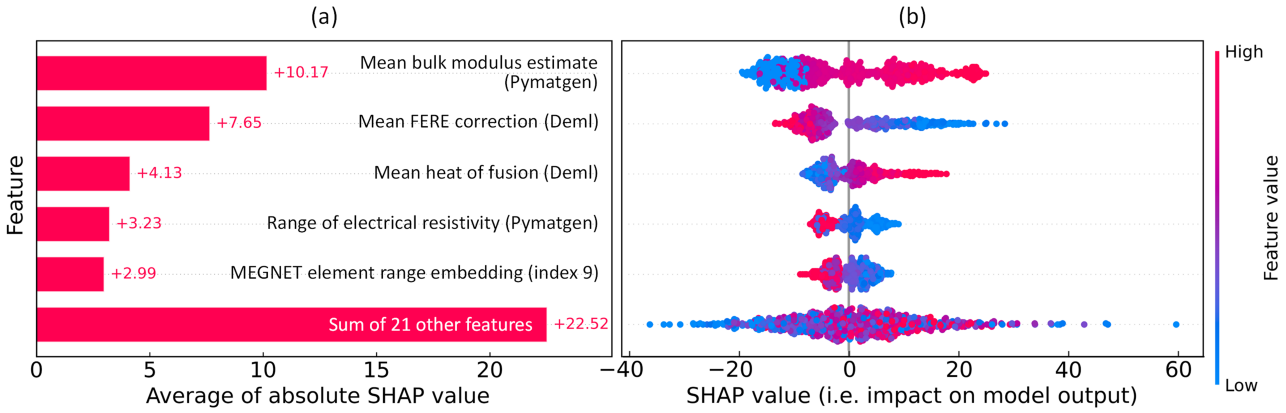


Fig. 4.3 The results based on the SHAP framework: (a) the average contribution (i.e., the mean absolute SHAP value) of five features that are identified as having the greatest contributions to the model output. A positive SHAP value indicates a positive contribution to the prediction of K values. (b) The beeswarm plot illustrates the impact of these features on the model output by plotting each instance as a single data point together with the SHAP value on the x-axis, where the y-axis is consistent with (a). The colour scheme corresponds to the original feature value and the broadening shows the density of instances (cf. the density plot).

We have previously discussed the challenges in interpreting MEGNet element embeddings. To gain clarity on how feature selection is influenced when these embeddings are omitted, we employed our GBFS workflow to train an alternative ML model. This analysis identified the average electronegativity among the constituent elements as one of the top five most relevant features. Electronegativity differences among atoms within a material are critical in determining the nature of bonding — ionic, covalent, or metallic. This, in turn, significantly impacts the mechanical properties of materials. For instance, materials characterised by ionic or covalent bonding, which involve atoms with greater electronegativity differences, tend to have higher bulk moduli because the strongly directional nature of these bonds effectively resists compression. Conversely, materials whose neighbouring atoms exhibit similar electronegativity usually form metallic bonds. These materials display varying bulk moduli in line with the density of the electron clouds and the nature of bonding involved. Such variations in bonding characteristics directly affect how internal stresses are distributed when external forces are applied to a material, thereby influencing the material's response to compression and its overall level of mechanical stability.

Predicting a Wide Range Of HEA Properties

With the foundations of our GBFS approach assured through its demonstrational efficacy in predicting K values, we sought to predict other properties for HEAs. These properties were derived from the cubic elastic constants through mathematical operations. For instance, the arithmetic Hill average for calculating the polycrystalline shear modulus (G) is given by

$$G = \frac{G_V + G_R}{2}, \tag{4.1}$$

where G_R and G_V represent the Reuss and Voigt shear modulus bounds, respectively. These bounds are expressed as

$$G_V = \frac{C_{11} - C_{12} + 3C_{44}}{5}, \tag{4.2}$$

$$G_R = \frac{5(C_{11} - C_{12})C_{44}}{4C_{44} + 3(C_{11} - C_{12})}, \tag{4.3}$$

with C_{11}, C_{12}, and C_{44} being the cubic elastic constants. Additional mechanical properties such as Young's modulus (E), Poisson's ratio (v), the Zener anisotropy

ratio (A_z), and Pugh's ratio (k) are derived through the following equations:

$$E = \frac{9KG}{3K + G}, \tag{4.4}$$

$$v = \frac{3K - 2G}{2(3K + G)}, \tag{4.5}$$

$$A_z = \frac{2C_{44}}{C_{11} - C_{12}}, \tag{4.6}$$

$$k = \frac{K}{G}, \tag{4.7}$$

noting that the crystal lattice is considered elastically isotropic when $A_z = 1$ and, in the context of cubic polycrystalline structures, the Voigt and Reuss bounds for K converge to the same value [182]; thus, $K = K_V = K_R$.

The additional properties examined in this study were sourced from a data set created using high-throughput Density Functional Theory (DFT) calculations. These calculations used exact muffin-tin orbitals and coherent potential approximation (EMTO-CPA) methodologies [183–185]. Our analyses exclusively employed features derived from chemical composition, in the same fashion as our predictions of K. For each target property, we processed the exploratory features through the GBFS workflow. This process facilitated the identification and selection of a distinct subset of features, which ultimately culminated in the development of a dedicated predictive model for each property evaluated. The regression outcomes on the out-of-sample test set are depicted in Figure 4.4. The result pertaining to the classification of HEAs based on their experimentally observed structural symmetries is presented in Table 4.2, where the microstructure is classified into face-centred cubic (FCC), body-centred cubic (BCC) or other atomic arrangements.

The GBFS workflow evidences its broad utility in predicting 13 diverse properties of HEAs based solely on features derived from their chemical composition (Figure 4.4). The comparison between GBFS-based predictions of these properties and values reported in the literature demonstrates a high degree of agreement, with the minimum R^2 value being 0.827 for the prediction of the Pugh ratio. Furthermore, the distribution of absolute errors predominantly aligns towards the lower spectrum of the error scale, with a subsequent reduction in the error counts at higher values. In the case of classifying HEAs based on their experimentally observed structural symmetries, we obtained an AUC-ROC of 0.97, an F1-score of 0.88, and an accuracy of 0.88. These observations underscore the predictive precision of our GBFS workflow.

Additionally, it is helpful to highlight the variance observed between the directly predicted K values and those deduced from the predicted elastic constants. Zhang et al. [185] observed that the maximum discrepancy between these approaches was about 8%, a level of variance deemed permissible considering the intrinsic uncertainties in DFT-calculated elastic constants, which can diverge from experimental measurements by up to ±15% [186]. Accordingly, this substantiates the application of the mathematical operations to approximate material properties using Equations 4.1 - 4.7, formulated based on the predicted cubic elastic constants. Verifying the predicted elastic constants through the GBFS workflow confirms the mechanical stability criteria, as evidenced by observing the following criteria: $C_{11} > 0$, $C_{44} > 0$, $C_{11} - C_{12} > 0$, and $C_{11} + 2C_{12} > 0$. These findings are consistent with the criteria established by the EMTO-CPA methodology.

The analysis of chemical compositions identified as outliers for the properties of K, C_{11}, C_{12}, and C_{44} offers valuable insights into potential patterns or specific elements that may be contributing to deviations from expected values. The initial observation is that the majority of these compositions pertain to quaternary non-equimolar HEAs. This high level of elemental diversity and complexity could lead to greater variability in material properties, influenced by differences in atomic sizes, bonding characteristics, and electronic structures, which are critical in determining mechanical properties. Additionally, the prediction of these properties is based solely on features derived from their chemical compositions, making such greater variability in results expected.

In examining the presence of specific elements within the outlier chemical compositions, it is notable that Tungsten (W) and Hafnium (Hf) frequently occur. These heavy transition metals are likely to influence the mechanical properties of the alloys, possibly due to their high density and strong bonding characteristics. Additionally, Aluminium (Al) is another element commonly found in these outlier chemical compositions, which is known for its lightweight nature and corrosion resistance. However, its presence in high proportions might be unpredictably influencing the elastic properties, especially when combined with heavier or more brittle elements. These observations suggest that including such metals could be pivotal in dictating the mechanical behaviour of the alloys.

In the context of predicting the K values, the presence of several outliers, particularly those with high W content, suggests that K is sensitive to the incorporation of heavy metals. This is potentially due to their inherently high modulus of elasticity. Nevertheless, the outliers encompass a variety of alloying elements, including chromium (Cr), niobium (Nb), titanium (Ti), cobalt (Co), copper (Cu), and manganese (Mn).

This high level of elemental diversity indicates that there is no single dominant pattern affecting the bulk modulus; rather, it appears that the combined interactions of multiple elements within a given alloy will significantly influence its mechanical properties.

In the analysis of elastic constants, a notable level of diversity in transition-metal composition is observed. Many outliers are characterised by a combination of transition metals, such as chromium (Cr), manganese (Mn), cobalt (Co), and nickel (Ni), which are recognised for their contribution to strengthening mechanisms but may also introduce greater variability in directional elastic properties, such as C_{11}, C_{12}, and C_{44}. Moreover, the presence of minor percentages of elements, such as zirconium (Zr), hafnium (Hf), and vanadium (V) in certain alloys, correlates with these outliers, indicating that even trace components can profoundly influence elastic constants through their effects on crystal structure and bonding. In particular, the predictions of C_{11} and C_{44} show heightened sensitivity to combinations of Al with transition metals or heavier elements, reflecting the susceptibility of these elastic constants to lattice distortions or variations in atomic packing. Conversely, C_{12} displays sensitivity to a variety of elemental mixtures, without a predominant single-element influence, suggesting that the overall alloying strategy and the ratios of elements are critical.

It is evident that the presence of heavy transition metals in complex multi-component systems markedly influences mechanical properties, with elements such as W, Hf, and Al playing crucial roles. The observed variability in properties such as elastic modulus and elastic constants across these alloys can likely be attributed to the intricate interactions among these elements within the alloy's crystal structure, potentially impacting bonding characteristics and electronic interactions. This level of complexity underscores the need for a nuanced understanding of how the elemental composition of advanced alloys affects their mechanical behaviour.

Upon extending our examination to include additional properties that are depicted in Figure 4.4, we identified consistent chemical profiles and trends among the outliers in the prediction of various properties, including Young's modulus, total energy, Zener anisotropy, Pugh ratio, Poisson ratio, and universal anisotropy. We noted a prevalent use of Al and a combination of heavy transition metals, rare earth elements, or refractory metals such as Hf, Zr, Ti, Nb, W, and V. The majority of these outliers were linked to non-equimolar HEAs.

Further discussion is necessary to better understand the model predictions of various properties by concentrating on specific regions of interest relevant to different applications. This approach is essential because an accurate representation of the model performance can be observed by focusing on a region of particular interest for

a given property. Analysing whether or not predictions within this defined interval exhibit lower prediction errors will provide deeper insights into the effectiveness of our ML models. Generally, the performance metrics that were calculated (i.e., R^2, MAE, and RMSE) reliably reflect the accuracy of our models across their entire respective ranges, demonstrating minimal or consistent deviations from the actual values for target properties such as elastic constants, Young's modulus, shear modulus, Wigner-Seitz radius, formation enthalpy, and total energy. Nevertheless, this consistency does not extend to properties such as Zener anisotropy, universal anisotropy, Pugh ratio, and Poisson ratio. Therefore, it is crucial to engage in further discussion regarding these properties to better understand the efficacy of our modelling approach.

Specifically, the Zener anisotropy or ratio, which is a dimensionless number used to quantify anisotropy in cubic crystals, indicates the extent to which a material deviates from isotropic behaviour, with a value of 1 representing perfect isotropy. For HEAs, a Zener ratio close to 1 is therefore desirable. This ratio signifies isotropic behaviour, indicating that the properties of the material will exhibit uniformity in all directions. A Zener ratio of 1 implies that the alloy will maintain consistent mechanical properties irrespective of the direction in which stress is applied, making it ideal for applications that demand uniform behaviour across various orientations. In Figure 4.4, a noticeable deviation between the ground truth and the predictions is evident for Zener anisotropy values greater than 5, which is attributable to the scarcity of materials within this range. As a result, the accuracy of our model decreases when designing materials with high anisotropy. Nonetheless, since the majority of materials fall within the desirable isotropic value of 1, the deviations observed at higher values do not significantly undermine the overall effectiveness of this predictive exercise.

The Zener anisotropy or ratio is applicable only to cubic crystals. To address this limitation, the universal anisotropy index was formulated, which is based on the variational principles of elasticity and tensor algebra. This index allows for the quantification of anisotropy in elastic crystals across all classes. For HEAs, a desirable universal anisotropy value is close to zero. A value of zero indicates isotropic behaviour, meaning that a material has uniform elastic properties in all directions, which is typically preferred for various applications, as previously mentioned. Similar to the Zener anisotropy, the prediction performance shows noticeable deviations from the ground truth for universal anisotropy values approximately beyond 3. This is expected given that the majority of data points fall below this range. Nevertheless, since the desirable value is zero, the deviations observed at higher values do not substantially impact the overall efficacy of this predictive exercise.

The Poisson ratio is defined as the ratio of transverse strain to axial strain when a material is subjected to stretching or compression. It offers important insights into the ability of a material to undergo dimensional changes that are perpendicular to the load direction. A higher Poisson ratio, nearing the theoretical limit of 0.5 for isotropic materials, signifies greater ductility, allowing the material to stretch or compress with minimal volumetric changes. For HEAs, the optimal Poisson ratio typically falls between 0.3 and 0.35. This range is preferred in engineering applications due to its balance between ductility and volume conservation during deformation. In HEAs, maintaining a Poisson ratio within this specified range is considered to be ideal, as it enables the alloys to withstand mechanical stresses without undue deformation or failure. Once again, we observe that large deviations occur outside the targeted range, especially below a value of approximately 0.275.

The Pugh ratio, defined as the ratio of the bulk modulus to the shear modulus, demonstrates that a value greater than 1.75 indicates increased ductility. This characteristic is generally beneficial for applications that require materials to withstand substantial deformation without fracturing. Although there is no strict upper limit to the Pugh ratio that categorically defines it as undesirable, the suitability of a higher or lower Pugh ratio greatly depends on the specific application and required performance characteristic of a given alloy. While Pugh ratios significantly above 1.75 are indicative of enhanced ductility, they may also lead to reduced strength and hardness. Therefore, in situations where greater strength and hardness are essential, an exceedingly high Pugh ratio may not be advantageous. That said, similarly to the other properties discussed, we observe that relatively lower errors occur within the range of values typically considered suitable for HEAs.

Table 4.2 Summary of the performance metrics for the classification of HEAs by their structural symmetries.

	Precision	Recall	F1-score
BCC	0.913	0.894	0.903
FCC	0.872	0.774	0.820
Other	0.871	0.914	0.892
Macro average	0.885	0.860	0.872
Weighted average	0.884	0.883	0.883

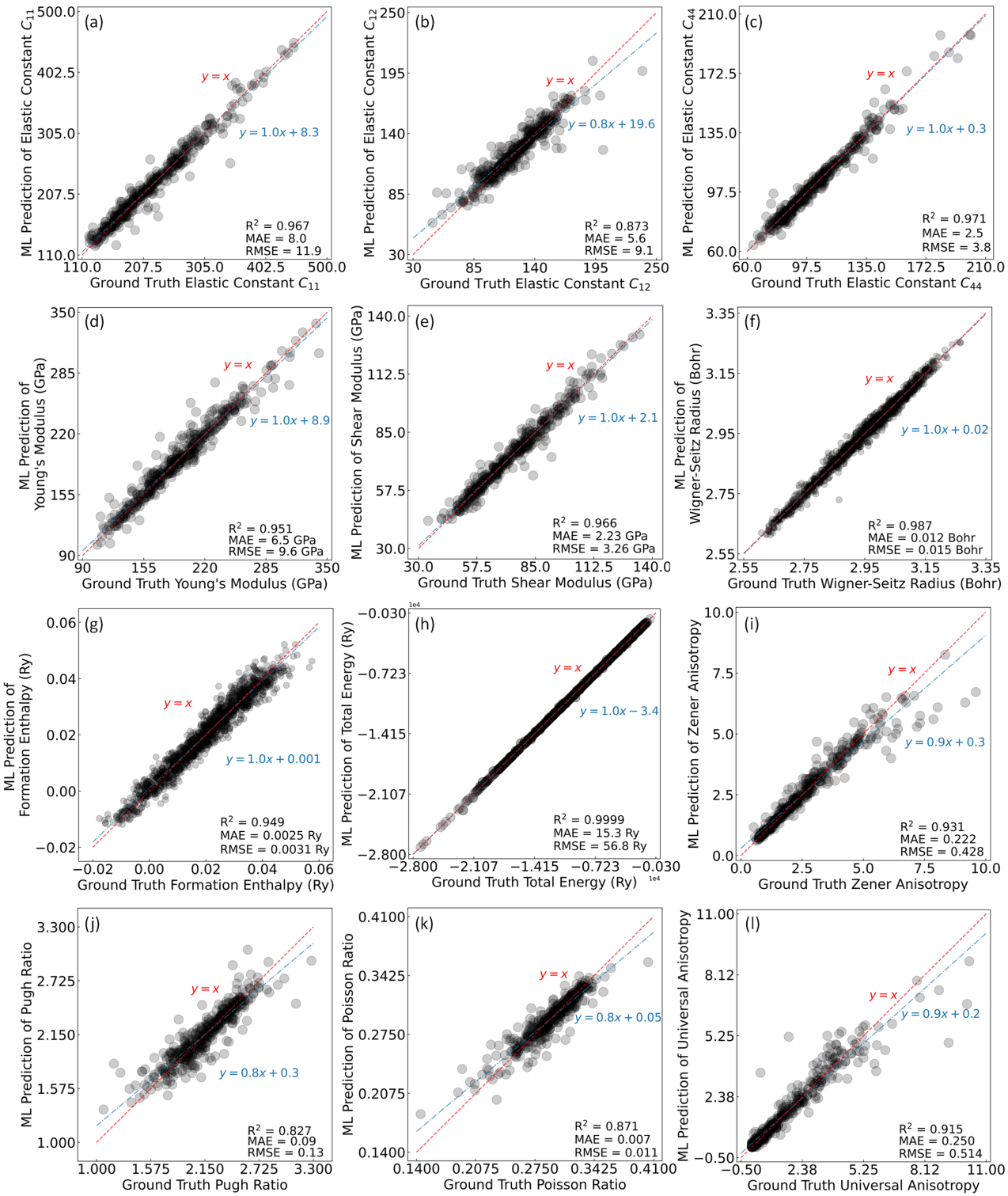


Fig. 4.4 Property predictions for HEAs. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the literature values. The blue dot-dash lines represent linear fits generated using the OLS method, applied to various materials properties: (a) elastic constant C_{11}, (b) elastic constant C_{12}, (c) elastic constant C_{44}, (d) Young's modulus, (e) shear modulus, (f) Wigner-Seitz radius, (g) formation enthalpy, (h) total energy, (i) Zener anisotropy, (j) Pugh ratio, (k) Poisson ratio, and (l) universal anisotropy.

Blind Test on Quinary HEA Composition

The aforementioned predictions of various material properties in comparison with values reported in the literature demonstrate notable statistical merit. However, the validation of these results extends beyond mere statistical performance; it enables an evaluation against specific, well-characterised HEA materials. We now showcase this model capability by performing a blind test study on the extensively studied quinary HEA system $NiCoFeCrAl_x$ [187], whose compositional range was previously unseen by the model; i.e., it is an out-of-sample test case.

This analysis emphasises the prediction of K values and the Wigner-Seitz (W-S) radius, with the outcomes being presented in Table 4.3 together with the comparable results from the Deep Sets ML-based prediction and DFT-calculated values that employed the EMTO-CPA methodology. The outcomes demonstrate a high level of agreement, which is particularly notable given that the GBFS-based model received no training on quinary HEA compositions. An auxiliary analysis of mean absolute deviation reveals that the GBFS workflow exhibits lower variance with respect to literature values when compared to the Deep Sets ML-based model. Specifically, regarding the prediction of K values, the mean absolute deviation is lower by ca. 46%, and for the W-S radius, this deviation is observed to be an order of magnitude lower between the two ML methodologies.

Table 4.3 The predicted Wigner-Seitz (W-S) radius and bulk modulus (K) for the $NiCoFeCrAl_x$ quinary system, as obtained through the GBFS workflow, are juxtaposed with findings derived from Deep Sets and EMTO-CPA calculations for the preferred structural symmetry.

x	K (GPa)			W-S Radius (Bohr)			Symmetry
	GBFS	Deep Sets	EMTO-CPA	GBFS	Deep Sets	EMTO-CPA	
0.1	198	199	200	2.622	2.620	2.611	FCC
0.25	197	195	197	2.627	2.628	2.619	FCC
0.3	197	194	196	2.633	2.631	2.622	FCC
0.375	194	192	194	2.634	2.635	2.626	FCC
0.5	188	189	190	2.637	2.641	2.632	FCC
1.25	171	166	171	2.671	2.680	2.667	BCC
1.3	170	166	170	2.673	2.678	2.670	BCC
1.5	163	161	167	2.678	2.689	2.675	BCC
2	159	151	159	2.692	2.706	2.690	BCC
2.5	144	145	153	2.711	2.721	2.701	BCC
Mean absolute deviation	1.8	3.9		0.0065	0.0116		

4.1.2 Amorphous Metallic Alloys

Having demonstrated the utility of our GBFS workflow in predicting HEA properties, we now direct our attention to the analogous application of GBFS to another type of alloy, AMAs. Thereby, we examine their glass-forming ability, treating it as a classification problem, as opposed to the regression problem that we showcased for HEA property predictions. This technically distinguishes the AMA investigation as a distinct case study.

Case Study Two: Glass-Forming Ability

The glass-forming ability of an alloy characterises its potential to solidify into an amorphous state, such as a bulk metallic glass (BMG). Alternatively, it may solidify into an amorphous ribbon (referred to as a ribbon metallic glass, or RMG), or a crystalline alloy (CRA). The materials data set of glass-forming ability encompasses a total of 6,471 unique alloy compositions, and includes 1,211 BMGs, 1,552 CRAs, and 3,708 RMGs. In instances where multiple reports on the glass-forming ability of specific alloys arise, the highest reported glass-forming ability was consistently selected for inclusion in our analysis.

The classification model we developed using the GBFS workflow demonstrated remarkable efficacy, outperforming several benchmarks in the literature. From an initial pool of ca. 650 exploratory features and ca. 130 engineered features, the model identified the 25 most pertinent features. It achieved an F1-score of 0.91, an AUC-ROC of 0.98, and an accuracy of 0.91. The extent to which the glass-forming abilities of these alloys have been correctly classified is delineated by the confusion matrix shown in Figure 4.5, with a summary of the corresponding precision, recall and F1-score evaluation metrics being provided in Table 4.4, while some of the most salient features identified by the GBFS workflow are given in Table 4.5.

Moreover, when applying a ten-fold cross-validation, our ML model realised an accuracy of 0.91 ± 0.01, a weighted AUC-ROC of 0.97 ± 0.01 and a weighted F1-score of 0.91 ± 0.01. In comparison, Xiong et al. [188] reported an accuracy of 89.52% utilising 36 features with a random-forest algorithm on the identical data set. This demonstrates the efficacy of our classification model which was developed via the GBFS workflow. Xiong et al. reported that the majority of their model performance is likely attributable to just 6 features. Our findings corroborate this, as we observed that the most significant loss reduction or variance gain was achieved with the first 5 features, particularly the top three, as depicted in Figure 4.6 (b). To explore this further,

we retrained our predictive model using only the first 6 features listed in Table 4.5, highlighting the most relevant features that are identified by our GBFS workflow. This method resulted in a classification accuracy of 89%, indicating a clear consistency between the two studies. Therefore, it is evident that our modelling approach is on par with state-of-the-art methodologies.

Precision quantifies the ratio of true positives relative to all positive predictions made by the model. For instance, a precision of 0.987 for BMGs implies that out of all the instances predicted as BMGs by the model, 98.7% were actually BMGs. Conversely, recall reflects the fraction of actual positive instances accurately identified by the model, with a recall of 0.946 for BMGs indicating that 94.6% of real BMGs were identified. The F1-score serves as an integral measure, amalgamating precision and recall into a singular metric via their harmonic mean. A high F1-score, approaching 1, denotes a perfect equilibrium between precision and recall. Notably, our F1-score of 0.966 for BMGs denotes both high precision and recall, which is indicative of its robustness in being able to predict the glass-forming abilities of AMAs. For RMG classifications, the precision, recall, and F1-score surpass the 0.9 threshold, attesting to the highly discriminative nature of our classifier. While the performance metrics for CRA classification are adequate, they are the lowest among the metrics, hinting at a relatively higher incidence of both false positives and false negatives.

The corresponding macro and weighted averages reveal the discriminative prowess of our classification model. Thereby, the macro average, which computes average scores across all classes without regard to class size, indicates balanced, high-precision performance with a value of 0.921. In contrast, the weighted average takes into account the support of each class (i.e., the number of instances in each class). The weighted average precision of 0.913 indicates that the model maintains a high level of precision while accounting for the varying populations of different classes.

The proximity of the macro and weighted averages indicates that our model performs uniformly across the different classes. This high level of uniformity is likely attributable to our application of an over-sampling technique in the training set such that each class has equal representations. Thereby, we employed random over-sampling (ROS) of the training set to ensure equitable class representation by synthesising feature vectors that resemble the characteristics of the minority classes within the high-dimensional feature space. Our implementation of ROS drew on a smoothed bootstrap resampling technique that effectively mitigated potential learning biases. This technique introduces a perturbation to the over-sampled feature vectors by adding Gaussian noise that

can be regulated by a smoothing matrix. This prevents the overfitting of a model on specific values of a feature, which helps to improve the generalisation of the model.

Feature Interpretation

Our GBFS workflow identified the most salient feature as configuration entropy, which is a measure derived from the statistical distribution of different elements within an alloy. It quantifies the disorder associated with the arrangement of different types of atoms or ions in a material structure and is based on the probability of finding a particular element at a given atomic point. This is calculated by

$$S_{config} = -k_B \sum_{i=1}^{n} x_i ln(x_i), \tag{4.8}$$

where k_B is the Boltzmann constant, x_i is the mole fraction of the i-th component in the alloy, and n is the number of different elements in the alloy. Equation 4.8 quantifies the contribution of configurational disorder to the entropy of an alloy. This entropy is influenced by the diversity and distribution of constituent elements; specifically, it increases as the number of distinct elements rises and their distribution approaches equimolarity. In the context of BMGs, configurational entropy plays a pivotal role in facilitating the formation of amorphous structures. High configurational entropy, arising from a significant diversity of elements and their equitable distribution within the alloy, acts to suppress the crystallisation process during cooling. This suppression helps to maintain the alloy in a glassy state by obstructing the orderly arrangement of atoms that typifies crystalline structures. Furthermore, elevated configurational entropy enhances the thermodynamic stability of these amorphous phases. This relationship is articulated through the Gibbs free energy equation:

$$\Delta G = \Delta H - T\Delta S, \tag{4.9}$$

where G represents the Gibbs free energy, H denotes the enthalpy, T is the temperature, and S signifies the entropy. An increase in entropy leads to a reduction of the Gibbs free energy for the amorphous phase, thereby favouring its formation over a crystalline phase. This principle underscores the preferential formation of the amorphous state in materials with high configurational entropy, which correlates with a greater glass-forming ability. The level of heterogeneity in atomic sizes and bonding characteristics among multiple elements disrupt the crystal nucleation process and slow grain growth, further favouring the development of amorphous structures over crystalline structures.

Therefore, configuration entropy is pivotal in predicting the glass-forming ability of AMAs and ensuring their stability as non-crystalline solids.

The second salient feature was identified as the mismatch of local atomic radii, which relates to local variations in atomic sizes among the constituent elements of the alloy. In other words, it quantifies the extent to which the atomic radii of the elements that make up the material differ from each other, affecting the physical and chemical properties of the alloy. A significant mismatch in the atomic radii among the constituent elements disrupts the regular atomic packing required for the formation of crystalline phases. This disruption impedes the nucleation and growth of crystal lattices, thereby enhancing the propensity of the alloy to form an amorphous structure. In addition, the variability in local atomic radii among the constituent elements typically elevates the viscosity of the melt. This increased viscosity impedes the atomic diffusion and rearrangement necessary for crystal growth, thereby enhancing the glass-forming ability of the alloy. This phenomenon can be conceptualised within the framework of energy landscape theory, where variations in atomic sizes contribute to a complex energy landscape with more local minima. These minima can entrap the structure in a non-crystalline, amorphous state as it cools. Consequently, an understanding and ability to manipulate this feature can significantly impact the ability to produce materials with high glass-forming ability.

The third prominent feature contributing significantly to the reduction in loss within our model is the calculated enthalpy of mixing (ΔH) for an alloy, as determined by Miedema's model [189]. Miedema's model, a foundational semi-empirical methodology in materials science, forecasts the formation enthalpies of alloys. It incorporates variables such as disparities in atomic sizes, electronegativity, and electron densities of constituent elements to estimate the enthalpic changes when different metals amalgamate to form an alloy. This predictive capability is crucial for understanding and engineering the thermodynamic properties of new alloy compositions.

More specifically, Miedema's model, through the ΔH parameter, predicts the thermodynamic nature of mixing among various metallic elements, discerning whether the interaction will be exothermic or endothermic. A negative value of ΔH (i.e., exothermic mixing) generally indicates that the components of the alloy are likely to form a stable solution or an amorphous phase rather than segregating or crystallising into intermetallic compounds. This inherent stability is conducive to the formation of amorphous structures. Consequently, alloys characterised by more negative values of enthalpy of mixing typically exhibit a higher glass-forming ability, as these favourable energy states can facilitate the bypassing of crystallisation pathways during the cooling

process. Alloys characterised by a significantly negative enthalpy of mixing typically exhibit increased viscosity in their molten state. This increased viscosity restricts atomic mobility which is essential for the formation of crystalline structures, thereby increasing the likelihood of glass formation as the alloy cools. ΔH appears to be a critical parameter for predicting and understanding the glass-forming ability of alloys. It provides a theoretical basis for anticipating how different elements will interact thermodynamically, influencing the stability and formability of amorphous phases. See Table 4.5 for additional features that are deemed pertinent in the classification of alloys by their glass-forming ability.

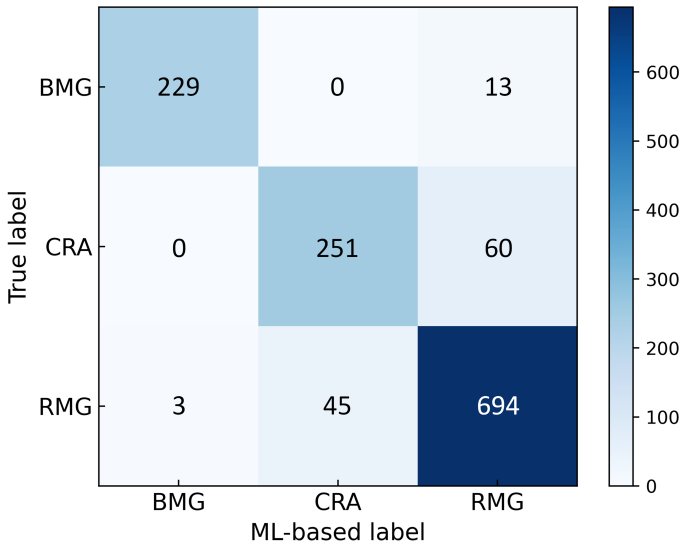


Fig. 4.5 Confusion matrix for the classification of alloys by their glass-forming ability.

Table 4.4 Summary of the performance metrics for the classification of alloys by their glass-forming ability.

	Precision	Recall	F1-score
BMG	0.987	0.946	0.966
CRA	0.869	0.810	0.839
RMG	0.907	0.945	0.925
Macro Average	0.921	0.900	0.910
Weighted Average	0.913	0.913	0.912

Table 4.5 List of most relevant features identified by our GBFS workflow.

No.	Feature description
1	Configuration entropy
2	Mismatch of local radii
3	Miedema's enthalpy of mixing
4	Average deviation in Goldschmidt's atomic volume per atom among the constituent elements of an alloy
5	Range of the Mendeleev number among the constituent elements of an alloy
6	Range of the coefficient of linear thermal expansion among the constituent elements of an alloy
7	Average deviation of electronegativity among the constituent elements of an alloy
8	Mean valence electron concentration (VEC)

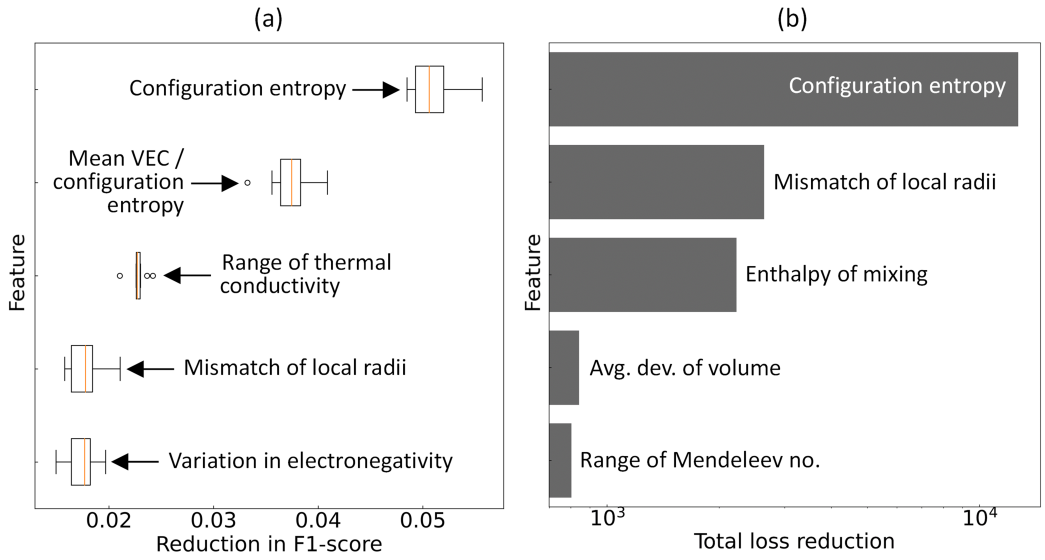


Fig. 4.6 (a) Permutation-based feature importance and (b) feature relevance plots plots for the classification analysis of alloys by their glass-forming ability, displaying the five most significant features.

Predicting Other Properties of AMAs

Having established a GBFS-based model for predicting the glass-forming ability of AMAs, we broadened the application of our modelling approach to create and train individual models that are capable of predicting a wider range of properties of AMAs; this includes characteristic transformation temperatures (CTTs), the critical casting diameter (D_{max}) and elastic moduli. The results are displayed in Figure 4.7.

The results pertaining to the prediction of CTTs are presented in Figure 4.7 (a) - (c), where T_g denotes the glass transition temperature, T_x the onset of crystallisation temperature, and T_l the liquidus temperature. CTTs are crucial for establishing criteria to evaluate the glass-forming ability of alloys [190]. The ground-truth values were derived from 674 measurements conducted via differential thermal analysis or differential scanning calorimetry at a consistent heating rate, primarily at 20 K/min. Different combinations of transformation temperatures can relate D_{max} to glass-forming ability criteria. Notably, this includes the supercooled liquid range ($\Delta T = T_x - T_g$) [190] and the reduced glass transition temperature ($T_{rg} = T_g/T_l$) [191, 192]. Hence, CTTs serve as critical parameters in the design and characterisation of AMAs.

The predictive outcomes for D_{max} are depicted in Figure 4.7 (d). D_{max} is the largest diameter or section thickness of an alloy that can be cast into a fully amorphous rod or plate. The ground-truth data set for D_{max} includes measurements for 5,934 alloys, incorporating only the critical copper-mold casting diameter values reported in the literature for BMGs. Generally, a slower cooling rate allows for a larger D_{max}, indicating a higher glass-forming ability, where the critical cooling rate is defined as the slowest cooling rate at which an alloy can solidify without crystallisation. Consequently, D_{max} serves as another important parameter for assessing the glass-forming ability of alloys.

The predictive results for the shear and bulk moduli of BMGs are presented in Figure 4.7 (e) - (f). The ground-truth data set comprises 278 unique BMGs, with their moduli measured using resonant ultrasonic spectroscopy. The mean value has been computed for cases where multiple measurements are reported in the literature. These elastic moduli, or the ratios thereof, are critical mechanical properties that influence the glass-forming ability and the overall behaviour of AMAs [193]. The shear modulus plays a pivotal role in theoretical models that assess the glass-forming ability of alloys. For instance, models such as the fragility index of glasses utilise the temperature dependency of the shear modulus to predict how easily a glassy state forms under varying cooling conditions. A lower shear modulus at high temperatures can indicate a greater likelihood of glass formation. Conversely, a higher bulk modulus indicates

a lower compressibility of the alloy, which typically correlates with increased density and potential stability of the glass. This enhanced stability is essential for preventing the crystallisation of the amorphous structure under various thermal and mechanical stresses.

Having broadened our analysis to encompass multiple target properties, it is crucial to conduct a comparative evaluation of our predictive capabilities against those documented in the existing literature. We align our results with state-of-the-art models reported in scholarly articles, employing identical data sets for these properties to facilitate a direct comparison.

In our examination of the elastic properties of AMAs, we conducted a detailed comparison with the findings of Xiong et al. [188]. Their work, an important benchmark in the field, reported an RMSE of 3.6 GPa and an R value of 0.98 for shear modulus predictions, alongside an RMSE of 9.5 GPa and an R value of 0.98 for bulk modulus predictions. These metrics highlight the high accuracy and robustness of their predictive analytics. Our ML models demonstrated comparable performance. Thereby, for the shear modulus, our models achieved an RMSE of 3.3 GPa and maintained an R value of 0.98 (or an R^2 of 0.97), indicating precise predictions with minimal variance. More notably, for bulk modulus predictions, our models showed an improved RMSE of 8.5 GPa and an R value of 0.99 (or an R^2 of 0.98). This indicates a statistically significant fit to the data, underscoring the efficacy of our modelling approach in accurately capturing the underlying patterns in the materials data. We also observe that the linear fits produced using the OLS method demonstrate strong consistency between the predicted values and the ground truth, as both target properties yielded a gradient of 1.0. There is a positive bias of 1.1 and 0.2 GPa at lower property values for shear and bulk modulus, respectively. However, the impact on the predictions appears minimal, especially within the range of feature values considered in this analysis.

Shifting our attention to CTTs, which are pivotal in assessing the glass-forming ability of alloys, our models demonstrated remarkable accuracy. Notably, the highest R^2 value achieved was 0.994 for predicting T_g, whereas the prediction of T_l achieved a slightly lower, yet still impressive, R^2 value of 0.992. These high values of R^2 indicate excellent model performance in predicting key thermal transitions in alloy systems. Upon a closer analysis of the most influential exploratory features within these predictions, the mean melting temperature of the constituent elements in a given alloy consistently emerged as the most critical predictor of CTTs. This feature significantly contributed to loss reduction and variance gain, illustrating its central role in the models' ability to accurately predict CTTs. Next in importance was the mode of

the ground-state volume per atom among the constituent elements, which also played a significant role in the prediction process. Additionally, the standard deviation of the velocity of sound travelling through these elements proved to be another key feature, reflecting how acoustic properties correlate with various forms of thermal behaviour in materials. These findings align closely with those reported by Xiong et al. [188], who also highlighted the predominant influence of the mean melting temperature on CTTs, noting an R value greater than 0.93 in correlation with CTT target values. They similarly emphasised the importance of atomic radius, which aligns with our observations regarding the significance of the volume per atom. Such consistent results across different studies reinforce the reliability of these features as predictors of alloy behaviour, particularly in the context of glass-forming ability, and underscore the interconnected nature of thermal, acoustic, and volumetric properties in materials science.

The final target property we will discuss in this analysis is D_{max}. Our predictive model for this target property achieved an R of 0.92 (or an R^2 of 0.85) and an RMSE of 0.9 mm. The linear regression, executed using the OLS method, presented a gradient of 0.9 and a positive bias of +0.1. The scatter plot of this regression reveals a relatively dispersed set of data points compared to other properties that we have examined, illustrating variability in predictive accuracy across the data set. Notably, the majority of the D_{max} values were observed to be below approximately 9 mm. This distribution characteristic significantly contributes to the lower model accuracy for larger D_{max} values, where deviations become more pronounced. When comparing our results with those documented in the literature [188], where an RMSE of 1.2 mm and an R value of 0.85 were reported for the same target property, our predictions show close concordance. These results affirm the robustness of our ML models, which were systematically refined through our GBFS workflow without requiring any human intervention in selecting the feature space. This automated approach proves highly effective in capturing and predicting diverse material properties, emphasising the capability of ML to adapt to and accurately model complex patterns within materials data.

We further analysed the chemical compositions that are associated with the extended tails in the error distributions for CTTs and D_{max}. For T_g, numerous alloys prominently incorporate rare earth elements such as La, Pr, Gd, and Nd. These elements are strategically utilised in alloys to improve properties such as corrosion resistance and high-temperature strength. Additionally, transition metals such as Ni, Cu, and Zr, along with refractory metals such as Nb and Ta, are frequently included in various compositions. Aluminium is also prominently featured, indicating an intention to

exploit its lightweight characteristics and capacity to form amorphous structures when alloyed with larger atoms from rare earth elements.

In the analysis of T_x and T_l, we observe a similar chemical profile among the outliers, which is characterised by a significant use of rare earth elements (e.g., La, Ce, Pr, Ho and Y). These elements are used for their distinctive properties, including enhanced stability at high temperatures and resistance to oxidation, which are typical benefits of rare earth metals. For instance, chemical elements such as La and Ce are incorporated in alloys to refine their grain structure, thereby enhancing the mechanical properties of alloys. The presence of multiple principal elements in diverse ratios suggests an approach aimed at exploiting the synergistic effects among different metals. The inclusion of Al and Mg in various alloys highlights its role in reducing overall weight while simultaneously improving mechanical strength.

When examining the chemical compositions associated with the heavy tail in the error distributions for D_{max}, a consistent chemical trend emerges. This encompasses the use of rare earth and exotic metals such as Nd and Gd, along with Al and Mg to optimise strength-to-weight ratios. A notable feature is the prevalent use of Cu and Zr, which are important chemical elements in the formation of BMGs. Additionally, there is a frequent inclusion of Ag, which likely aims to enhance the durability of the alloys, while the inclusion of Be is likely due to its stiffness and thermal stability. Once more, the majority of these chemical compositions have been crafted with complex, multi-component formulae designed to exploit the synergistic effects of various elements to enhance overall properties. Yet, this high level of complexity can introduce challenges in accurately predicting their properties.

We now examine the model predictions by focusing on specific regions of interest that are relevant to various engineering applications, similar to our approach for the properties associated with HEAs. In general, the calculated performance metrics (i.e., R^2, MAE, and RMSE) provide a good indication of the model performance across their respective ranges, showing minimal or consistent deviations from the ground truth for the CTTs and elastic moduli. However, a different pattern emerges for the prediction of D_{max}, as illustrated in Figure 4.7 (d). Here, we observe a low prediction error within regions below 7 mm. In contrast, the model accuracy diminishes for predictions concerning D_{max} values greater than 7 mm, particularly when D_{max} exceeds 10 mm. This observation suggests that for applications that require materials with large D_{max}, the effectiveness of the model should be assessed based on its accuracy within this higher range. Therefore, it is imperative to evaluate the predictions by focusing on specific regions of interest for applications, beyond just the overall model accuracy.

For AMAs (or BMGs), the desired D_{max} value is ideally as large as possible. This critical measurement reflects the maximum thickness at which the alloy can be cast and still maintain its amorphous structure without undergoing crystallisation as it cools. This indicates that the model predictions may not be reliable for designing materials with a target D_{max} value that exceeds approximately 7 mm. It is important to make this limitation clear to the reader to ensure transparency regarding the capabilities of the model developed through our proposed GBFS workflow. The reduced performance at higher values can be attributed directly to the scarcity of data points available within this range. When directly comparing our results with those of Xiong et al. [188], who employed the same data set and established the benchmark for this study, we noted similar trends at higher values. Xiong et al. implemented various methodologies, from linear regression to random forest, achieving their best prediction of D_{max} with an R of 0.85 and an RMSE of 1.2 mm. In contrast, our model achieved an R of 0.92 (or R^2 of 0.85) and an RMSE of 0.9 mm. Specifically, our model exhibits comparatively smaller deviations in predicting D_{max} values greater than 7 mm. This observation suggests that there are added advantages to adopting a systematic approach to feature analysis, selection, modelling, and optimisation, as proposed in this study. With the availability of more data in this range, we anticipate significant improvements in both modelling approaches.

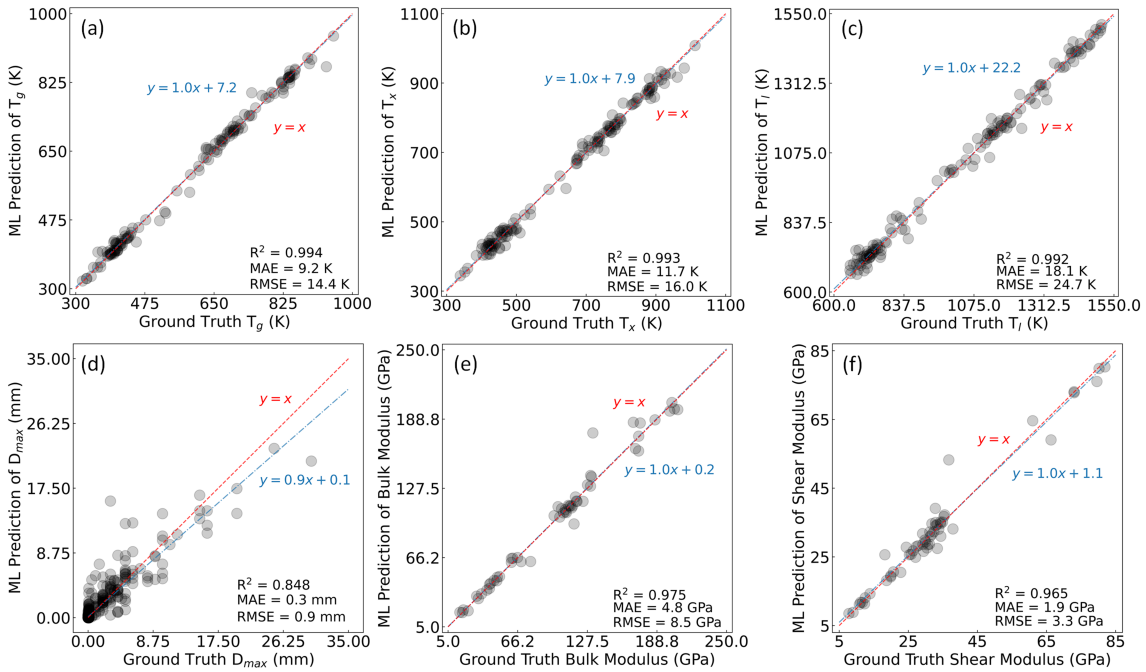


Fig. 4.7 Property predictions of AMAs. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the literature values. The blue dot-dash lines are linear fits generated using OLS method, applied to various materials properties: (a) glass transition temperature T_g, (b) onset of crystallisation temperature T_x, (c) liquidus temperature T_l, (d) critical casting diameter D_{max}, (e) bulk modulus, and (f) shear modulus.

4.2 Predictions of Critical Temperatures of Superconductors

4.2.1 Classification of Superconductors

Overall Performance Evaluation

In our classification analysis, we excluded chemical compositions that lacked T_c data. Such compositions were considered inappropriate for analysis under the assumption of $T_c = 0$, as this potentially imposes an inadequate presumption for these compounds. For chemical entries that were documented multiple times, we calculated the median T_c value. The median was selected over the mean due to its relative robustness against outliers. Our data set comprised approximately 12,000 unique chemical compositions.

We assigned binary classification labels based on a critical temperature threshold of 10 K. (i.e., $T_{\text{thres}} = 10$ K). This threshold was chosen as it effectively reduces class imbalance and optimises classification metrics (refer to Equations (5) - (7)). This approach aligns with the findings of Stanev et al. [194] and is supported by earlier studies [195, 196], thus ensuring an equitable distribution of classification labels. To further mitigate the effects of class imbalance at both the model development and evaluation phases, we applied the smoothed random oversampling (smoothed-ROS) technique with a shrinkage factor of 0.35. Smoothed-ROS is an extension of random oversampling with Gaussian noise. The addition of stochastic noise helps to prevent model overfitting on specific feature values by increasing the intraclass sample variability, thereby improving model generalisation. Our results indicated that smoothed-ROS alleviates potential learning biases most effectively when compared to other oversampling strategies.

For the partitioning of our data set, we implemented a stratified train-to-test split to preserve the class ratio consistency within the training and test sets as observed in the original data set. Each chemical composition was randomly selected from the initial data pool during this process. An 85%:15% ratio was maintained between the training and test sets. The resulting training set was subjected to feature processing through the GBFS workflow, which was instrumental in optimising our model performance.

The classification model we developed via the GBFS workflow demonstrated excellent efficacy, outperforming several benchmarks in the literature. Starting with an initial set of approximately 730 exploratory features and about 90 engineered features, the model identified a subset of the 29 most pertinent features. It achieved a weighted average F1-score of 0.912, an AUC-ROC of 0.986, an average precision (AP) of 0.919, and an accuracy of 0.947. The performance of the classification model is further

illustrated in Figure 4.8 (a) to (c), along with the confusion matrix in Figure 4.8 (d). A detailed breakdown of precision, recall, and F1-score metrics for each target label is presented in Table 4.6. Here, we use the AP score to summarise the precision-recall (PR) curve shown in Figure 4.8 (b) rather than the area under the curve. This decision was based on the potential overestimation of the area under the PR curve when calculated using the trapezoidal rule via linear interpolation. The AP is calculated as the weighted average of precision at each classification threshold, with the weights being the increments in recall from the previous threshold. These measures indicate that our classification model is highly discriminative, effectively distinguishing chemical compositions likely to exhibit T_c above 10 K.

These results were realised without adopting specific modifications or treatments to account for polymorphism — the presence of distinct 3-D crystal structures of the same chemical compound. Our approach strictly trains the model to remain indifferent to polymorphic variations. The ML model exhibits a high level of predictability, notwithstanding its ignorance of the various 3-D crystal structures that a compound may manifest, owing to the myriad ways in which chemical elements can arrange within unit cells of the crystalline lattice of each material.

We now explore specific evaluation metrics in Table 4.6. The table summarises the performance metrics for our classification model designed to predict whether observations fall above or below $T_{\text{thres}} = 10$ K. The metrics displayed include precision, recall, and F1-score for the categories 'Below T_{thres}' and 'Above T_{thres}', as well as aggregated statistics in the form of macro and weighted averages across these categories.

Precision measures the proportion of true positives relative to all positive predictions made by the model. A precision of 0.937 for instances classified as 'Above T_{thres}' suggests that 93.7% of the predictions made by the model for being above the threshold were indeed correct, with a minimal incidence of false positives. Conversely, recall assesses the percentage of actual positive instances correctly identified by the model. A recall of 0.960 for the 'Above T_{thres}' category indicates that the ML model successfully identified 96.0% of true positives (i.e., superconductors with $T_c > 10$ K). For predictions categorised as 'Below T_{thres}', the model demonstrates a precision of 0.959, affirming its high accuracy in predicting true negatives. The corresponding recall of 0.934 signifies that the model correctly identified 93.4% of all actual negatives. These metrics highlight the ability of the model to distinguish between the two target classes, ensuring reliable performance in applications where precise threshold-based classification is vital.

The F1-score, calculated as the harmonic mean of precision and recall (as delineated in Equation (5)), serves as a comprehensive metric that encapsulates the balance

between these two evaluation measures. An F1-score nearing 1 denotes an optimal equilibrium between precision and recall. Notably, our ML model afforded F1-scores that are commendably high in both categories, with 0.947 for 'Below T_{thres}' and 0.948 for 'Above T_{thres}', suggesting that the model effectively balances the trade-off between precision and recall. The results attest to the highly discriminative ability of our classifier, particularly highlighted by the substantial recall rate of 0.960 for the 'Above T_{thres}' category. This score is significant as it emphasises the model's proficiency in accurately identifying the majority of positive cases, which is often critical in applications where there are severe consequences for failing to detect positive instances.

Table 4.6 also presents the macro and weighted averages for precision, recall, and F1-score. The macro average calculates the mean performance across both categories, irrespective of class size. Conversely, the weighted average incorporates the support of each class, accounting for the number of instances in each category, which aligns performance evaluation with class prevalence. These averages reflect good model performance across both categories, ensuring a balanced model output regardless of class distribution. The close correspondence between the macro and weighted averages suggests that the class distribution does not adversely affect the model's performance metrics. This consistency is indicative of excellent model generalisation across different levels of class imbalance. Moreover, the high uniformity across metrics is likely due to the strategic adjustment of the T_{thres} value and the application of an oversampling technique in the training set. Specifically, the use of the smoothed-ROS method ensured equal representation of each class by generating synthetic feature vectors that mimic the characteristics of minority classes within the high-dimensional feature space. Lastly, a ten-fold cross-validation process afforded an F1-score of 0.95 ± 0.01. These performance metrics confirm the robustness and reliability of our classifier in threshold-based decision-making scenarios.

Table 4.6 Summary of the performance metrics for the classification of superconductors by their T_c values relative to threshold $T_{\text{thres}} = 10$ K.

	Precision	Recall	F1-score
Below T_{thres}	0.959	0.934	0.947
Above T_{thres}	0.937	0.960	0.948
Macro Average	0.948	0.947	0.910
Weighted Average	0.948	0.947	0.912

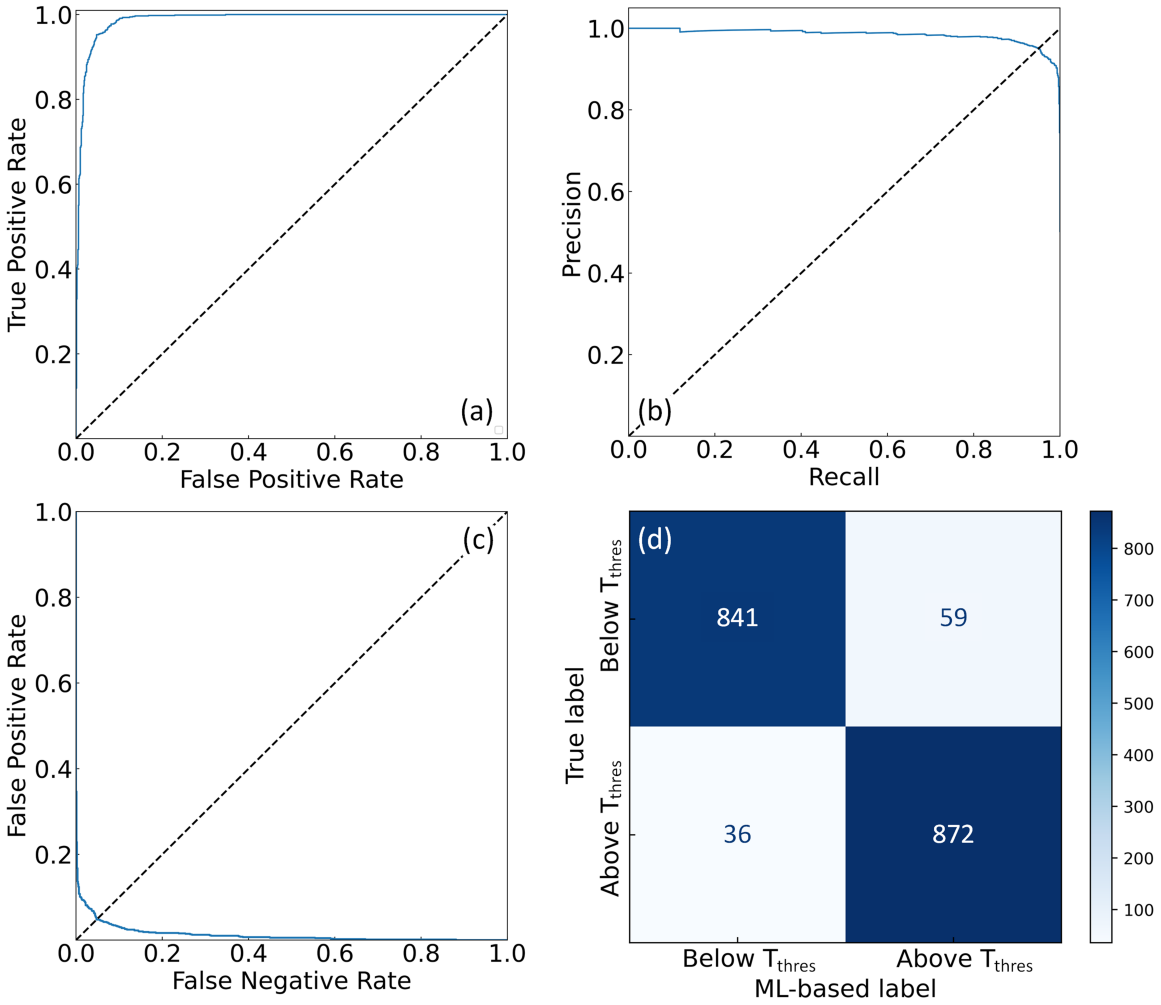


Fig. 4.8 Classification performance on the test set. (a) Receiver operating characteristic (ROC), (b) precision-recall (PR), (c) detection error trade-off (DET) curves and (d) confusion matrix. An AUC-ROC of 0.986, an average precision (AP) score of 0.919 and an equal error rate (ERR) of 0.05 are realised.

Feature Interpretation

In the classification analysis, our GBFS workflow identified the range of periodic table column (or group) numbers among the elements in a chemical composition as the most salient feature. Upon analysing the relationship between this feature and the median T_c values in the training set, we observed a correlation coefficient of around 0.52, where higher values of this feature generally correspond to higher T_c values. For instance, a range of periodic table column numbers less than 10 are typically associated with superconductors having a median T_c below 15 K. Conversely, we observe that the median T_c exceeds 40 K when the feature value surpasses 10. The peak of the feature distribution (i.e., the highest median T_c value) is observed when the range of periodic table column numbers is 14. This result is consistent with findings from other research, including a notable study by Stanev et al. [194], which identified the standard deviation of the periodic table column number as the most relevant predictor in their classification analysis. This finding is complemented by the importance of statistical measures related to the electronegativity, melting temperature, and atomic weight of the elements comprising the composition.

A possible interpretation appears to be associated with the influence of the observed variations in electron configuration and bonding characteristics on the physical properties of materials. For instance, elements with a single valence electron, such as alkali metals, exhibit distinct properties from transition metals, which often have partially filled *d*-orbitals. These *d*-orbitals contribute unique electronic characteristics that can be crucial for superconductivity. Furthermore, a wide range of the periodic table column numbers of the elements in a compound's formulation suggests high diversity in their electronic properties. This high diversity appears to be pivotal, as superconductivity is intricately linked to the electronic structure of the material. Elements from various groups influence the band structure differently, thereby affecting the material's conductive properties. Consequently, a broader range in column numbers not only signifies greater chemical complexity but may also indicate a more complex electronic band structure, which can be favourable for achieving higher T_c, as we will discuss later. This high level of complexity can enhance the superconducting properties of materials, making the range of periodic table column numbers a significant predictor in superconductivity studies. A clear example can be observed via cuprate materials, which have complex compositions and an extensive variety of constituent elements, making them likely to exhibit a broader range in this feature.

The second salient feature was identified as minimum molar volume. The molar volume of a material, which measures the volume occupied by one mole of a substance,

appears to be a key feature for predicting the T_c of superconductors due to several reasons related to the material's structural and electronic properties.

Molar volume impacts the electron-phonon coupling in a material, an important mechanism for superconductivity. In superconductors, the atomic spacing affects phonon modes — quantised modes of vibration within the crystal lattice that are essential for the electron pairing mechanism. This interaction is particularly crucial in conventional superconductors, where electron pairs (i.e., Cooper pairs) are mediated by phonon interactions. The Debye frequency of phonons underpins the well-established relationship $T_c \sim 1/\sqrt{m}$, where m is the ionic mass [197–199]. An increase in molar volume typically corresponds to larger interatomic distances, which can influence both the density and mobility of electrons. The spatial expansion affects the electron-phonon interaction, hindering the ability of electrons to pair and form Cooper pairs, an essential process for achieving superconductivity. The role of electron-phonon coupling and electron correlation strengths in superconductivity is supported by the experimental discovery of the isotope effect [197–200], providing strong empirical confirmation for these mechanisms, while the BCS (Bardeen-Cooper-Schrieffer) theory of superconductivity that explains the isotopic effect expresses the superconducting critical temperature as:

$$k_B T_c = 1.14\hbar\omega_c e^{-\frac{1}{N_0 V}}, \tag{4.10}$$

where k_B is the Boltzmann constant, $\hbar$ is the reduced Planck constant, ω_c is a cutoff frequency, N_0 is the electronic density of states near the Fermi level, and V is the electron-phonon coupling potential [201].

The molar volume also relates directly to the lattice constants of a material, which influences the material's electronic band structure. Changes in the band structure can affect the density of states at the Fermi level (Equation 4.10), a relevant factor for the superconducting properties of materials. Moreover, variations in molar volume affect the internal chemical pressures within a material. This can affect the hybridisation of electronic orbitals and thus influence superconductivity. In high-T_c superconductors, such as cuprates and iron-based superconductors, the internal pressure can modify the electronic structure, thereby affecting T_c. It is well-established that external pressure often elevates T_c in some superconducting materials; similarly, internal pressures induced by changes in molar volume can exert a comparable effect. By reducing the molar volume (i.e., increasing pressure) in some chemical compounds, superconductivity was observed in a number of high-pressure phases even when the chemical compound was not superconducting at ambient pressure [202]. This phenomenon is corroborated

by empirical data showing that lower values of minimum molar volume are somewhat associated with higher median T_c values.

The third prominent feature contributing significantly to the reduction in loss within our model is the average deviation of melting temperature among constituent elements in a chemical composition. The melting temperature appears to be an important feature for predicting the T_c of superconductors due to several interconnected physical and chemical properties.

The melting temperature is indicative of the strength of atomic bonds within a material, and variations in this temperature can influence electron density and mobility. The strength of these bonds affects the dynamics of electrons and phonons, which are fundamental to the mechanisms of superconductivity, particularly in conventional superconductors where phonon interactions mediate electron pairing. Consequently, disparities in melting temperatures can create varied local electronic environments, leading to a spectrum of phonon behaviours within the material. These variations, in turn, impact the electron-phonon interactions that are fundamental to the mechanism of superconductivity, as articulated by the BCS theory of superconductivity (Equation 4.10).

Furthermore, the analysis indicates that variations in melting temperature among the constituent elements correlate with the median T_c value up to a certain value. This indicates a possible enhancement of lattice vibrations that may facilitate the formation of Cooper pairs. Nevertheless, this trend reverses as the distribution of this feature exhibits a broad peak around 500 K. Such a pattern implies an optimal standard deviation in the melting temperatures among the constituent elements, which appears to maximise the positive influence on T_c predictions. Alternatively, a high standard deviation in melting temperatures may be interpreted as an indicator of considerable variability in the thermal stability of the constituent elements. Such variability could imply a lack of homogeneity in the material's composition, potentially leading to inhomogeneities in local electronic and phononic properties. These disparities across different regions of the material may influence its overall capability to transition into a superconducting state. The findings suggest that an optimal value exists for melting temperature variability, which could maximise the material's superconducting potential. This observation indicates a complex relationship between melting temperature and superconductivity.

It is necessary to recognise that due to the non-linear characteristics of our ML model, the likelihood of a chemical composition exhibiting a T_c value greater than 10 K is greatly influenced by the interaction among various features, some of which may

exhibit notable correlations. Consequently, these interactions can result in a distinct feature profile differing from those observed in other studies. The complex nature of this matter mandates careful consideration and analysis of feature interdependencies within the predictive modelling framework. Our examination of these features emphasises the intricate relationships that govern superconductivity and further accentuates the multifaceted nature of predictive modelling in superconductor research.

4.2.2 Regression Analysis of Critical Temperatures

Overall Performance Evaluation

Having established an effective classification model, we now advance to the more intricate task of developing a regression model to predict the critical temperatures (T_c) of superconductors based on their chemical composition alone. This endeavour is considered more complex because it involves quantifying the magnitude of T_c rather than merely categorising compounds based on their potential to exhibit discernible superconducting properties. Specifically, the development of a regression model not only deepens our understanding of the factors influencing T_c in known superconductors but also serves as a crucial element in initiatives aimed at discovering new superconducting materials.

We employed the same set of features used to create the classification model to formulate four distinct regression models. Initially, an ML model was trained on the entire data set using our GBFS workflow, which we will refer to as the 'all-T_c' model. In addition, we tailored specific regression models for distinct families of superconductors, namely cuprates and iron-based superconductors. A fourth ML model was then developed to handle data that did not fall into these categories, which we will refer to as 'low-T_c' group owing to their relatively smaller T_c values. We exclusively considered superconductors with valid T_c records, where we computed the median T_c values for duplicated chemical entries, and employed a random splitting methodology to divide the data. We maintained an 85%:15% train-to-test split ratio and strictly trained the model to be agnostic to crystalline polymorphs. This approach enables modelling without requiring 3-D crystallographic data, thereby permitting a methodical analysis across a wide set of superconducting materials.

We begin our analysis with a performance evaluation of the all-T_c regression model. The results of the random data partitioning are depicted in Figure 4.9, which displays the distribution of T_c values among known superconductors. Notably, the majority of the T_c records are concentrated below approximately 50 K, with a rapid decline in the

frequency of superconductors at higher T_c values. These distributions highlight the skewed nature of T_c within the data set, compiled from the scientific literature.

Figure 4.10 (a) presents the ML-based predictions of T_c plotted against experimental values that have been generated by the all-T_c model. This model is based on a Bayesian-optimised gradient-boosting algorithm that was refined using 34 features selected from an initial set of about 730 exploratory features via our GBFS workflow. The blue dot-dash line represents the line of best fit, established through the Ordinary Least Squares (OLS) method. This line illustrates the relationship between our ML-based predictions and the ground truth. The linear fit has a gradient of 0.94 and a y-intercept of 1.5 K. The y-intercept of the linear fit indicates a small systematic bias for lower T_c values, while the gradient reveals an almost ideal alignment between the predictions of our model and the experimental measurements, albeit with a minor underestimation at higher temperatures within the examined range.

The apparent deviations in model predictions at higher temperatures are understandable, considering the predominant concentration of data within the lower temperature range. The reduced data density at higher temperatures likely contributes to the discrepancies between the ML predictions and the ground truth within this range, as evidenced by the line of best fit (blue) in Figure 4.10 (a), which is positioned below the diagonal lines (red). Despite these challenges and the absence of any 3-D structural information, our ML approach has consistently provided accurate predictions of T_c, demonstrating its utility in modelling superconductors across varied conditions. With a closer examination of the error distribution plots presented in Figure 4.10 (b), it is evident that the majority of the prediction errors are below 10 K. Further analysis of the chemical compositions within the test set reveals that approximately 91% of cases exhibit an absolute error below 10 K, about 79% display errors below 5 K, and around 39% maintain errors below 1 K.

We observed that 13 chemical compositions out of approximately 1,800 in the test set exhibited deviations greater than 30 K between predicted and experimental values. These compounds predominantly belong to various families of cuprate superconductors which are known for their layered crystal structures that incorporate copper-oxide (Cu–O) planes; planes which are critical for their superconducting properties. These include YBCO (Yttrium Barium Copper Oxide), where Yttrium may be substituted by other elements; BSCCO (Bismuth Strontium Calcium Copper Oxide), characterised by their phase diversity and varying numbers of copper oxide planes; TBCCO (Thallium Barium Calcium Copper Oxide) and Tl-based compounds, where Thallium substitutes Bismuth; Mercury-based superconductors, noted for achieving high critical tempera-

tures; and Lanthanum-based superconductors, which involve complex substitutions affecting their superconducting characteristics. This finding highlights the intricate relationship between chemical composition and superconducting properties in these superconductors.

These results are further substantiated by an R^2 of 0.945, an MAE of 3.54 K and an RMSE of 6.57 K, reflecting a strong correlation between predictions and experimental measurements. The higher RMSE reflects greater penalisation for predictions that deviate from their true values. The distribution of absolute errors is shown in Figure 4.10 (b), further detailing the performance nuances of this predictive model. During a ten-fold cross-validation procedure, the model afforded an R^2 of 0.94 ± 0.01, an MAE of 3.73 ± 0.13 K, and an RMSE of 7.45 ± 0.65 K. These metrics highlight the consistent predictive accuracy and reliability of our model across multiple subsets of the data. We attribute the performance result to the implementation of our GBFS workflow, which effectively minimised feature redundancy and maximised relevance to the target variable. The efficacy of our modelling strategy enabled us to surpass the performance of other studies by employing a systematic approach to feature analysis and selection, and notably, without resorting to the use of regularisation techniques during model optimisation. The results on both the test set and the cross-validation process further affirm the generalisability of our model, without over-fitting to the training set.

To rigorously evaluate the results of our study against those obtained from other methods, we have compiled a table that encapsulates key performance metrics from various studies. Table 4.7 presents a comparative analysis of test-set results from diverse ML algorithms used to predict superconductivity in materials, examining evaluation metrics such as R^2, MAE, RMSE, Accuracy, F1-Score, and AUC-ROC. This comparison delineates the distinctions and parallels in model performance, providing insights into how our study aligns in terms of accuracy, generalisability, and predictive precision.

The findings confirm that our GBFS-based model, employing a train-to-test split of 85%:15%, achieves superior performance with an R^2 of 0.945, signifying its ability to explain 94.5% of the variance within the test set. Furthermore, this model exhibits robust classification capabilities, which indicate a well-balanced precision and recall and illustrate its proficient discrimination capacity. The random-forest model, with an identical train-to-test ratio, displays a substantial R^2 of 0.885 and consistent classification performance with both accuracy and F1-score at 92%, albeit slightly lower than that of the GBFS model.

In comparison to deep-learning methodologies such as the Deep-CNN model, which operates with a tighter train-to-test ratio of 95%:5%, it is observed that its regression capabilities are nearly on par with the GBFS-based model, achieving an R^2 of 0.92. Additionally, the DeepSet model, with a train-to-test ratio of 80%:20%, also demonstrates excellent regression performance with an R^2 of 0.93. These results underscore the capability of our GBFS-based model to achieve excellent outcomes, even when juxtaposed with more complex deep-learning approaches. It should be highlighted that the classification results associated with deep-learning methods included a test set from an additional source [203], thereby precluding a direct comparison. Moreover, the Deep-CNN model employed a unique testing scheme, specifically the temporal separation testing scheme.

Having established that a regression model trained on a comprehensive data set encompassing all types of superconductors performs robustly, we now aim to investigate the impact of training and testing the model on specific groups of superconductors. By adhering to the same modelling approach via our GBFS workflow, we demonstrate the adaptability of our methodology. Such targeted analysis not only tests the flexibility of the modelling technique but also provides insights into how prediction outcomes vary among different superconductor families, which is attributed to the distinct mechanisms that drive superconductivity in each group. This nuanced approach helps to understand the differential behaviour and underlying materials-property characteristics that are specific to each type of superconductor.

The high performance of our regression models is consistently observed across different groups of superconducting materials, specifically when models are trained separately on (a) cuprate, (b) iron-based, and (c) low-T_c superconductors. The results of these models are shown in Figure 4.11, and the corresponding distributions of absolute errors are depicted in Figure 4.12.

The performance of our regression models varies by the type of superconducting material used for training. The model based on cuprate superconductors achieved the lowest R^2 value of 0.84, accompanied by an MAE of 7.26 K and an RMSE of 11.87 K. In comparison, the iron-based model demonstrated slightly lower errors, with an MAE of 4.44 K, an RMSE of 6.27 K, and an R^2 of 0.88. Meanwhile, the model trained on low-T_c superconductors, which constitutes the larger data set with approximately 6,600 unique chemical compositions, showed the strongest performance with an R^2 of 0.90, an MAE of 1.15 K, and an RMSE of 1.92 K. This outcome is unsurprising given that the low-T_c model not only has the benefit of a larger data set but also targets superconductors within a relatively smaller T_c range. Conversely, the model

based on cuprate superconductors is trained using approximately 4,000 unique chemical compositions, with the test set encompassing a temperature range that extends up to 150 K. Meanwhile, the iron-based model is trained on a data set of around 1,500 unique compositions, covering a temperature range that falls between those of the low-T_c and cuprate models.

It is noteworthy that the regression models tailored to specific families of superconductors do not outperform the general model that incorporates all superconductivity data. This observation can be attributed primarily to a compositional effect, whereby each group of superconductors predominantly contributes to a unique range of T_c values. The composition effect denotes the portion of the observed inter-group variance in the distribution of the target variable (i.e., T_c) that can be attributed to disparities in the distribution of covariates among the groups. This phenomenon may arise from the data set containing multiple groups, each with distinct characteristics, thereby influencing the overall analytical results due to the unique attributes of each group. Consequently, the aggregate regression model, which encompasses a broader spectrum of data across extended temperature intervals, is more effectively calibrated and yields a more accurate determination of T_c values. Each family-specific model employs a distinct profile of features for making T_c predictions. This diversity in feature utilisation highlights the unique characteristics of each superconductor family and enables a detailed analysis of model efficacy within each specific material group.

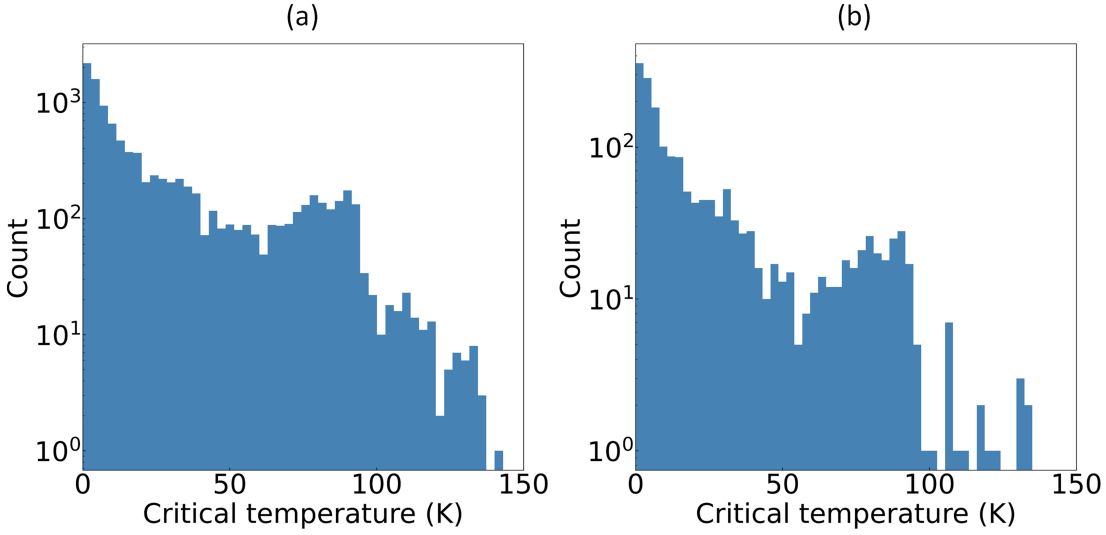


Fig. 4.9 Distribution of superconducting critical temperature (T_c) values in (a) the training set and (b) the test set.

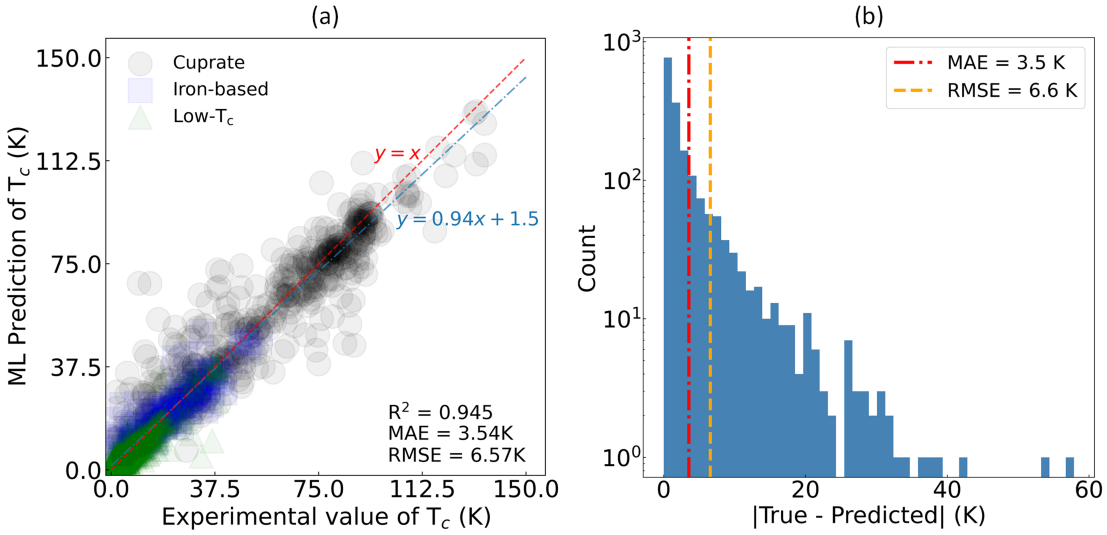


Fig. 4.10 (a) ML-based predictions of critical temperature (T_c) plotted against the experimental values (ground truth) for the regression model that has been trained and Bayesian-optimised on the final subset of features selected by our GBFS workflow. The dashed red line illustrates the hypothetical scenario in which the ML-based predictions perfectly match the experimental values. The blue dot-dash line represents a linear fit produced via the Ordinary Least Squares (OLS) method. The scatter plot is colour-coded to represent the different types of superconductors, where black circles correspond to cuprate, blue squares to iron-based, and green triangles to low-T_c superconductors. The accompanying panel in (b) displays the distribution of absolute errors in the predictions, where the dashed red line (−.) signifies the MAE and the dashed orange line (−−) represents the RMSE. These results pertain to the out-of-sample test set.

Table 4.7 Summary of test-set results for the regression and classification analyses using different ML methods to predict superconductivity of materials. * indicates that the corresponding classification analysis incorporated a test set from an additional source, specifically Ref 203.

Algorithm	Train-to-test split	R^2	MAE (K)	RMSE (K)	Accuracy	F1-Score	AUC-ROC
GBFS (this study)	85%:15%	0.945	3.54	6.57	0.947	0.912	0.94
Random Forest [194, 177]	85%:15%	0.885	-	-	0.92	0.92	-
Deep-CNN* [204]	95%:5%	0.92	-	-	0.95	0.77	0.94
Ensemble DeepSet* [205, 206]	80%:20%	0.93	-	9	0.84	0.71	-

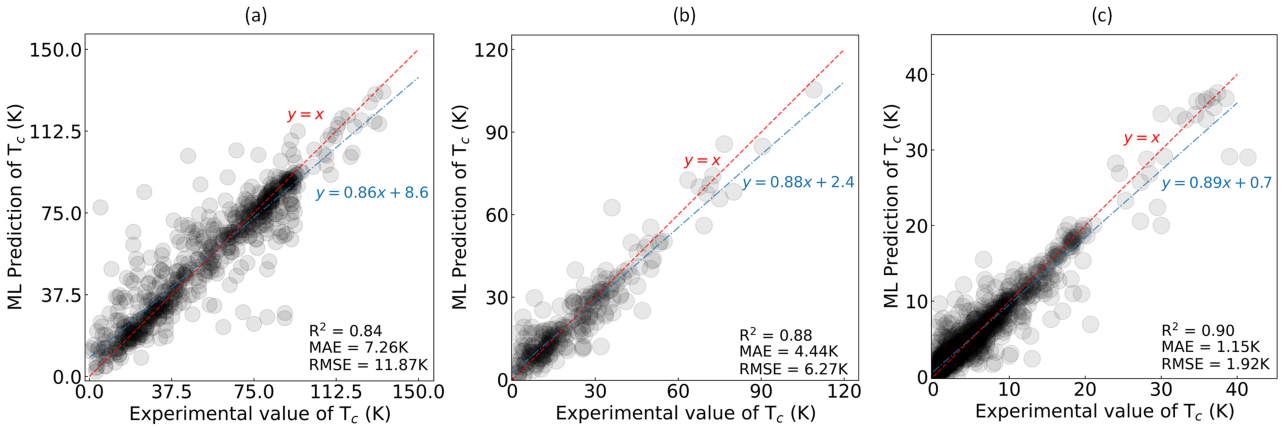


Fig. 4.11 ML-based predictions of critical temperature (T_c) plotted against experimental values (ground truth) in a series of regression models, each trained on distinct groups of superconducting materials: (a) cuprate, (b) iron-based, and (c) low-T_c superconductors. Each test set is confined to the respective superconducting group upon which it was trained. The dashed red line depicts the ideal scenario where ML-based predictions align perfectly with the experimental values. The blue dot-dash line indicates a linear fit generated using the Ordinary Least Squares (OLS) method. These results pertain to the out-of-sample test set.

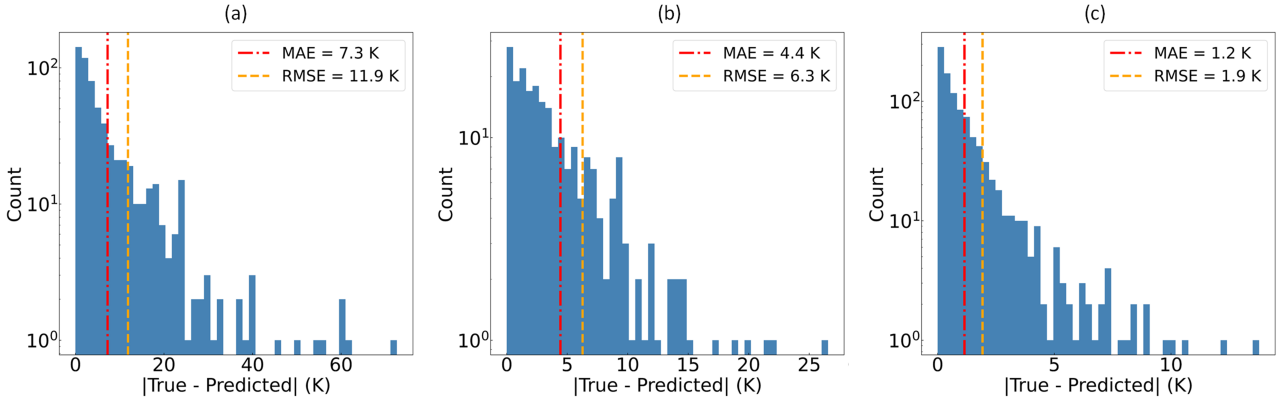


Fig. 4.12 Distribution of absolute errors in the ML predictions critical temperature (T_c), where each model is trained on distinct groups of superconducting materials: (a) cuprate, (b) iron-based, and (c) low-T_c superconductors. The dashed red line ($-.$) signifies the MAE and the dashed orange line ($--$) represents the RMSE. These results pertain to the out-of-sample test set.

Feature Interpretation

Table 4.8 summarises the most salient features that contributed to the results of our regression analyses, including those pertinent to family-specific models. We now seek to rationalise their significant role. The foremost influential feature, as denoted by the realised total loss reduction, is the range of thermal conductivity among the constituent elements within the chemical composition. The second most influential feature concerns the standard deviation of molar volume among these elements. Further features demonstrating notable relevance to the target variable include a variety of statistical measures related to the number of unfilled orbitals, the presence of calcium and iron, the periodic column (or group) number, the proportion of *d*-valence electrons, and the total number of electrons. We examined the significance of molar volume and periodic table column numbers. Consequently, this section will focus on exploring other features that have played pivotal roles in our predictive analysis.

The feature exhibiting the highest relevance in our analysis is the range of thermal conductivity among the constituent elements within the chemical composition. A correlation coefficient of 0.68 between the feature and the target variable (i.e., the median T_c values) indicates a strong positive trend. This observation is corroborated by SHAP analysis results, which demonstrate that higher values of this feature predominantly correspond to positive SHAP values. There appears to be a clear interplay of thermal and electronic properties of superconducting materials that govern this association.

This stands to reason because superconductors are often used in applications that are exposed to high currents or external magnetic fields, whereby efficient heat dissipation is an essential property of the superconducting material due to the substantial heat loads these conditions can generate. High thermal conductivity also helps to maintain the superconductor below its T_c by effectively dissipating heat generated by electrical resistance or external sources. Materials with higher thermal conductivity are generally more adept at stabilizing their temperature, thereby helping to maintain superconductivity under varying operational conditions. Thermal conductivity in materials is also intricately linked to phonon (lattice vibration) activity. As previously mentioned, phonons mediate electron pairing which is necessary for the superconducting state through electron-phonon interactions. High thermal conductivity generally indicates efficient phonon transport, which can improve electron-phonon coupling and thereby potentially elevate T_c. This relationship is especially significant in conventional superconductors, where, according to the BCS theory, superconductivity is phonon-mediated. The ability of phonons to efficiently transport heat, a direct consequence

of high thermal conductivity, can play a crucial role in influencing the dynamics of electron pairing, which is essential for the onset of superconductivity.

Moreover, a considerable variation in thermal conductivity among the constituent elements of a material may indicate distinct phonon behaviours within the substrate. Such variation can create unique or uneven electron-phonon interactions, which could influence the stability and uniformity of the superconducting state. This effect is particularly pronounced in materials where superconductivity is highly sensitive to local phononic environments. Variability in this trait could lead to localised regions where superconductivity might be either suppressed or enhanced, thereby affecting the overall T_c value of a material. Such heterogeneity is apparent in the empirical data. Empirical analysis of the distribution of this feature reveals multiple distinct peaks in the profile, suggesting the existence of various crystal-lattice environments. These environments, facilitated by conditions related to thermal conductivity, may be conducive to efficient phonon-mediated electron pairing.

Additional factors warranting consideration in the context of thermal conductivity include low-temperature thermal conductivity, effects of anisotropy, the superconducting order parameter, and the two-fluid model. For instance, at temperatures approaching absolute zero, the thermal conductivity of superconductors tends to be dominated by electron transport. The characteristics of thermal conductivity near T_c can provide valuable insights into the strength of electron-phonon coupling, which directly impacts T_c. Regarding factors pertinent to experimental observations, anisotropic effects exist in high-T_c superconductors, such as cuprates. The anisotropy in thermal conductivity, characterised by varying values in different crystallographic directions, can provide information about the anisotropy of the superconducting gap. Additionally, the superconducting order parameter is a central feature of the BCS theory, describing how the electron pairs condense into a coherent quantum state with lower energy than the normal (non-superconducting) state. Fluctuations in the superconducting order parameter can influence thermal conductivity near the transition temperature — a phenomenon observed in various high-T_c materials. More specifically, the magnitude of the order parameter corresponds to the energy gap in the electronic structure of a superconductor that forms at the Fermi surface. This gap represents the energy required to break Cooper pairs into individual electrons, thus disrupting the superconducting state. Below T_c, thermal conductivity can also be interpreted using the two-fluid model, which conceptualises the material as comprising both normal (non-superconducting) and superconducting electrons. The proportion of electrons in the superconducting state significantly affects the overall thermal conductivity, as superconducting electrons

do not contribute to thermal transport in the same manner as normal electrons. While the relationship between thermal conductivity and T_c is inherently linked through the material's thermal and electronic properties, it is modulated by various factors including the material's structural characteristics, purity, and the nature of superconducting pairing. These factors must be considered to fully understand how T_c can be optimised or predicted based on thermal conductivity measurements.

Another salient feature is associated with the number of unfilled orbitals, the fraction of *d*-valence electrons and the total number of valence electrons, for which a clear rationale exists. The number of unfilled orbitals pertains to electron configurations of constituent elements within a chemical composition. Analysis of the training set reveals a considerable correlation of -0.49 between T_c and the mean number of unfilled orbitals, with the distribution of the feature peaking at around a value of one. This suggests that an optimal value of one unfilled orbital, on average, is conducive for superconductivity. Transition metals are characterised by multiple unfilled d and f orbitals, contributing to intricate electronic structures that are advantageous for high-T_c superconductivity. This characteristic is prominently observed in cuprates and iron-based superconductors. For instance, copper is a key component in cuprate superconductors and it has one unfilled orbital, as indicated by its electronic configuration $[Ar]4s^1 3d^{10}$ (or $[Ar]4s^2 3d^9$ before forming an oxide).

It is, therefore, unsurprising that such a feature would be highly relevant for the cuprate-based model, given that structures incorporating multiple superconducting copper oxide (Cu–O) planes per unit cell tend to exhibit higher T_c values. Specifically, a higher proportion of Cu and O per formula unit correlates with a reduction in the average number of unfilled orbitals — typically two for O and one for Cu. Furthermore, cuprates typically comprise layers of Cu-O structures, which are stabilised through the addition or substitution of cations like Ba^{2+} and La^{3+}. These cations possess a large number of unfilled orbitals, thereby elevating the average number of unfilled orbitals within the compound. This suggests that the ability of interlayer cations to supplement charge to the Cu–O planes could be a crucial factor in influencing the superconducting properties of these materials.

The relationship between *d*-valence electrons and the T_c value for superconductors is complex and is rooted in the electronic structure and behaviour of the materials. The corresponding *d*-orbitals tend to be partially filled in transition metals and their compounds. These *d*-orbitals contribute to the density of states at the Fermi level, a key determinant in superconductivity. A heightened density of states at the Fermi level can lead to an increase in the T_c value. Additionally, *d*-electrons are known for their strong

electron-electron interactions, which arise from their close proximity in energy levels and spatial distribution. These interactions are pivotal in facilitating unconventional superconductivity, as commonly seen in high-T_c superconductors including cuprates and iron-based superconductors. In the realm of conventional superconductors, d-electrons can enhance electron-phonon coupling. Such enhanced coupling mediated by d-electrons can elevate T_c via the mechanism described by the BCS theory of superconductivity.

Another consideration may be linked to magnetic fluctuations and crystal-field effects. For example, d-orbitals are known to contribute to magnetic fluctuations. In high-T_c superconductors, such as cuprates and iron-based superconductors, magnetic interactions mediated by d-electrons are integral to the electron-pairing mechanism. Moreover, the stabilisation of the crystal field in materials containing d-valence electrons influences the electronic environment by causing a split in the d-orbital energy levels. This splitting alters the electronic band structure, which can affect the superconducting properties of the material.

Empirical studies have frequently demonstrated that materials featuring d-valence electrons are capable of achieving higher T_c values under specific conditions. Notably, materials such as cuprates (which incorporate copper d-orbitals) and iron-based superconductors (which involve iron d-orbitals) are known for their high T_c values and intricate phase diagrams. These characteristics are influenced by the properties of their d-valence electrons. Furthermore, experimental modifications to the electronic configurations of d-valence electrons, achieved through methods such as doping or the application of external pressure, have been effectively employed to optimise and elevate T_c. Our analysis of the training set suggests that the optimal fraction of d-valence electrons for superconductivity lies approximately between 0.2 and 0.5. It is also noteworthy to mention that theoretical models that incorporate the behaviour of valence electrons, such as resonating valence bond (RVB) theory which builds upon Hubbard and $t - J$ models, are instrumental in enhancing our understanding and prediction of the superconducting properties of these complex, strongly correlated materials [207–213]. These theoretical models provide insights into the electronic interactions that may govern superconductivity in materials characterised by d-valence electrons.

The significance of the number of valence electrons in influencing the T_c value of superconductors is underscored by an empirical relationship first identified by Matthias in 1955 [214]. Matthias observed a relationship between T_c and the number of valence electrons per atom, noting that optimal conditions for superconductivity tend to occur at 5 and 7 valence electrons per atom. The study led to a conclusion that

a larger number of valence electrons ($\gtrsim 4$) can be favourable for the occurrence of superconductivity, although an exception exists for electronic configurations near six valence electrons. This anomaly is characterised by a noticeable decrease in the T_c value, which is thought to arise from reductions in paramagnetic susceptibility and electronic specific heat. These phenomena are attributed to a reduced effective number of free electrons.

The prominence of chemical elements such as copper and iron in the context of superconductivity is self-explanatory, especially in relation to cuprates and iron-based superconductors which have been highlighted throughout this discussion. Nevertheless, it is noteworthy that features based on calcium have also emerged as prominent in relation to T_c. This observation warrants a more detailed exploration to elucidate the specific role and mechanisms by which calcium influences superconducting properties.

Calcium is commonly found in various superconductors. Examples include cuprate superconductors based on the Bi-Sr-Ca-Cu-O system [215], cuprate superconductors based on the Hg–Ba–Ca–Cu–O system [216–218], calcium intercalated graphite compounds [219], calcium superhydrides [220], among many others [221–223]. In the Supercon database, approximately 14% of the chemical compositions contain calcium, with T_c values ranging up to about 140 K. Besides, calcium is frequently used as a dopant in superconducting materials. During the doping process, calcium can substitute for other elements within the crystal lattice, thereby modifying the electronic structure and potentially elevating T_c. When introduced into superconductors, particularly in cuprates or iron-based superconductors, it can alter the carrier density within the material. An increase in carrier density can then elevate T_c by enhancing the superconducting pairing mechanism. Additionally, calcium plays a key role in stabilizing the crystal structure of iron-based superconductors, which is vital for maintaining the pathways necessary for effective electron flow and pairing. The incorporation of calcium can adjust the balance between electron-like and hole-like carriers, an influential factor in the mechanism of superconductivity in these materials.

The final note in this section underscores the distinct feature profiles for each sub-family of superconducting groups, as outlined in Table 4.8. This differentiation in feature profiles is anticipated given the different underlying material-property mechanisms across these families of superconductors. It is noteworthy to mention that the most relevant features for the cuprate and iron-based models align with those identified by Stanev et al. [194], whose modelling approach differs from that employed in this study. These results indicate that our modelling approach effectively captures the distinct characteristics of different families of superconductors, each

driven by unique mechanisms of superconductivity. However, the feature profile for the low-T_c model is different, which is not surprising given that this group was formed by aggregating the residual materials after segregating cuprates and iron-based superconductors. Consequently, the low-T_c data set lacks a specific mechanism governing superconductivity, as it comprises a heterogeneous mix of various materials. We expect to observe improvements in predictive accuracy for these analyses when additional features that extend beyond mere chemical composition are incorporated. Specifically, the incorporation of features derived from crystallographic data, density of states, doping levels, as well as experimental conditions such as external pressure, is expected to refine the prediction accuracy of T_c values against their experimental measurements. These expanded feature sets are likely to capture more complex interactions and dependencies within these materials, offering a deeper understanding of superconducting behaviours.

Table 4.8 List of most relevant features identified by our GBFS workflow for the regression of T_c, including sub-family analyses.

	Model type			
Feature rank	all-T_c	cuprates	iron-based	low-T_c
1	Thermal conductivity	No. of unfilled orbitals	Periodic column (group)	Magnesium
2	Molar volume	Praseodymium	Packing efficiency	Space group no.
3	No. of unfilled orbitals	*d*-valence electrons	Oxygen	No. of unfilled *p*-orbitals
4	Calcium & Iron	Calcium	Thermal conductivity	Niobium
5	Periodic column (group)	Periodic block	Molar volume	First ionisation energy
6	*d*-valence electrons	Volume per atom	Mendeleev no.	Volume per atom

4.2.3 Predictive Capabilities for ML-Model Applications

Case study 1: Analysis of Out-of-Sample T_c Predictions

The findings reported thus far provide predictions of T_c values juxtaposed against experimental measurements, yielding promising statistical figures-of-merit. Nonetheless, it is important to validate these results by considering the predictions within a distinct range of temperatures of specific interest rather than merely demonstrating their collective statistical quality in an anonymised form. Therefore, we conducted a comparative analysis using experimental measurements of high-T_c superconductors that had not been previously encountered by our ML model during the training phase. Such an analysis allows us to take a closer examination to assess its predictive capability.

The predicted T_c values were compared against experimental measurements, with the analysis of 50 examples provided in Table 4.9. These examples were randomly

selected from instances where the experimental measurement of a T_c value exceeded 85 K. This selection criterion is important, as materials displaying superconductivity at elevated temperature (cf. relative to the liquid-nitrogen temperature) are highly desirable for various practical applications.

The propensity of our model to underestimate T_c values against the experimental measurements, particularly at higher temperature ranges, is clearly observed. This trend is evident in 38 out of 50 cases, where the predictions show a negative percentage difference from the experimental values. This underestimation becomes more apparent at higher T_c values, as illustrated by the line of best fit (blue) in Figure 4.10 (a); this line sits below the diagonal with a gradient of 0.94. We attribute this anomalous pattern to the scarcity of data about high-T_c superconductors in our training set, as well as in the broader scientific literature. Such scarcity can limit the ability of ML models to effectively learn and predict within the higher temperature range. We anticipate that by augmenting the data set with a greater number of data points that represent these higher temperature ranges, one would significantly improve the predictive accuracy of our model.

Despite the apparent negative bias, which is modest in proportion, our ML model demonstrates strong performance in predicting T_c values on an absolute scale. The mean absolute percentage difference from the ground truth is approximately 4.5%. Specifically, around 88% of these predictions deviate from the ground truth by less than 8% on the absolute scale, and approximately 54% exhibit an absolute deviation of less than 5%. In Table 4.9, the absolute percentage difference of each predicted T_c value from its experimental measurement is colour-coded according to the classifications green ($0-4\%$), amber ($4-8\%$) or red ($>8\%$). Only 6 out of the 50 examples were classified in the red category; thereby, demonstrating the efficacy of our ML model in accurately predicting T_c values. Some of the lowest absolute percentage differences were observed for chemical compounds such as $Er_1Ba_2Cu_3O_7$, $Bi_{0.6}Nd_{0.2}Tb_{0.2}Pb_1Sr_2Ca_2Cu_3O_{10}$, $Y_1Ba_{1.9}La_{0.1}Cu_3O_{6.975}$, and $Hg_{0.8}W_{0.2}Ba_2Ca_1Cu_2O_{6.4}$. Meanwhile, some of the highest absolute percentage differences were noted for chemical compounds such as $Tl_2Ba_2Ca_1Cu_2O_{7.984}$, $Nd_1Ba_{1.8}Sr_{0.2}Cu_3O_{6.96}$, and $Y_{0.6}Eu_{0.4}Ba_2Cu_3O_7$, whose absolute percentage differences from the ground truth exceed 10%.

Case study 2: Electronic Density of States of $Er_1Ba_2Cu_3O_7$

We now take a closer look at one of the chemical compositions in Table 4.9, namely $Er_1Ba_2Cu_3O_7$. This compound is a type of high-T_c cuprate superconductor and is part of the generic $YBa_2Cu_3O_7$ (YBCO) structural family, where yttrium (Y) is substi-

tuted by another rare-earth element erbium (Er) in this instance. The orthorhombic crystal structure of this material mirrors that of YBCO and contains Cu–O planes interspersed with layers containing Ba and Er, where the Cu–O planes are essential for superconductivity. The electronic density of states (DOS) for $Er_1Ba_2Cu_3O_7$ is depicted in Figure 4.13. This was derived from density functional theory calculations sourced from the Materials Project [5].

In the context of superconductivity, a flat band refers to an energy band in which the energy levels remain relatively constant across a range of momentum values, indicating a high degree of electronic-state localisation. Such flat bands markedly influence the electronic DOS at particular energy levels. The DOS is inversely proportional to the gradient of the energy-band dispersion. Therefore, in flat bands where the dispersion curve is horizontal (indicative of zero gradient), the DOS becomes singularly large. This explains why the presence of a high DOS at the Fermi level is important for superconductivity, particularly according to the BCS theory, which posits that a higher DOS at the Fermi level increases the likelihood of Cooper pair formation under favourable conditions. In cuprate superconductors such as $Er_1Ba_2Cu_3O_7$, the electron pairing mechanism may involve more complex interactions than those characterised by the conventional BCS theory. In such systems, the presence of flat bands just below the Fermi level can enhance these interactions, thereby enhancing electron pairing probabilities. Notably, the DOS for $Er_1Ba_2Cu_3O_7$ exhibits large peaks immediately below the Fermi energy level, which corroborates the theoretical perspectives discussed in the preceding section on feature interpretation.

Furthermore, the role of flat bands in superconductivity is frequently discussed alongside the concept of a Van Hove singularity in the DOS [224, 225]. A Van Hove singularity arises when the gradient of the band structure approaches zero, resulting in a pronounced peak in the DOS. The close proximity of such a singularity to the Fermi level can elevate T_c by intensifying the electronic interactions that facilitate superconductivity. Studies on various high-T_c cuprates have demonstrated that alterations in the band structure, induced by chemical doping or the application of pressure, can shift the relative positions of flat bands and Van Hove singularities with respect to the Fermi level, thereby affecting T_c. This relationship highlights the importance of DOS profiles in understanding the electronic properties and behaviours essential to superconducting states.

Case study 3: Out-of-Distribution T_c Predictions for $(Ca_{1-x}La_x)FeAs_2$ and $(Ca_{1-x}La_x)Fe(As_{1-y}Sb_y)_2$

We expand upon the predictive capabilities of our ML model beyond benchmarking against state-of-the-art reports. Specifically, we undertook a case study, whereby we focused on predicting how the T_c value of $(Ca_{1-x}La_x)FeAs_2$ varies with the lanthanum (La) concentration, denoted by x. For this analysis, the T_c values reported by Kudo et al. [226] serve as the ground truth, while an additional 73 T_c records were sourced from Hosono et al. [203], expanding the training set. To ensure the integrity of these out-of-distribution predictions, we ensured that there was no overlap between the training and test sets. Here, out-of-distribution predictions refer to model predictions made on chemical compositions that are either underrepresented or entirely absent from the training data. This approach challenges our ML model to navigate unexplored material spaces by extrapolating from established chemical-property relationships learned during training. This endeavour not only tests the robustness of our model but also enhances our understanding of its predictive power and potential utility in practical applications.

Figure 4.14 presents the ML-based predictions of the superconducting transition temperature T_c for $(Ca_{1-x}La_x)FeAs_2$ as a function of the lanthanum (La) content (x), juxtaposed with literature values. The predictions illustrate T_c as calculated based on the proportional composition of La. The figure includes an orange bar chart that depicts the residuals, which are defined as the absolute differences between the ML-predicted values (represented by black diamonds) and the ground-truth measurements (depicted by grey circles). This visualisation highlights the levels of accuracy and discrepancies of the predictions relative to established empirical data. The results indicate that the majority of the T_c predictions made by our ML model for the previously unseen chemical compositions in this family of iron-based superconductors correspond closely with values reported in the literature, effectively capturing the overall downward trend of T_c as the La content increases to 0.22. The 112 phase of the chemical family, $(Ca_{1-x}La_x)FeAs_2$, is achievable at $x \geq 0.15$. Notably, the T_c value is at its highest with 35 K when the composition, x, is at its lower boundary of $x = 0.15$; moreover, T_c decreases monotonically with increasing La content.

While not depicted in Figure 4.14, additional computations were performed for another case study that involved an antimony (Sb)-doped variant of this chemical family, whose generic chemical formula is $(Ca_{1-x}La_x)Fe(As_{1-y}Sb_y)_2$, where we specifically focused on a compound with $y = 0.01$ and $x = 0.15$. Our ML model predicted a T_c value of 40 K for this compound, which closely aligns with the 43 K reported by

Kudo et al. [226]. This observation reinforces the T_c trend and it is attributable to the introduction of Sb doping. In particular, simultaneous Sb doping facilitates a reduction in La content down to $x = 0.12$, yielding an increase in T_c to 47 K. Generally, such chemical substitutions not only modify the number of charge carriers but also induce a form of chemical pressure, influencing the electronic properties and superconducting behaviour of the material.

The enhancement of T_c in the $(Ca_{1-x}La_x)Fe(As_{1-y}Sb_y)_2$ family of chemicals can be attributed to two key factors. Firstly, the reduction in La content influences T_c due to its role in modifying charge carrier density. Kudo et al. found that, despite the doping of La, the unit cell volume and lattice parameters of $(Ca_{1-x}La_x)FeAs_2$ remain largely unchanged, which is attributed to the similar ionic radii of La^{3+} and Ca^{2+} [226]. This observation indicates that a reduction in La content directly depletes the number of charge carriers. Secondly, the incorporation of Sb introduces a distinct effect. Simultaneous Sb doping exerts a negative chemical pressure, leading to an increase in unit cell volume [203, 226]. This is due to the fact that the ionic radius of As^{3-} is smaller than that of Sb^{3-}. The resultant increase in unit cell volume primarily affects the in-plane crystallographic lattice parameters a and b, while the c parameter remains relatively stable. This analysis underscores the complex interplay of features that influence the prediction of T_c in superconductors, demonstrating that T_c is highly contingent upon a multitude of factors.

3.3.4 Case study 4: Virtual Compositional Search of Hg-Pb-Ba-Ca-Cu-O Series of Compounds

In our fourth case study, we explored the potential utility of ML models to assist in the compositional search for a particular set of chemical compounds. Specifically, we examined whether ML models could be employed to modify the doping levels of certain chemical elements to predict or identify trends in material properties, such as critical temperature of superconductivity, T_c. Herein, we focus on the Hg-Pb-Ba-Ca-Cu-O series of compounds, starting from the compound with chemical composition $Hg_{0.66}Pb_{0.34}Ba_2Ca_{1.98}Cu_{2.9}O_{8.4}$, which holds the record for the highest T_c value within the Supercon database. In this chemical compound, the elements mercury (Hg) and lead (Pb) are typically considered to be dopants. While Hg is a part of the main structure in some high-temperature superconductors, its fractional composition (i.e., 0.66) in this specific chemical stoichiometry suggests a significant role in optimising its crystal lattice structure or enhancing its superconducting properties, possibly through partial substitution or supplementation of other elements. Moreover, Pb, indicated

by the stoichiometric fraction 0.34, is likely serving with a doping function aimed at modifying material properties to improve material stability or modify electronic structures, with an aim to enhance T_c.

To investigate these hypotheses, we performed a virtual compositional search by independently adjusting the concentrations of Hg and Pb in a systematic manner, with an increment of 0.01 in the elemental content (x), to identify possible trends in T_c profiles. It is essential to underscore that this analysis presupposes the chemical validity and stability (i.e., synthesizability) of all variants, without consideration of any specific structural forms or alterations. Our method focused exclusively on modifying the chemical composition to establish a predictive profile of T_c for these virtual compounds. Consequently, the findings should be interpreted cautiously, serving more as an illustration of the potential utility of ML models rather than as definitive results. The outcomes of this investigation are depicted in Figure 4.15.

The marked sensitivity of T_c to variations in the Hg and Pb concentrations indicates that these elements are crucial to the superconducting properties of the material, supporting the fact that achieving optimal doping levels is essential for maximizing T_c. As anticipated, in both cases, the ML model identified a single pronounced peak near the ground truth (starting compound), which was seen by the ML model during the training phase, with no other significant peaks observed. This potentially suggests the absence of other chemical variants with T_c values comparable to that of $Hg_{0.66}Pb_{0.34}Ba_2Ca_{1.98}Cu_{2.9}O_{8.4}$, at least within the compositional space in which we have searched herein. In Figure 4.15 (a), variations in the Pb content (x) resulted in a profile showing relatively greater fluctuations in the T_c profile, potentially indicating complex interactions or phase transitions within the virtual superconductor that impact its superconductivity. Moreover, a notable decrease followed by smaller peaks in T_c as Pb concentration increases beyond the primary peak may suggest possible instability or variations in superconducting properties driven by chemical composition modifications. Conversely, Figure 4.15 (b) presents a smoother decline in T_c after the peak, implying that variations in the Hg content might lead to a more gradual adjustment in superconducting properties compared to changes in the fractional content of Pb. This differential response underscores the intricate relationship between chemical composition and the superconducting characteristics of the material.

Despite the fact that this chemical compound search is based on a systematic virtual compositional survey, which relies on several scientifically unwarranted assumptions as previously noted, these results nevertheless illustrate the potential utility of ML models to perform compositional searches around an experimentally verified chemical material

to predict superconductive properties from chemical compositions. This could be pivotal for designing new superconductors without extensive empirical testing. Insights from analyses such as these could guide the synthesis of new superconductors with tailored properties by adjusting the concentrations of dopants within a compound. Moreover, these figures illustrate the fundamental role of elemental composition in determining the superconducting properties of cuprate superconductors. They also demonstrate the utility of ML in forecasting material properties based exclusively on chemical compositions, potentially expediting the discovery and refinement of new superconducting materials.

Table 4.9: Critical Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of Critical Temperature (i.e., Model Output), Along With the Percentage Difference From the Experimental T_c Values (i.e., Ground Truth)[a].

Chemical composition	Experimental T_c (K)	ML prediction of T_c (K)	% diff.
$Hg_1Ba_2Ca_2Cu_3O_{8.29}$	135	126.0	-6.6
$Hg_{0.8}Pb_{0.2}Ba_2Ca_2Cu_3O_8$	132	130.4	-1.2
$Hg_{0.85}Re_{0.15}Ba_{1.66}Sr_{0.34}Ca_2Cu_3O_8$	131.3	129.7	-1.2
$Hg_{0.8}Ge_{0.2}Ba_2Ca_1Cu_2O_{6.2}$	124	114.5	-7.6
$Hg_{0.8}W_{0.2}Ba_2Ca_1Cu_2O_{6.4}$	118	117.2	-0.7
$Bi_{1.65}Pb_{0.35}Sr_2Ca_2Cu_3O_{10}$	108.6	99.7	-8.2
$Bi_{0.6}Nd_{0.2}Tb_{0.2}Pb_1Sr_2Ca_2Cu_3O_{10}$	107.5	107.0	-0.5
$Bi_{1.7}Pb_{0.3}Sr_2Ca_2Cu_3O_{10}$	107.25	99.8	-7.0
$Ga_1Sr_2Ca_3Cu_4O_{11.1}$	107	100.4	-6.2
$Tl_2Ba_2Ca_1Cu_{1.98}Fe_{0.02}O_8$	106	97.7	-7.9
$Hg_1Ba_2Ca_1Cu_2O_{6.13}$	102	100.4	-1.6
$Hg_1Ba_2Ca_2Cu_3O_{8.17}$	96	100.1	4.2
$Tl_1Ba_2Ca_{0.6}Y_{0.4}Cu_2O_7$	95.8	86.3	-10.0
$Hg_1Ba_2Cu_1O_{4.18}$	95.16	88.0	-7.5
$Nd_1Ba_2Cu_3O_{6.82}$	95	90.5	-4.8
$Gd_1Ba_2Cu_3O_{6.98}$	94.2	92.8	-1.4
$Y_1Ba_{1.92}La_{0.08}Cu_3O_{6.96}$	94	93.4	-0.7
$Er_{0.97}Pr_{0.03}Ba_2Cu_3O_7$	93.9	91.3	-2.8
$Y_1Ba_{1.9}La_{0.1}Cu_3O_{6.975}$	93.9	93.3	-0.6
$Y_1Ba_{1.7}Cs_{0.3}Cu_3O_{6.6}$	93.75	86.3	-8.0
$Y_1Ba_{1.97}La_{0.03}Cu_3O_{6.96}$	93.2	94.0	0.9
$Y_{0.6}Eu_{0.4}Ba_2Cu_3O_7$	92.6	83.1	-10.3
$Y_{0.925}Ba_{1.975}Ca_{0.1}Cu_4O_8$	92.3	86.6	-6.2
$Dy_1Ba_2Cu_3O_{6.9}$	92.3	89.8	-2.7
$Er_1Ba_2Cu_3O_7$	92.0	92.6	0.6

Continued on next page

Table 4.9: Critical Temperature - Examples of Chemical Composition (i.e., Model Input) and the Corresponding Prediction of Critical Temperature (i.e., Model Output), Along With the Percentage Difference From the Experimental T_c Values (i.e., Ground Truth)[a]. (Continued)

$Y_1Ba_2Cu_{2.97}Co_{0.03}O_{6.96}$	92	89.9	-2.3
$Er_1Ba_2Cu_3O_{6.91}$	92	90.4	-1.7
$Tl_1Ba_2Ca_{0.9}Gd_{0.1}Cu_2O_7$	91.8	90.3	-1.6
$Y_{0.6}Cd_{0.4}Ba_2Cu_3O_7$	91.7	90.9	-0.8
$Y_1Ba_2Cu_{2.96}Co_{0.04}O_7$	91.5	85.9	-6.1
$Y_{0.95}Pr_{0.05}Ba_2Cu_3O_{6.95}$	91.5	89.3	-2.4
$Dy_{0.95}Pr_{0.05}Ba_2Cu_3O_{6.9}$	91.4	85.9	-6.0
$Nd_1Ba_{1.8}Sr_{0.2}Cu_3O_{6.96}$	91.1	81.8	-10.2
$Tl_2Ba_2Ca_1Cu_2O_{7.984}$	91	100.1	10.0
$Y_{0.68}Tb_{0.32}Ba_2Cu_3O_7$	91	83.8	-7.9
$Y_1Ba_2Cu_3F_{0.25}O_{6.7}$	91	88.0	-3.3
$Er_1Ba_2Cu_3O_{6.93}$	91	90.2	-0.9
$Y_1Ba_2Cu_{2.98}Fe_{0.02}O_7$	90.7	88.9	-2.0
$Y_1Ba_{1.8}Rb_{0.2}Cu_3O_{6.56}$	90.55	97.5	7.6
$Y_1Ba_2Cu_{2.98}Zr_{0.02}O_{6.84}$	90.2	82.5	-8.6
$Y_1Re_{0.1}Ba_2Cu_{2.9}O_{7.08}$	90	92.5	2.8
$Gd_1Ba_2Cu_3O_{6.95}$	90	91.6	1.8
$Bi_2Sr_{2.2}Ca_{0.8}Cu_2O_8$	89.9	96.2	7.0
$Tm_1Ba_2Cu_3O_{6.99}$	89.8	90.6	0.9
$Y_1Ba_{1.95}La_{0.05}Cu_3O_{6.98}$	89.65	94.5	5.4
$Y_1Ba_2Cu_{2.982}Zn_{0.018}O_{6.9}$	89.2	82.9	-7.1
$Bi_2Sr_{1.8}Ca_1Dy_{0.2}Cu_2O_{8.26}$	89.2	90.4	1.4
$Bi_2Sr_2Ca_1Cu_2O_{8.291}$	89.1	83.8	-6.0
$Y_1Ba_2Cu_3O_{7.19}$	89	91.8	3.1
$Y_{0.9}Ca_{0.1}Ba_2Cu_3O_{6.9}$	88.6	82.4	-7.0

[a]The results are sorted by the magnitude of T_c values in descending order. The absolute percentage difference between the predicted values and the experimental measurements are colour-coded in: green $(0-4\%)$, amber $(4-8\%)$ or red $(>8\%)$.

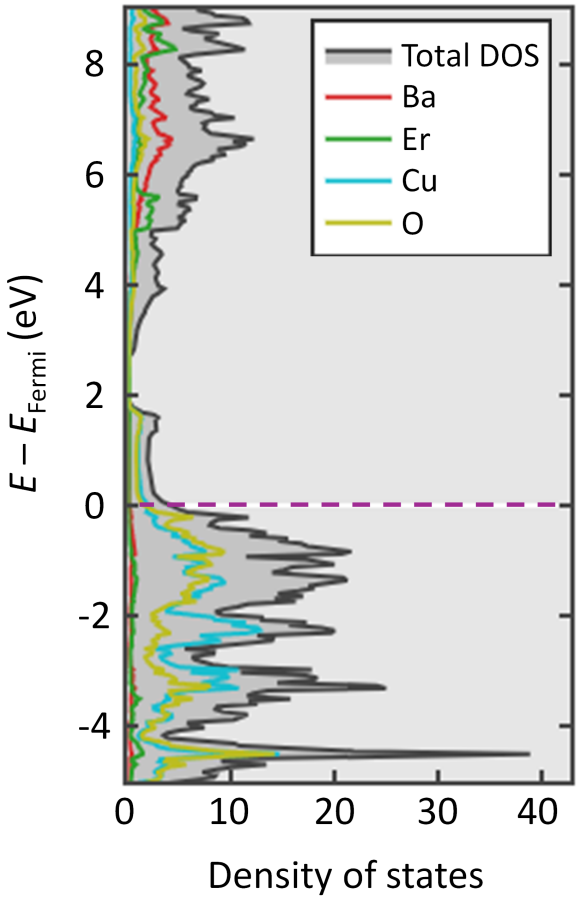


Fig. 4.13 Electronic density of states (DOS) of $Er_1Ba_2Cu_3O_7$ derived from density functional theory calculations sourced from the Materials Project [5].

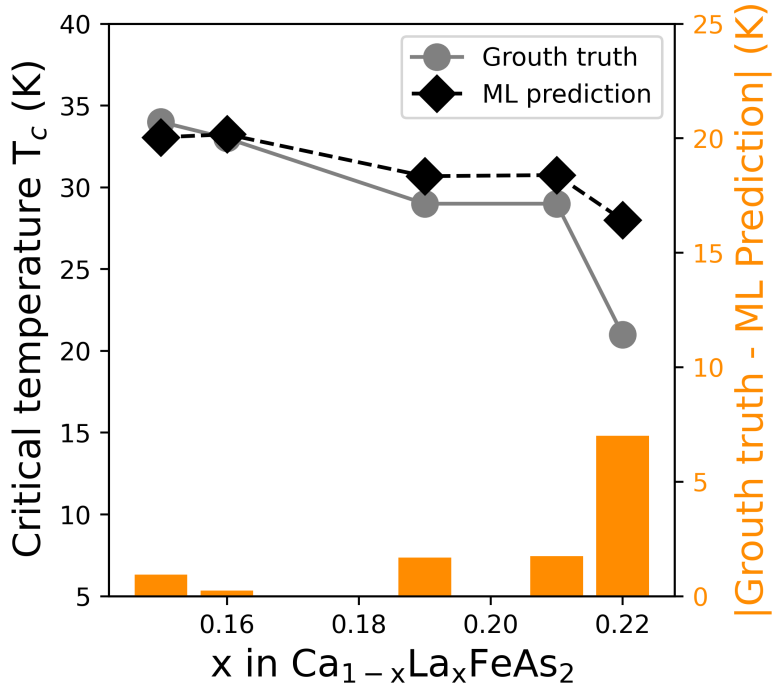


Fig. 4.14 ML-based predictions of the superconducting transition temperature T_c of $(Ca_{1-x}La_x)FeAs_2$ on the La content x against literature values [226]. The orange bar chart illustrates the residuals, defined as the absolute discrepancies between the predicted values (black diamonds) and the ground truth (grey circles) measurements.

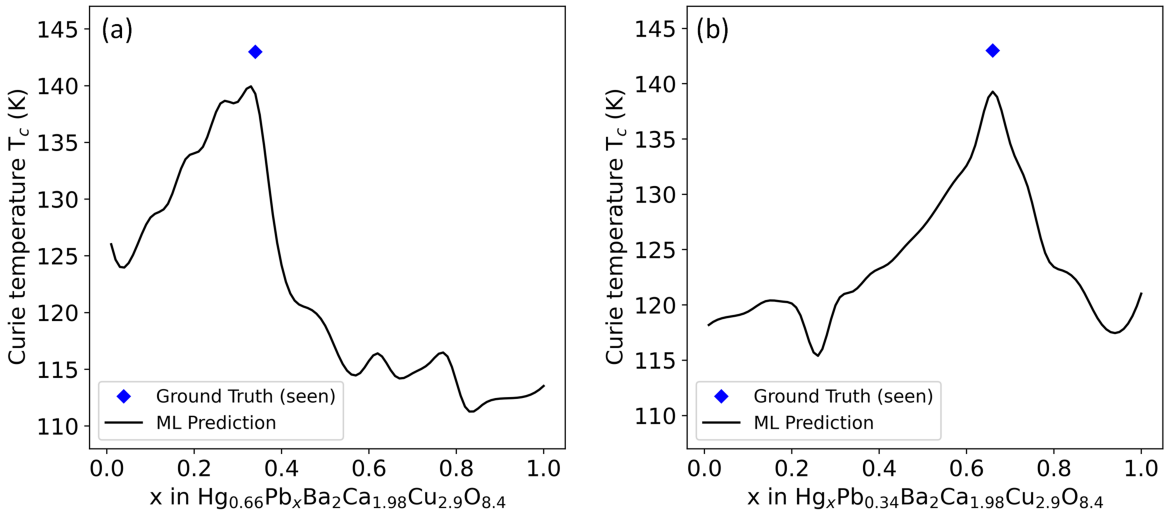


Fig. 4.15 ML-based virtual compositional analysis of Hg-Pb-Ba-Ca-Cu-O series of compounds through independent and systematic variation of the concentrations of (a) Pb and (b) Hg to explore potential trends in the T_c profiles of virtual chemical compounds.

Chapter 5

Property Prediction of Organic Materials

This chapter explores the application of the proposed GBFS workflow for property prediction of organic materials and highlights its versatility when combined with other machine learning algorithms. The effectiveness of the modelling methodology is assessed across diverse data sets, addressing both regression and classification tasks in the context of molecular and drug property prediction. The workflow is extended to incorporate a semi-supervised strategy that leverages substructure vector embeddings alongside the ML-based feature selection process. Additionally, a case study is presented to predict the lipophilicity of chemical molecules, illustrating the robustness and applicability of the proposed approach. To further demonstrate the compatibility of the workflow with other deep learning methodologies, a study is conducted to predict optical properties. This involves the development of a workflow that integrates deep residual convolutional neural networks (CNN) with the gradient-boosting feature selection framework to predict peak optical absorption wavelengths from SMILES representations of dye molecules and solvents. The study enhances prediction accuracy by utilising extensive data sets derived from DFT calculations and experimental measurements. This integrated approach aims to accelerate molecular design targeted toward specific optical characteristics, offering a promising avenue for advancing material discovery and design.

This chapter is based on:

1. **Jung, S. G.**, Jung, G., Cole, J. M., Automatic Prediction of Molecular Properties Using Substructure Vector Embeddings within a Feature Selection Workflow. *J. Chem. Inf. Model.*, 2025, 65 (1), 133-152;

2. **Jung, S. G.**, Jung, G., Cole, J. M., Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks. *J. Chem. Inf. Model.*, 2024, 64 (5), 1486-1501.

Prof. Jacqueline Cole and I conceived the overarching project and designed the study. I created the workflow, designed the CNN architecture, performed data preprocessing, featurisation and hyperparameter optimisation, in addition to analysing the data under the supervision of Prof. Jacqueline Cole. Guwon Jung assisted with the design of the CNN and contributed to the hyperparameter optimisation. All figures and tables are reprinted with permission from J. Chem. Inf. Model., Copyright 2024 American Chemical Society.

5.1 Automatic Prediction of Molecular Properties Using Substructure Vector Embeddings within a Feature Selection Workflow

5.1.1 Introduction

In organic materials, ML models use the SMILES (Simplified Molecular Input Line-Entry System) representation of chemical molecules [39]. SMILES serves as a chemical notation system that enables a precise specification of molecular structures through natural grammar. The adoption of learning techniques on SMILES has been widespread for predicting molecular properties, owing to the succinct representation that it provides. Notably, SMILES explicitly encapsulates the presence of chemical elements, diverse bond types, branches, aromaticity, cyclic structures, and local chirality. Such a representation has paved the way for molecular property prediction through language models within the self-supervised setting [36].

Recently, significant progress has been made in various downstream natural-language-processing tasks by transformer-based language models pretrained on extensive unlabelled data sets. To this extent, Ross et al. [37] introduced molecular embeddings through the training of a MOlecular Language transFORMER (MOLFORMER) that features a linear attention mechanism. This model was trained on sequences of SMILES strings that correspond to a corpus of 1.1 billion unlabelled molecules. Their study illustrated the efficacy of molecular-representation learning and achieved competitive performance against other supervised and unsupervised language models and GNNs. These findings offer promising indications that large-scale molecular

language models can capture sufficient chemical and structural information to predict a range of molecular properties through their fine-tuning on a data set specific to each task. It is noteworthy that studies have not found any discernible shortcomings in SMILES compared to alternative chemical notation systems, such as self-referencing embedded strings [227], in terms of optimisation capability and sample efficiency in molecular optimisation tasks within the learned representation space [228].

However, there are certain issues associated with large-scale molecular language models or other modelling methodologies that rely on a vast corpus of unlabelled data within the unsupervised or self-supervised framework. The main drawback is the demand for extensive computational resources. For instance, MOLFORMER [37] is regarded as an efficient model scalable due to its use of adaptive batch bucketing, parallelisation, and a linear attention mechanism. Its pretraining process used 16 graphics processing units (GPUs) and required 208 hours to complete 4 epochs of loss reduction with a batch size of 1600 molecules per GPU. Without bucketing and linear attention, batch sizes would be constrained to fewer than 50 molecules per GPU, necessitating over 1000 GPUs.

The primary objective of this study is to validate the broader applicability of our proposed workflow as a general-purpose tool across various research domains, with a particular emphasis on its utility in organic chemistry and drug discovery. This study illustrates the adaptability of our workflow across different areas of organic chemistry, a capability that has not been previously demonstrated. With enhancements that accommodate organic compounds, our workflow can now be applied more extensively across a range of scientific fields and can be seamlessly integrated with other computational methods, such as vector embeddings. To illustrate the versatility of our methodology, we explored an array of more than 20 diverse material properties, demonstrating that it extends beyond the constraints of specific domains. We enriched our data set by incorporating a variety of sources from the scientific literature, which enables direct comparisons and emphasises the efficacy of our modelling approach in tackling complex predictive tasks, including those related to quantum mechanical properties. Moreover, we utilised data sets originating from experimental sources. Our models adeptly predict properties using only the SMILES notation as input. This capability is exemplified in a case study on lipophilicity, which showcases the adaptability of our workflow in developing and implementing domain-specific models.

Additionally, we seek to understand the impact of improving the predictability of ML models by incorporating systematic analyses at the feature level. This approach contrasts popular practices that prioritise intricate model development, often at the

cost of increased computational demand. Our methodology emphasises the significance of meticulous feature analysis and selection, as opposed to relying solely on complex modelling techniques. We illustrate that a simple tree-based model trained on SMILES-based features, which have been selected and engineered by our GBFS workflow using a single central processing unit (CPU) such as the AMD Ryzen 9 or Intel Core i9, can produce results comparable or superior to those reported in the literature, many of which require multiple CPUs, GPUs or high-performance supercomputers. ML methods or workflows that realise such performance gains while requiring considerably fewer computational resources are vital in light of the global need to achieve energy sustainability. Moreover, our workflow offers valuable insights into feature interactions and their importance in predicting target variables, thus advancing our understanding of material properties.

We further explore methods to improve the predictability of our models by integrating substructure vector embeddings generated by an unsupervised ML approach. We adopt the implementation of Jaeger et al. [229] to learn vector representations of molecular substructures. These vector representations are designed to align in similar directions for chemically related substructures, with the final embeddings being computed as a vector addition of the individual substructures. The underlying concept is that vectors of closely related SMILES representations are positioned close to each other in the vector space. We will demonstrate an enhancement in the predictive performance of our workflow via the inclusion of such substructure vector embeddings. This leads to results that outperform many supervised or unsupervised state-of-the-art benchmarks in molecular-property predictions.

We have provided the code that supports the creation of models applicable to any data set, facilitating the prediction of novel chemical properties. This capability is exemplified through specific implementations for inorganic compounds, as detailed in the work of Jung et al. [123]. In this study, we have enhanced the utility to encompass organic compounds and integrated the use of chemical descriptors derived from an unsupervised learning framework, specifically Mol2Vec [229]. Consequently, our code surpasses domain-specific constraints and is adaptable across various research environments. Furthermore, the integration of Mol2Vec underscores the compatibility of our workflow with established methodologies, demonstrating that our tool is not merely an isolated tool but is designed to be seamlessly integrated into broader research ecosystems.

5.1.2 Methods

Data Sets

The data sets employed in this study were compiled from diverse literature sources, as summarised in Table 5.1. Among the data sets used is the quantum-chemistry QM9 data set [230], which is a subset of the GDB-17 database [231] and comprises about 130,000 stable organic chemical compounds. These compounds consist of H, C, F, N, and O atoms, with a maximum of 9 heavy atoms in each molecule. The QM9 data set also contains 12 molecular properties that have been computed for each molecule at the B3LYP/6-31G(2df,p) level of quantum chemistry. These properties are summarised in Table 5.2. In contrast to studies by others, we did not include geometric properties such as atomic coordinates in our features. Instead, we limited our features to only those derived from the SMILES representation. This decision was made considering that comprehensive three-dimensional (3D) molecular geometric information is often absent in the literature, so removing any reliance on such information streamlines the materials-property prediction process. Such a methodological choice also allows us to evaluate the efficacy of our modelling approach, given that predicting quantum-chemical properties solely from SMILES representation is considered non-trivial.

Other data sets employed in our study include: (i) Estimated SOLubility (ESOL) [232], a data set of water solubility data; (ii) Free Solvation Database (FreeSolv) [233], a data set of experimentally measured and calculated hydration free energy of small molecules in water; (iii) Lipophilicity [234, 235], a data set of experimental results on the octanol-water distribution coefficient of molecules, which is a feature of a drug molecule that concerns membrane permeability and solubility; (iv) Photovoltaic Efficiency (PE) [236], a data set of photovoltaic efficiency of organic molecules; (v) Human Immunodeficiency Virus (HIV) [237], a data set sourced from the Drug Therapeutics Program AIDS Antiviral Screen to test the ability of drugs to inhibit HIV replication; (vi) β-Secretase (BACE) [238], a data set of experimental binding results for a set of inhibitors for β-secretase 1; (vii) Blood-brain-barrier Penetration (BBBP) [239], a data set about the blood-brain barrier penetration or permeability properties of chemical compounds; (viii) Toxicity-21 (Tox21) [240, 241], a data set of toxicity measurements on 12 different targets, including nuclear receptors and stress response pathways; (ix) Side Effect Resource (SIDER) [235, 242], a data set of marketed drugs and adverse drug reactions, delineating the side effect of drugs on 27 different organ classes; (x) Clinical Toxicity (ClinTox) [235, 243–245], a data set consisting of data on clinical trial

toxicity and Food and Drug Administration (FDA) approval status of drugs. Once again, we confine our features to those generated from SMILES representations alone.

The data-splitting method employed for these data sets precisely replicates that of Ross et al. [37], who implemented a train-to-test split ratio of ca. 9:1. The training and the test sets are, therefore, identical for each task that we perform unless explicitly stated otherwise. Consequently, this enables a like-for-like comparison to the literature benchmarks, and the performance metrics reported in this study offer a direct comparison to the findings of Ross et al. The sole exception to this procedure is in a case study that examines the lipophilicity of chemical molecules, where we intentionally reduced the size of the training set to impose more rigorous criteria for model assessment. We chose to expand the size of that test set to include a broader range of chemical molecules, increasing the proportion of the test set to 20% of the data set, up from its original 10%. This adjustment allows for a more thorough assessment and comprehensive evaluation of our proposed approach. Moreover, we have utilised the scaffold splitting method for the analyses involving the following data sets: BBBP, Tox21, ClinTox, HIV, Bace, and SIDER. A scaffold-splitting strategy, based on molecular substructures, imposes a more rigorous test on the ability of the model to generalise to new, unseen data.

More specifically, to evaluate the generalisation capabilities of our ML models beyond the training distribution, we employed a scaffold-splitting strategy as delineated by Ramsundar et al. [246]. This method involves grouping molecules based on their molecular substructures, or scaffolds, as defined by Bemis and Murcko [247]. The most prevalent scaffolds are placed into the training set, while the validation and test sets are composed of structurally distinct molecules. This scaffold splitting approach presents a significant challenge to learning algorithms, far exceeding that of random splits, by guaranteeing that structurally unique subsets are explicitly separated. The implementation of such a splitting methodology is available in the RDKit library [72].

Gradient-boosted statistical feature selection workflow

The GBFS-Mol2Vec implementation that we employed in this work is portrayed in Figure 5.1, with the GBFS workflow introduced in Chapter 2.2. Substructure vector embeddings were generated using a molecule-to-vector (Mol2Vec) technique [229], which takes SMILES strings as inputs. These embeddings were then directly incorporated into the model optimisation phase, being concatenated with the descriptor features created via our GBFS workflow.

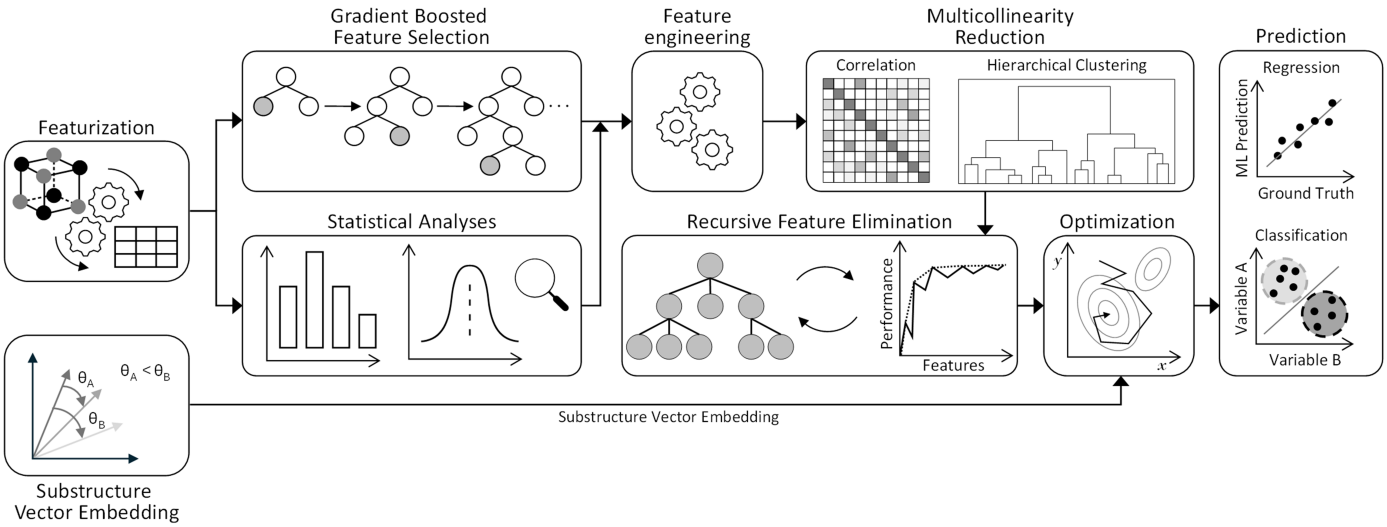


Fig. 5.1 The operational workflow of our GBFS-Mol2Vec pipeline. For an in-depth explanation of the processes involved, the reader is referred to Jung et al. [123]. Adapted with permission from Ref. 123, Copyright 2023 AIP Publishing.

Table 5.1 Summary of data sets used in this study

Subject Category	Name of Dataset	No. & Type of Tasks		No. of Compounds	Description of Dataset Contents
Quantum mechanics	QM9	12	Regression	133,885	12 quantum mechanical calculations of organic molecules
Physical chemistry	ESOL	1	Regression	1,127	Water solubility
	FreeSolv	1	Regression	641	Hydration free energy of molecules in water
	Lipophilicity	1	Regression	4,200	Octanol-water distribution coefficient of molecules
Photovoltaics	PE	1	Regression	29,978	Photovoltaic efficiency of organic molecules
Biophysics	HIV	1	Classification	41,126	Ability of drugs to inhibit HIV replication
	BACE	1	Classification	1,512	Binding results for a set of inhibitors for β-secretase 1
Physiology	BBBP	1	Classification	2,038	Blood-brain barrier penetration ability of compounds
	Tox21	12	Classification	7,830	Toxicity measurements of compounds on 12 different targets
	SIDER	27	Classification	1,426	Side effect of drugs on 27 different organ classes
	ClinTox	2	Classification	1,476	Clinical trial toxicity and FDA approval status of drugs

Table 5.2 Chemical properties contained within the QM9 data set

No.	Property	Description
1	μ	Dipole moment
2	α	Isotropic polarizability
3	E_{HOMO}	Energy of highest occupied molecular orbital (HOMO)
4	E_{LUMO}	Energy of least unoccupied molecular orbital (LUMO)
5	E_{gap}	Gap ($E_{\text{LUMO}}E_{\text{HOMO}}$)
6	$< R^2 >$	Electronic spatial extent
7	ZPVE	Zero point vibrational energy
8	U_0	Internal energy at 0 K
9	U	Internal energy at 298.15 K
10	H	Enthalpy at 298.15 K
11	G	Free energy at 298.15 K
12	C_v	Heat capacity at 298.15 K

5.1.3 Results and Discussion

Evaluation of model performance

We assessed the efficacy of our modelling technique using 5 data sets across 16 regression tasks, and 6 data sets across 44 classification tasks. Our findings are compared with those of state-of-the-art reports in the following discussion.

Regression tasks - Table 5.3 presents the comparative performance of various material-property predictions from different ML models on the QM9 test set, as judged by reports of MAE values. Each row corresponds to a specific material property, while each column represents a different ML model utilised for the predictions. The ML models evaluated comprise: an attentive-FP (A-FP) [248] model; a graph convolutional (GC) network [249]; a multitask neural network encoding the Coulomb Matrix (CM) [30]; a deep-tensor neural network (DTNN) [250]; a message-passing network (MPNN) [35]; Chemical RoBERTa (ChemBERTa) [36]; MOLFORMER (QM9 only) [37]; MOLFORMER-XL [37]; GBFS; GBFS-Mol2Vec. Here, GBFS refers to our original GBFS workflow [123], while GBFS-Mol2Vec denotes our GBFS workflow that incorporates substructure vector embedding.

In general, MOLFORMER-XL, GBFS-Mol2Vec, and GBFS exhibit the most favourable average MAE values across the spectrum of material properties, which indicates their superior predictive capabilities compared to the other evaluated models. Conversely, models such as CM and ChemBERTa display higher average MAE values across the 12 properties, suggesting relatively inferior performance in predicting the quantum-chemical properties of organic molecules. Notably, our proposed modelling approach outperforms the MOLFORMER (QM9 only) model in all prediction tasks. The MOLFORMER (QM9 only) model is a molecular-language transformer model trained solely on the QM9 data set, while MOLFORMER-XL has been pre-trained on a data set comprising ca. 1.1 billion molecules; these include data from PubChem [251] and ZINC [252]. It is unsurprising to observe that MOLFORMER-XL achieves the lowest average MAE, although it is noteworthy that the GBFS-based models appear to perform with greater accuracy than the MOLFORMER approach when the data set is limited to QM9 alone. There exists a clear trade-off between model predictability and the volume of data employed. It is important to remember that the use of more data results in higher computational demands. This is not to disregard the advantages of employing self-supervised learning with a vast corpus of unlabelled data. However, it is imperative to select the appropriate modelling method according to

energy-sustainability considerations, as well as the practical availability and use of computational resources for these prediction tasks.

A closer examination of these MAE statistics reveals that the DTNN and MPNN models afford commendable performance in predicting the dipole moment (μ), with respective MAE values of 0.244 D and 0.358 D. Similarly, the DTNN model exhibits the lowest MAE in predicting the electronic spatial extent ($< R^2 >$). These results are anticipated given that both models are geometry-based, integrating 3D molecular geometric data into their feature sets. Furthermore, we observe that our GBFS-based models both achieve markedly lower MAE values than alternative methods when predicting internal energies at 0 K and 298.15 K (U and U_0), enthalpy at 298.15 K (H), and free energy at 298.15 K (G). These MAE values are also on par with those obtained by the higher-order GNN model developed by Morris et al. [253] referred to as 1-2-3-GNN, who achieved MAE values of a similar magnitude. Nevertheless, it is worth noting that the more expressive and potent 1-2-3-GNN-based models are difficult to scale [254], even when compared to MOLFORMER-XL, which employs a linear attention mechanism to ensure linear time complexity [37]. Despite the predictive capabilities of the 1-2-3-GNN, we have decided to exclude this model from our comparative analysis hereafter, as it is well-documented that such networks are challenging to scale.

Table 5.3 also offers valuable insights into the relative efficacy of diverse models in forecasting various material properties. Forecasting ability is crucial when choosing the most appropriate model for specific applications in materials science. It is essential to highlight that the selection of an ML model can profoundly influence the precision of material-property predictions. Overall, our GBFS and GBFS-Mol2Vec models emerge as promising modelling methodologies in the prediction of quantum-mechanical calculations of organic molecules; cf. statistics in Table 5.3 demonstrate their competitive performance across numerous material properties and models, and do so despite the fact that they do not require significant computational resources.

Before delving further into our analysis, it is necessary to be aware of conformational effects in the context of molecular chemistry [76]. These effects manifest as variations in atomic arrangements within a molecule, stemming from different spatial orientations. While spectroscopic data can offer insights into these conformational effects by representing a Boltzmann average of all conformers present in a solution, SMILES strings lack the capacity to account for such variations. Given that this study does not differentiate between cases involving single or multiple conformations in the calculations, nor addresses potential modifications to the conformational ensemble

due to interactions with the environment, it is imperative to consider these factors when citing or referencing the methodology and the findings presented herein. This is pertinent not only to this study but also to other research endeavours that employ SMILES representations; subject to the caveat that this may not be applicable to certain properties such as solubility, which is a property of a molecule rather than its specific conformers.

In addressing the limitations of SMILES representations for capturing the diverse spatial orientations and conformations of molecules, evaluating the implications for predictive accuracy in molecular modelling becomes critical. The inherent inability of SMILES to account for conformational variability restricts the capability of ML models to discern spatial variations. This limitation is particularly detrimental for predicting molecular properties that are profoundly influenced by molecular geometry and conformation, such as reactivity, which may exhibit substantial variation with changes in molecular conformation. As SMILES representations lack the capacity to convey information on molecular dynamics or energetically favourable structures, models that rely exclusively on SMILES representations are prone to oversimplifying molecular complexity, potentially leading to inaccurate predictions by generating uniform outputs irrespective of the conformational state of a molecule. This shortfall is notably critical in pharmaceutical research and materials science, where a comprehensive understanding of a molecule's conformational ensemble is essential for accurately predicting interactions with biological targets or other chemical entities. Without the ability to account for the various conformations that a molecule may adopt, SMILES-based models are likely to overlook crucial interactions, thus hindering effective drug design and materials synthesis.

Additionally, the lack of environmental consideration further constrains the generalisability of these models. Molecular conformations are susceptible to alterations during interactions with solvents or in complex biological formations. Consequently, without adaptations or additional data to accommodate these interactions, SMILES-based models will tend to underperform when extrapolated beyond the specific conditions or data sets on which they were originally trained. While SMILES representations may suffice for modelling certain properties such as solubility, which is generally less sensitive to precise atomic configurations as mentioned earlier, they are inadequate for applications that demand detailed 3D structural data. Therefore, the utility of SMILES representations is restricted to a subset of molecular properties and should be selectively applied based on the property under examination.

In light of these considerations, it is imperative for researchers to exercise caution in using SMILES representations as the sole structural representation in molecular property modelling. There is a pressing need to incorporate additional data forms that can capture 3D structural details and dynamic conformational changes. Employing techniques that integrate molecular descriptors reflecting 3D spatial or geometric information could substantially enhance the predictive accuracy of ML models. For achieving comprehensive and reliable predictions, a hybrid approach that amalgamates multiple molecular representations is often essential. This topic is suggested for further exploration as a potential avenue for subsequent research. In this study, we exclusively utilise the SMILES representations to facilitate a direct like-for-like comparison with state-of-the-art benchmarks.

Additional analyses were conducted to demonstrate the versatility of the proposed methodology as a general-purpose tool. Table 5.4 presents an overview of the model performance measured by root-mean-squared error (RMSE) across various regression tasks, encompassing data from repositories about water solubility (ESOL), hydration free energy in water (FreeSolv), and the octanol-water distribution coefficient of molecules (Lipophilicity). The models assessed in this study include a range of approaches: GC [249]; A-FP [248]; MPNN [35]; a variant of MPNN known as DimeNet, which integrates spatial interactions within molecules and incorporates directional message passing [255]; a self-supervised molecular-graph-representation model pre-trained using 2D graph topology and 3D conformal geometry (GeomGCL) [256]; a geometry-enhanced molecular representation learning (GEM) [257], which incorporates atom-bond-angle relations in the graph; MOLFORMER-XL [37]; and our GBFS and GBFS-Mol2Vec models.

For the regression tasks that aim to predict the water solubility (ESOL) of molecules or their hydration free energy in water (FreeSolv), the MOLFORMER-XL model displayed the lowest RMSE, followed by the A-FP model and our GBFS or GBFS-Mol2Vec model. The remarkable performance of the MOLFORMER-XL model in these tasks is possibly due to its use of a vast data set that was sourced from PubChem [251] and ZINC [252], as mentioned earlier. Our GBFS-based models performed competitively, positioning between the supervised A-FP and the geometry-informed, self-supervised GeomGCL models. However, geometry-informed models may not be universally applicable or generalisable in scenarios where 3D molecular geometric data are not readily available. This limitation combined with the higher computational requirements of these models could pose challenges, especially for large molecules.

Concerning the regression-based task of predicting lipophilicity, our GBFS-Mol2Vec model achieved the lowest RMSE value, closely followed by GBFS and MOLFORMER-XL. The GBFS-based models and the MOLFORMER-XL model rely exclusively on SMILES representations, and they outperform all other aforementioned models that rely on geometry-informed approaches in the forecast of the octanol-water distribution coefficient of molecules. Furthermore, these results clearly demonstrate the generalisability of our GBFS-based models.

The MSE values corresponding to different ML models employed for the prediction of photovoltaic efficiency (PE) were also analysed. Table 5.5 displays the findings which compare six ML models: neural fingerprints (NFPs) [32], designed to extract data-driven features by employing a back-propagation algorithm; molecular graph convolutional networks (GCNs) [258], which leverage a deep-learning architecture to represent small molecules as undirected graphs of atoms; both the bidirectional long short-term memory (BiLSTM) and self-attention mechanism BiLSTM (SA-BiLSTM) [259]; and our two GBFS-based models. Among the models evaluated, the GBFS model stands out with the lowest MSE, closely followed by GBFS-Mol2Vec and GCN models. These results suggest that both of our GBFS-based models exhibit excellent performance in predicting photovoltaic efficiency. Conversely, models such as NFP and BiLSTM show relatively higher MSE values, indicating comparatively poorer predictive accuracy for this particular material property. Overall, these findings provide valuable insights into the efficacy of different modelling approaches in the context of photovoltaic efficiency prediction, which requires an analysis of a complex feature space. This may help to explain our analytical findings given that a major challenge in neural-network-based models is the creation of unintuitive features that are often hard to interpret. In contrast, our GBFS-based models do not depend on this type of feature learning process.

Considering the overarching analysis of these regression tasks, the MOLFORMER-XL model consistently exhibits competitive performance, particularly excelling in predicting water solubility and hydration free energy in water. The GBFS-Mol2Vec model also shows promising results, especially in predicting lipophilicity and photovoltaic efficiency. These findings underscore the effectiveness of self-supervised models, such as the MOLFORMER-XL model, as well as supervised and semi-supervised approaches, such as our GBFS and GBFS-Mol2Vec models, in capturing complex molecular properties for diverse research applications in materials science. Such findings demonstrate the versatility and applicability of our proposed methodology.

Classification tasks - Table 5.6 presents the AUC-ROC values for various ML models across different classification tasks using scaffold splitting. These tasks employed data sets that inform blood-brain barrier penetration (BBBP), toxicity in the 21st Century (Tox21) on 12 different targets, clinical trial toxicity and FDA approval status of drugs (ClinTox), ability of a drug to inhibit HIV replication (HIV), binding results for inhibitors for β-secretase 1 (BACE), and the side effect of marketed drugs and adverse drug reactions (SIDER).

In the evaluation of these classification tasks, we included random forest (RF) [177] and support vector machine (SVM) [261] classifiers trained on molecular fingerprints, in addition to the previously mentioned ML models, as well as other graph-based networks and self-supervised models. These include a multilevel graph convolutional network (MGCN) [262] designed to process molecular graphs, preserving their intrinsic structure, and it uses a hierarchical graph-neural network to extract features directly from the spatial and conformational information, followed by multilevel interactions; directed-MPNN (D-MPNN) [263], which differs from the generic MPNN, by using messages associated with directed edges instead of messages linked to vertices; a graph-isomorphism network (GIN) [264] which is a graph-neural network pre-trained on molecular graphs with edge features that uses a multilayer perceptron and a weighted sum of node features; the molecular contrastive learning representation (MolCLR) [38] which is a self-supervised learning framework based on GIN with contrastive loss; N-Gram [265] that employs an unsupervised approach for molecular representation, embedding nodes in a graph and constructing a representation by assembling the vertex embeddings in short graph walks; and GraphMVP-C, the graph multi-view pre-training framework [256], which is a self-supervised learning framework that leverages the relationship between 2D topological structures and 3D geometry views.

The AUC-ROC analysis reveals interesting insights into performance variation between different models across a range of classification tasks. Among the existing supervised and semi-supervised ML models (i.e., excluding our GBFS and GBFS-Mol2Vec workflows), RF and SVM models trained on molecular fingerprints demonstrate moderate performance across most tasks, with varying degrees of accuracy. They scored relatively high AUC-ROC scores on data sets such as Tox21, HIV, BACE and SIDER compared to MGCN, D-MPNN, and DimeNet. These models achieve this performance despite their lack of awareness regarding 3D molecular geometric information. Additionally, we observe that the semi-supervised MGCN model and the supervised D-MPNN model outperform RF, SVM and the supervised DimeNet models on BBBP and ClinTox data sets; while these models generally yield satisfactory results,

Table 5.3 Performance of various ML models on the QM9 test set. MAE values are colour-coded in shades of pink, green, and blue to denote the top-performing, second-best, and third-best models, respectively. Performance benchmarks were obtained from multiple sources [37, 235, 248, 260] .

Model →	Graph-Based		Geometry-Based			SMILES-Based				
Property ↓	A-FP[248]	GC[249]	CM[30]	DTNN[250]	MPNN[35]	ChemBERTa[36]	MOLFORMER (QM9 only)[37]	MOLFORMER-XL[37]	GBFS	GBFS-Mol2Vec
μ	0.451	0.583	0.519	0.244	0.358	0.4659	0.4380	0.3616	0.4820	0.4557
α	0.492	1.37	0.85	0.95	0.89	0.8510	0.5078	0.3327	0.4868	0.4746
E_{HOMO}	0.00358	0.00716	0.00506	0.0038	0.00541	0.0044	0.0042	0.0029	0.0038	0.0035
E_{LUMO}	0.00415	0.00921	0.00645	0.0051	0.00623	0.0041	0.0041	0.0027	0.0038	0.0037
E_{gap}	0.00528	0.01126	0.0086	0.0112	0.0066	0.0052	0.0057	0.0038	0.0049	0.0046
$< R^2 >$	26.839	35.97	46.00	17.00	28.5	86.150	24.0785	17.062	23.726	22.877
ZPVE	0.00207	0.00299	0.00207	0.0017	0.00216	0.0023	0.0011	0.0003	0.00028	0.00023
U_0	0.898	3.41	2.27	2.43	2.05	3.9811	1.1462	0.3211	0.0585	0.0421
U	0.893	3.41	2.27	2.43	2.00	4.3768	1.0454	0.2522	0.0618	0.0427
H	0.893	3.41	2.27	2.43	2.02	4.0853	1.1579	0.2522	0.0593	0.0476
G	0.893	3.41	2.27	2.43	2.02	4.1295	0.9985	0.3362	0.0685	0.0450
C_v	0.252	0.65	0.39	0.27	0.42	0.4234	0.1589	0.1447	0.1564	0.1588
Avg MAE	2.6355	4.3536	4.7384	2.3505	3.1899	8.7066	2.4622	1.5894	2.0927	2.0130

Table 5.4 Model performance on addition regression tasks. Model performance was assessed using RMSE, with colour-coding indicating the top-performing model in pink, second-best in green, and third-best in blue. Performance benchmarks were obtained from various studies [37, 235, 248] .

Model ↓ Dataset →	ESOL	FreeSolv	Lipophilicity
GC[249]	0.970	1.40	0.655
A-FP[248]	0.503	0.736	0.578
MPNN[35]	0.58	1.150	0.719
DimeNet[255]	0.633	0.978	0.614
GeomGCL[256]	0.575	0.866	0.541
GEM[257]	0.798	1.877	0.660
MOLFORMER-XL[37]	0.2787	0.2308	0.5289
GBFS	0.5608	0.8428	0.5128
GBFS-Mol2Vec	0.5456	0.8656	0.4989

Table 5.5 Model performance on photovoltaic efficiency of organic molecules. Model performance was assessed using MSE, with colour-coding indicating the top-performing model in pink, second-best in green, and third-best in blue.

Model ↓ Dataset →	PE
NFP[32]	1.43
GCN[258]	1.10
BiLSTM[259]	1.4
SA-BiLSTM[259]	1.18
GBFS	1.098
GBFS-Mol2Vec	1.125

they may fall short in terms of the robustness required for precise predictions across all tasks.

We note that existing self-supervised or unsupervised ML methodologies, including pre-trained molecular language models, generally outperform the aforementioned supervised ML models, exhibiting higher AUC-ROC scores across all classification tasks. For example, the MOLFORMER-XL model consistently ranks in the top three for AUC-ROC scores in five out of six classification tasks. Additionally, there are notable performances that surpass the MOLFORMER-XL model. The N-Gram model achieves one of the highest AUC-ROC scores on the HIV data set, while the MolCLR model records an AUC-ROC of 89.0 on the BACE data set, and the geometry-aware GeomGCL model attains the highest AUC-ROC score on the Tox21 data set. These scores underline their efficacy in capturing intricate molecular and drug properties across various data sets. In particular, the MOLFORMER-XL model excelled in achieving exceptionally high AUC-ROC values across all tasks, highlighting its strength and effectiveness in prediction across a variety of data sets.

We now introduce GBFS and GBFS-Mol2Vec into this comparative analysis and find that they emerge as top-performing ML models, demonstrating high performance across all classification tasks. Our models showcase superior predictive capability, particularly evident in classification tasks related to HIV, BACE and SIDER data sets. This highlights their versatility in capturing complex chemical-structural relationships with the target classes. Notably, both GBFS-based models consistently rank among the top-three performers across all data sets, as indicated by the colour-coded results in Table 5.6. While graph-based models such as GraphMVP-C, GeomGCL, and GEM show promising performance in specific tasks, they may not consistently demonstrate the same level of robustness across all data sets when compared to the MOLFORMER-XL, GBFS, and GBFS-Mol2Vec models.

In summary, the results of this analysis emphasise the importance of model selection in classification tasks. They also demonstrate the effectiveness of sophisticated molecular language models, such as MOLFORMER-XL, alongside supervised or semi-supervised methods, such as our GBFS and GBFS-Mol2Vec workflows, in capturing intricate chemical-structural characteristics across diverse data sets. These findings have significant implications for drug discovery and toxicity assessment in pharmaceutical research and development, thereby paving the way for diverse modelling strategies.

Table 5.6 Model performance on classification tasks. Model performance was assessed using AUC-ROC on scaffold splits, with colour-coding indicating the top-performing model in pink, second-best in green, and third-best in blue. Performance benchmarks were obtained from multiple sources [36–38, 256] .

Model ↓ Dataset →	BBBP	Tox21	ClinTox	HIV	BACE	SIDER
RF[177]	71.4	76.9	71.3	78.1	86.7	68.4
SVM[261]	72.9	81.8	66.9	79.2	86.2	68.2
MGCN[262]	85.0	70.7	63.4	73.8	73.4	55.2
D-MPNN[263]	71.2	68.9	90.5	75.0	85.3	63.2
DimeNet[255]	-	78.0	76.0	-	-	61.5
GIN[264]	70.8	78.7	78.9	80.2	85.9	65.2
N-Gram[265]	91.2	76.9	85.5	83.0	87.6	63.2
MolCLR[38]	73.6	79.8	93.2	80.6	89.0	68.0
GraphMVP-C[256]	72.4	74.4	77.5	77.0	81.2	63.9
GeomGCL[256]	-	85.0	91.9	-	-	64.8
GEM[257]	72.4	78.1	90.1	80.6	85.6	67.2
ChemBERTa[36]	64.3	-	90.6	62.2	-	-
MOLFORMER-XL[37]	93.7	84.7	94.8	82.2	88.2	69.0
GBFS	91.5	83.6	93.7	83.4	88.6	71.5
GBFS-Mol2Vec	91.8	83.6	94.2	83.4	90.3	72.6

5.2 Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks

5.2.1 Introduction

We propose a deep learning method that generates learned representations of dye molecules and their cognate solvents using deep residual convolutional neural networks (DR-CNNs), where the inputs to the method are feature matrices that are created exclusively from SMILES strings. We adopt a multifidelity modelling approach to train an ML model on experimental measurements, whose predictive accuracy is improved by auxiliary ML models derived using computational calculations of peak vertical excitation energies in an optical spectrum. Simultaneously, we employ a gradient-boosted and statistical feature selection (GBFS) workflow for material property predictions [123] to identify descriptor features that afford minimal feature redundancy and maximal relevance to the target variable. The incorporation of such a workflow further reinforces the predictability of the final predictive model based on the Bayesian-optimised gradient-boosting algorithm. The proposed methods are benchmarked against the state-of-the-art methods reported in the scientific literature, and we demonstrate the efficacy and generalisability of the aforementioned methods by applying them on several of the most extensive open-source experimental data sets using two splitting strategies.

Both the GBFS and DR-CNN workflows are designed for general-purpose use, emphasising their versatility beyond the confines of this study. Our paper on the GBFS workflow [123] showcased its adaptability by predicting various material properties. These predictions were achieved by fine-tuning selected features or representations to different target variables, all within the same workflow. A comparable adaptability is inherent in DR-CNN, designed to generate deep feature representations for a given SMILES string. The overarching nature of our approach, geared toward general applicability, allows us to employ our workflows to a broad spectrum of properties across diverse projects. This encompasses the scope of the present study, whereby these representations are fine-tuned for the UV–vis spectroscopic property of interest, enabling the prediction of optical absorption peaks against their experimental measurements.

5.2.2 Methods

Workflow Overview

The operational workflow used in this study is shown in Figure 5.2. There are two distinct subworkflows. These include (i) the GBFS workflow based on our previous work [123] that has been configured to identify a subset of features which affords minimal feature redundancy and maximal relevance to TD-DFT-calculated peak vertical optical excitation energies and (ii) the deep learning method using DR-CNNs to generate a deep feature representation (DFR) of dye molecules and their associated solvents for the prediction of molecular absorption maxima in optical spectra. The feature matrix that comprises the input to the DR-CNNs is generated solely from the SMILES strings of the dye molecules and their associated solvents, as illustrated in Figure 5.3.

The aforementioned subworkflows are utilised to process both TD-DFT and experimental data simultaneously. We adopt a multifidelity approach, whereby four models are initially generated in parallel: two from computational data and two from experimental data; these are eventually consolidated into one final ML-based predictive model. On the one hand, the auxiliary DR-CNN outputs a DFT-aware DFR, while an auxiliary GBFS model identifies a subset of selected descriptor features that achieve the maximal loss reduction in predicting the TD-DFT peak vertical optical excitation energies. On the other hand, additional DR-CNN and GBFS models are trained using experimental measurements, and they output an experiment-aware DFR as well as a subset of selected descriptor features that realise the maximal loss reduction; this maximises their relevance to experimental measurements of the optical absorption peaks. For a given SMILES string, the output of the four models is concatenated. This is fed into the final predictive model based on a gradient-boosting algorithm and is Bayesian-optimised using Gaussian processes [52, 53]. Note that a different ML algorithm can be used as the predictive model. In this study, we employed the light gradient-boosting machine, originally developed by Microsoft [42].

The training and optimisation of our models were conducted in the energy domain. This choice aimed to alleviate nonlinear artifacts, particularly in error calculations, and to avoid the reported performance from exhibiting nonlinearity within the spectral domain considered herein. Performing the model training in the wavelength domain would entail assigning equal importance to both small and large errors in energy, contradicting the intended functionality of an objective function. While this practice is marginally acceptable for spectra that absorb red light with errors ranging from ca. 50 to 100 meV, it is notably inadequate for spectra that absorb blue light, where

errors surpass ca. 0.5 eV. This disparity in the magnitude of the error has substantial consequences during model training, hence the decision to train in the energy domain. Nevertheless, we present most of the results in wavelength units, as it is the preferred unit for expressing UV–vis absorption spectra among chemists and spectroscopists. Additionally, by reporting in wavelength, we can benchmark our performance directly against relevant studies in the literature, where results are often quoted in wavelength units.

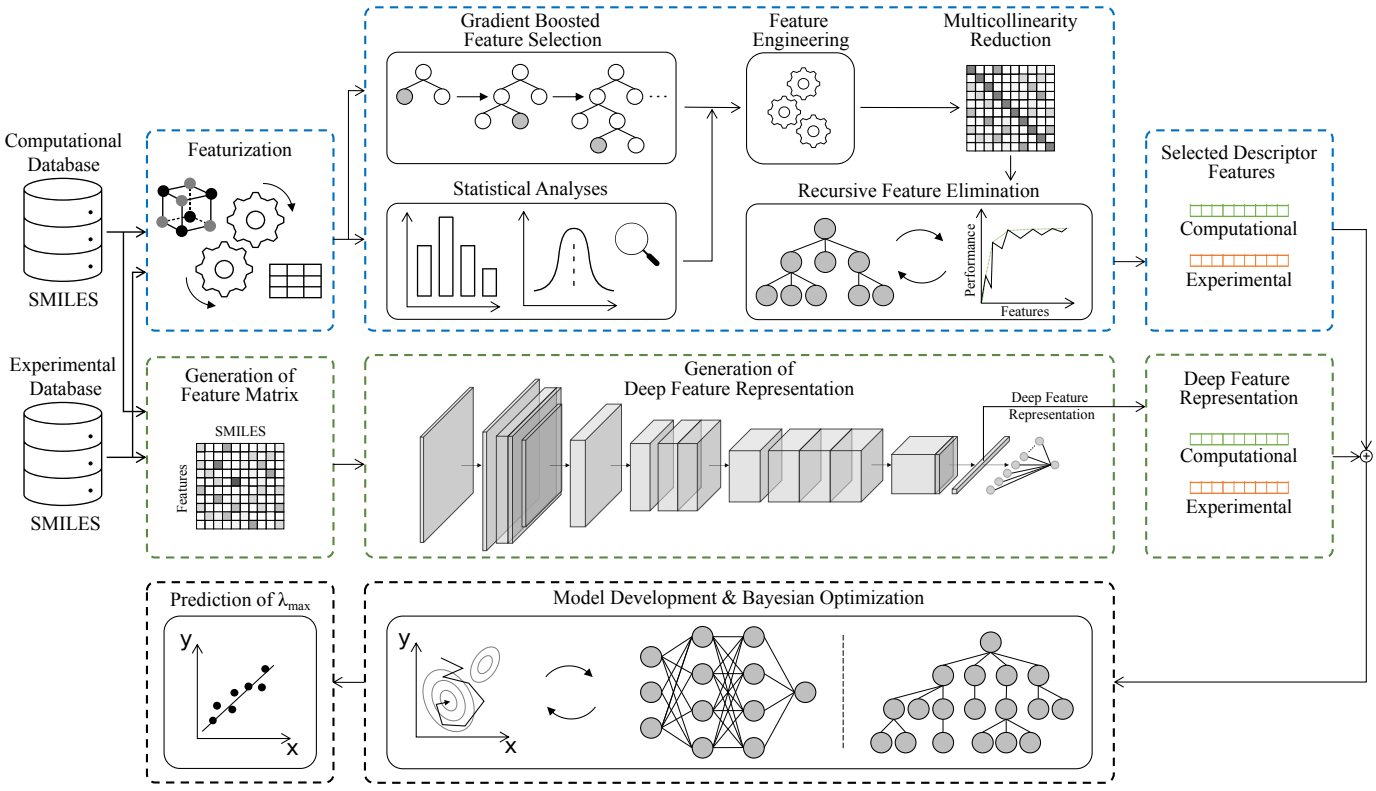


Fig. 5.2 Overview of our operational workflow compartmentalised into two distinct subworkflows. The subworkflow compartmentalisation in blue indicates the GBFS framework designed to identify a subset of features that affords minimal feature redundancy and maximal relevance to the target variable. The subworkflow encapsulated in green shows the deep learning method. This includes feature matrix generation using SMILES strings, followed by the generation of their DFR via DR-CNNs. The selected subset of descriptor features and DFRs are concatenated, and they serve as inputs to the final predictive ML model. The last stage is the development and optimisation of the final predictive ML model via Bayesian optimisation using Gaussian processes.

Data Sources, Acquisition, and Splitting

To adopt a multifidelity modelling approach, both experimental and computational data were acquired. The experimental UV–vis absorption data used herein were aggregated from five extensive open-source data tools, sources, or repositories. These included (i) ChemDataExtractor [266], (ii) Dye Aggregation [267], (iii) ChemFluor [268], (iv) Deep4Chem [269], and (v) the Dye-Sensitised Solar Cell Database (DSSCDB) [270]. These data sets were chosen due to their size and formatting style. Each chemical data entry contains a SMILES string of a dye molecule, accompanied by the peak optical absorption wavelength and solvent information. Not every solvent in these data records was presented in SMILES format. When such a SMILES string was unavailable, it was acquired from the data set compiled by Greenman et al., who identified the corresponding SMILES via the mapping between solvent information and a manually curated dictionary of SMILES strings. The data set was manipulated via the filtration of valid dye and solvent SMILES strings using RDKit [72], in addition to the elimination of entries with multiple dye molecules or molecular clusters. This resulted in a data set that comprised 28,734 sets of experimental measurements as summarised in Table 5.7; this is consistent with the data in ref [271]. Further data cleaning was undertaken to deal with duplicate entries and inconsistent wavelength measurements at a tolerance level of 5 nm. The ChemDataExtractor-generated data set was put aside from the combined experimental data set. This was done to achieve a like-to-like comparison with the state-of-the-art in the literature, namely, the work of Greenman et al., who did not use the ChemDataExtractor-generated data set. This data-removal process led to the remaining 26,395 records.

Table 5.7 List of Data Sources and the Corresponding Number of Experimental Measurement of Peak Optical Absorption Wavelengths[a].

Data tool or source	**Number of entries**
ChemDataExtractor	1 915
Dye Aggregation	3 626
ChemFluor	4 170
Deep4Chem	16 585
DSSCDB	2 438
total	28 734

[a]This table summarises the number of data entries after filtering; it does not reflect the entire data entries from each data tool or source and the duplication that exists between them.

The computational data used in this study were obtained from Beard et al. [266] and Greenman et al. [271] The data set provided by Beard et al. was realised via high-throughput electronic structure calculations that employed DFT within the simplified Tamm–Dancoff approximation (sTDA-DFT) and traditional TD-DFT. Chemicals were selected for these calculations where experimental data about their optical absorption wavelength had been mined from scientific literature using the "chemistry-aware" text-mining tool ChemDataExtractor [21, 266]. Thereby, the calculated data provided a theoretical counterpoint to the chemically matching experimental data. We used the entire set of 6,142 sTDA-DFT calculations for this study.

The computational data set provided by Greenman et al. included 28,772 geometrically optimised TD-DFT calculations. These calculations employed the Tamm–Dancoff approximation (TDA) [272] ωB97X-D3[273]/def2-SVPD level of theory using the ORCA software [274]. In preparation for such calculations, initial molecular geometries were generated by using RDKit [72] to convert SMILES strings into Cartesian coordinates. These initial geometries were refined employing semiempirical tight-binding DFT (GFN2-xTB[275]). These geometries were subsequently optimised at the BP86 [276]-D3 [277]/def2-SVP [278] level of theory. Solvent corrections were also employed using the integral equation formalism polarisable continuum model (IEFPCM) in the Gaussian software [279]. Cognate experimental data in solution exist for 19,409 or 10,409 of these molecular structures calculated under vacuum or solution, respectively. We note that 21 SMILES strings of dye molecules were incompatible with certain descriptors used in the GBFS workflow.

The proposed methodology was evaluated using two types of splitting strategies for the experimental measurements. This was to ensure that we conducted a fair assessment of the generalisability of the ML models since the choice of the splitting strategy can sometimes lead to an underestimation of the predicted errors [280, 281]. The strategies adopted were (i) random splitting by dye–solvent pairs and (ii) scaffold splitting using the Bemis–Murcko scaffold framework [247]. The former approach ensures that there is no overlap of dye–solvent pairings between the training and test sets given the removal of duplicated entries. Such a strategy, however, overlooks the potential correlation between measurements of the same dye in different solvents, which may result in an overestimation of predictive accuracy. Scaffold splitting addresses the potential overestimation of a model's performance. Thereby, a Bemis–Murcko scaffold was employed using a constraint that any dye molecules with an identical scaffold are present only in one of the data sets. This ensures that the realised errors are more reflective of the true model generalisability.

For both splitting strategies, the split ratio of 9:1 was used for the training and test sets, respectively. 20% of the training set was then retained as an out-of-sample validation set. Moreover, to ensure a fair and unbiased comparison with the state-of-the-art models reported in the literature, the results pertaining to the scaffold-splitting strategy presented hereafter are computed using the same data sets as those in the study by Greenman et al. [271].

Featurisation

There are two separate featurisation steps that run in parallel; the workflow for each is shown in Figure 5.2. For the featurisation step that involves a DR-CNN workflow, a 2D feature matrix or image is created from a SMILES string. The idea stems from image-processing methods that use convolutional neural networks, whereby high-dimensional, usually 2D or 3D, images are used to train and generate low-dimensional representations or feature vectors of the original images. These low-dimensional representations are subsequently fed into a fully connected network that has been optimised to perform either a classification or a regression analysis. The application of such a concept to computational chemistry means that low-dimensional representations of chemical molecules can be acquired from their corresponding feature matrices or images. Furthermore, these representations can be optimised to discriminate between target classes or to predict a target variable; this results in context-aware representations. This approach contrasts with fixed feature types that rely on standardised chemical or structural descriptors.

In this work, we create a feature matrix exclusively from SMILES strings. The SMILES representation is constructed from symbols that represent certain chemical and structural information about a molecule [39]. A SMILES string is broken up into a sequence of strings using a rule-based tokeniser, and the individual string (i.e., a symbol, element, or ion) forms the columns of the feature matrix, as shown in Figure 5.3. The resulting matrix is subsequently populated when certain features are detected, where each row represents one of the features listed in Table 5.8. The feature matrix is resized via a zero-padding operation. In order to incorporate solvent information, further zero-padding is applied, and the tokenised SMILES strings of the solvents in question are used to create additional columns in the feature matrix. The rows of this part of the matrix are populated in a manner similar to the part associated with the dye molecule.

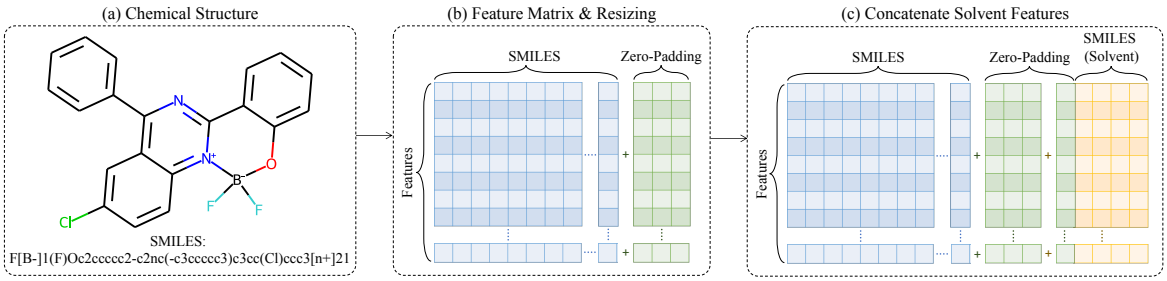


Fig. 5.3 Diagram depicting the process of feature matrix generation from a SMILES string of a chemical molecule. (a) SMILES string is broken up into a sequence of strings using a rule-based tokeniser. (b) Individual symbols, elements, and ions of the chemical molecule form the columns of the feature matrix. The rows of the matrix are subsequently populated when the features defined in Table 5.8 are detected. The matrix is resized to the required dimensions by applying zero-padding as illustrated. (c) To incorporate the information on a solvent, further zero-padding is applied, and the tokenised SMILES string of the solvent is used to create additional columns to the feature matrix. The rows are populated using the same process as described for (b).

Table 5.8 List of Features Used to Generate the Feature Matrix, as Shown in Fig. 5.3[a].

Feature Type	Feature Symbol or Name	Description	Length
SMILES	(	Start of branch	1
	)	End of branch	1
	[	Start of atom group	1
	]	End of atom group	1
	.	Ionic bond	1
	:	Aromatic bond	1
	=	Double bond	1
	#	Triple bond	1
	\	cis	1
	/	trans	1
	@	Chirality	1
	+	Positive charge	1
	-	Negative charge	1
	Ionic charge	No. of ionic charge (2-7)	6
	Ring	Start of ring	1
	Ring	End of ring	1
Chemical	Element	H, C, O, N, P, S or others	7
	No. of H	Total no. of hydrogen	1
	Degree	Degree of unsaturation	1
	Charge	Formal charge	1
	Valence	Total valence	1
	Ring	Within a ring	1
	Aromatic	Within an aromatic structure	1
	Tetrahedral Chirality	Clockwise, counter-clockwise or unrecognised	3
	Hybridisation	s, sp, sp2, sp3, sp3d, sp3d2 or unrecognised	7

[a]The resulting feature matrix is used as an input to the DR-CNNs in Figure 5.4. These features are computed using RDKit [72] following the tokenisation of the SMILES string.

Each feature matrix generated from a SMILES string contains 44 rows or features, as summarised in Table 5.8. The features are categorised into two types: (i) symbol-based SMILES and (ii) chemical-based information. The rows of the feature matrix are set up to detect 21 SMILES symbols and 23 chemical-based features, which include element types, number of hydrogen, degree of unsaturation, formal charge, total valence, ring, aromatic, chirality, and hybridisation. The four features listed in Table Table 5.8 (i.e., the number of ionic charges, type of element, chirality, and hybridisation) were treated as categorical features by representing them using one-hot encoding. These features are detected using RDKit [72], and additional features can be appended as a user option. It should also be noted that ambiguities can arise for compounds that can be represented in many different SMILES formats. These ambiguities can be eliminated during the preprocessing stage of the DR-CNN workflow, either manually or via the use of a normalisation algorithm, to ensure that unique SMILES representations for each chemical compound are supplied as input to the feature matrix generation stage.

Although the number of rows (features) is fixed, the number of columns (symbols) is determined by the longest SMILES string in the data set, by default. However, if solvent representations are disregarded, the user can predetermine the maximum number of columns, while SMILES strings shorter than the predetermined limit can be regulated using zero-padding. For studies where solvent representation is of importance, the limit can also be predetermined by considering the limit for the maximum SMILES length of the dye molecule and the solvent as well as the zero-padding between the two representations. In our algorithm, the default limiting number of columns in the feature matrix is set by the combination of the longest SMILES string of the dye molecule, the predetermined zero-padding layers, and the longest SMILES string of the solvent. Predetermining this limit so that the size of the feature matrix is known prior to carrying out any calculations is helpful because its dimensions can be used to optimise the computational cost associated with developing the DR-CNN. Within this scope, it can also minimise the number of zero-padding layers that can be applied.

Deep Residual CNN Architecture

Our DR-CNN architecture takes the 2D feature matrix described in Section 5.2.2 as input. Its architecture, inspired by deep residual learning [282], consists of five stages as shown in Figure 5.4a. The last four stages each include a convolutional block (ConvBlock) or a combination of a ConvBlock and an identity block (IdentBlock). The architecture of these two types of blocks is depicted in Figure 5.4b,c, along with operations performed and the number of filters used. In common with the popular

ResNet models, skip or residual connections are implemented with nonlinearities (e.g., ReLU) and batch normalisation.

The effect of network depth is highly influential in pattern or visual recognition tasks, with deep networks having the ability to learn complex patterns within a given data set. However, deep networks are prone to problems that are directly associated with their depth. In particular, the problem of the vanishing or exploding gradient needs to be addressed, as it hinders convergence [283, 284]. Consequently, batch normalisation is used to circumvent the convergence issue [285], although convergence with batch normalisation causes another complication, namely, the degradation problem. The degradation problem manifests itself with increasing network depth as accuracy degrades rapidly from the point of saturation; this is not a result of overfitting. Therefore, to benefit from the characteristics of a deep network while minimising the aforementioned problems, we employ the deep residual learning approach with a residual or skip connection. These residual connections ensure that we do not degrade the model when additional layers are introduced, which, in turn, increases the size of the searchable function space. This is equivalent to nesting the function space such that the model approximation does not move away from the current search space; thereby, guaranteeing that while the model can improve with more layers, it will not do any worse as regularisation will skip over them if the addition of layers does not prove to be useful.

The key difference between ConvBlock and IdentBlock is that there is no change in the input and output dimensions of the latter. In ConvBlock, there is a convolutional layer in the residual connection. This changes the dimensions such that the addition operator is applied to two matrices of equal size, and the output matrix has reduced dimensions compared to the input matrix.

It is important to emphasise that similar to the GBFS workflow, DR-CNN is a general-purpose workflow. This means that it is not designed specifically for the prediction of molecular absorption peaks; rather, its primary purpose is to generate deep representations of given molecules by incorporating a diverse set of features that are not exclusively tailored to a particular material property. These representations are subsequently fine-tuned against a user-defined target property or variable. Therefore, we do not constrain the selection of features to those directly pertinent to a specific material property as the workflow can be adapted for fine-tuning against other variables. In instances where certain features may be irrelevant to the target, the workflows are designed to handle this by eliminating such features (in GBFS) and either mitigating or disregarding their influence through regularisation techniques (in DR-CNN). Indeed,

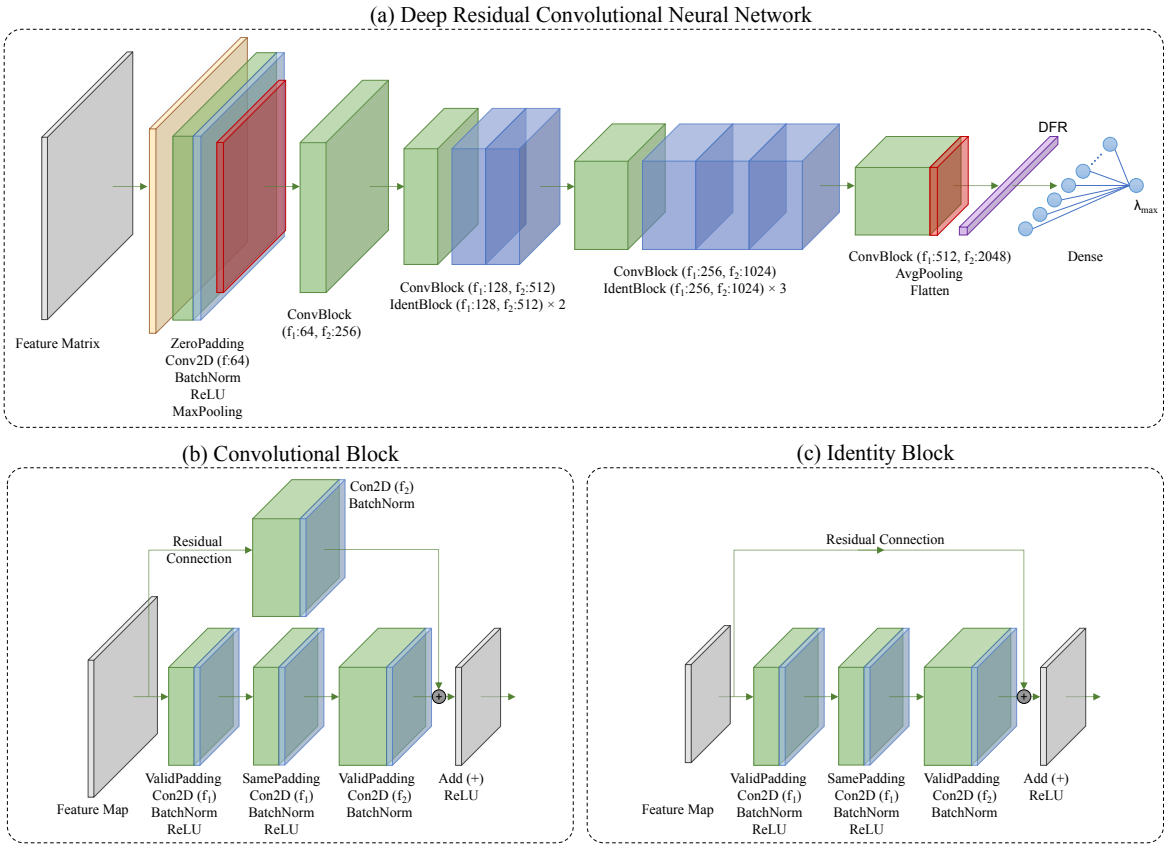


Fig. 5.4 (a) Schematic of the DR-CNN model architecture. As an input, the network takes in a feature matrix, as defined in Figure 5.3, whose features are listed in Table 5.8. The first convolutional layer consists of ZeroPadding, Conv2D with 64 filters (i.e., f:64), BatchNorm, ReLU, and MaxPooling operations, as denoted in the diagram. The subsequent stages of the DR-CNN model architecture consist of (i) convolutional block (ConvBlock) with f_1:64 and f_2:256; (ii) ConvBlock with f_1:128 and f_2:512 and two identity blocks (IdentBlock) each with f_1:128 and f_2:512; (iii) ConvBlock with f_1:256 and f_2:1024 and three IdentBlock each with f_1:256 and f_2:1024; and (iv) ConvBlock with f_1:512 and f_2:2048, AvgPooling, and Flatten operations, leading to a deep feature representation (DFR). A DFR feeds into a fully connected layer with a ReLu activation function and an output node that has a linear activation. (b,c) Architecture of the ConvBlock and IdentBlock, respectively, where f_1 and f_2 denotes two different filter sizes, as defined in (a).

the fact that our workflow automatically eliminates features that are known to be irrelevant to a given property of interest can help in the validation stage of a GBFS-based study; their selective removal demonstrates that our workflow can successfully discriminate features that should be discounted from a scientific standpoint.

5.2.3 Results and Discussion

Efficacy of Auxiliary ML Models: Predicting λ_{max} Values Against DFT Data

The prediction of vertical excitation energies and their corresponding peak wavelengths (λ_{max}) made by the Bayesian-optimised gradient-boosting models on the test set are shown in Figure 5.5. The inputs to these models were the DFT-aware features generated by either the DR-CNN or the GBFS subworkflows, whose inputs were the aforementioned feature matrices and descriptor features, respectively, generated using TD-DFT data described in Section 5.2.2. The DR-CNN and GBFS subworkflows achieved λ_{max} predictions with a mean absolute error (MAE) of 22.7 and 14.5 nm, a root-mean-square error (RMSE) of 34.9 and 23.1 nm, and a coefficient of determination (R^2) of 0.75 and 0.89 on the out-of-sample test set, respectively. The equivalent regression analyses of the DR-CNN and GBFS-based predictions of λ_{max} using the sTDA-DFT data set led to a test set MAE of 22.8 and 14.1 nm, an RMSE of 31.2 and 23.3 nm, and an R^2 of 0.71 and 0.88, respectively. The regression details associated with these model predictions can be found in Appendix B.1.

For this part of the analysis, we did not concatenate the outputs of the two subworkflows to develop a single predictive model for each of the DFT data sets. The rationale was to perform separate regression analyses to examine the difference in their performance levels. We observed that lower errors were realised when predictions of λ_{max} values were made using the GBFS workflow. However, certain descriptors failed to process some of the SMILES strings, while the generation of feature matrices for the DR-CNNs did not suffer from such an issue.

Moreover, we adopted a random splitting strategy to train the DFT-based ML predictions of λ_{max} values with a rationale that the development of the ML-based methods is decoupled from the process of DFT-based calculations. Indeed, the computational data were readily available prior to designing a framework to predict the experimental optical absorption peak wavelength of unseen chemical dyes.

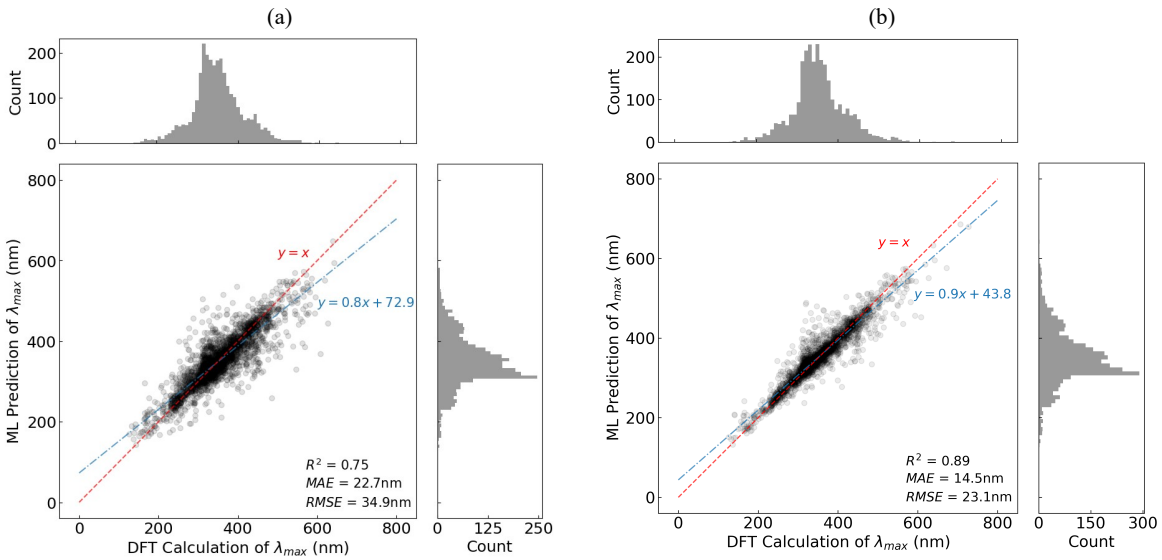


Fig. 5.5 Regression analysis of the ML-based prediction of vertical excitation energies, and their corresponding peak wavelengths ($\lambda_{\max}$) against the DFT-based calculations. The two separate predictions were made by two Bayesian-optimised gradient-boosting models trained on (a) DFR from the DR-CNN subworkflow and (b) descriptor features from the GBFS subworkflow. The analyses were conducted using the 28,751 vacuum TD-DFT calculations sourced from ref [266, 271] with a train-to-test split ratio of 9:1. The solid blue line is a linear fit between the DFT- and ML-based predictions, generated using the ordinary least squares. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the DFT-based calculations.

Efficacy of Multifidelity ML Models: Predicting λ_{max} Values Against DFT Calculations and Experimental Measurements

The combined data set of experimentally measured λ_{max} values was passed through the operational workflow of our system architecture, whereby the DR-CNN and GBFS models were trained to optimise a λ_{max} prediction model. By applying the test set of these data to our method with the random splitting strategy, our method realised an MAE of 16.5 and 23.3 nm, an RMSE of 33.1 and 44.4 nm, and an R^2 of 0.90 and 0.82, using the GBFS and DR-CNN subworkflows, respectively. When the scaffold-splitting strategy was used to create the test set, our method achieved an MAE of 28.2 and 35.9 nm, an RMSE of 41.2 and 49.5 nm, and an R^2 of 0.84 and 0.77, through the GBFS and DR-CNN subworkflows, respectively. The realised errors and the coefficient of determination for each subworkflow and splitting strategy are summarised in Table 5.9. The corresponding regression plots can be found in Appendix B.2.

These results reflect the effectiveness of the experimentally aware DFR and feature descriptors in the prediction of the optical absorption peak. We subsequently concatenated the two representations along with those generated by the auxiliary models, to afford the input to the final gradient-boosting model that was refined by Bayesian optimisation. Our multifidelity ML model realised an MAE, an RMSE, and an R^2 of 14.6, 31.3 nm, and 0.91, respectively, when applied to a test set that was generated using the random splitting approach, while a test set generated using the scaffold-splitting strategy afforded an MAE, an RMSE, and an R^2 of 24.7, 39.2 nm, and 0.86, respectively. These results are summarised in Table 5.9, while the regression analysis is shown in Figure 5.6 along with the corresponding error distributions in Figure 5.7.

We evaluated the performance of our proposed method against state-of-the-art ML models that have been reported in the literature. To ensure a consistent comparative analysis and assess the generalisability of our method and other approaches, we prioritised model evaluation using the scaffold-splitting strategy where possible, and we highlight the data sets that were used. Ju et al. realised a test set from the ChemFluor data set, whose MAE was 10.46 nm when applying GBRTs with the random splitting strategy [268]. Greenman et al. verified this result by illustrating that the GBRTs can indeed achieve an MAE of approximately 10 nm on a test set of the Deep4Chem data set with random splitting, before further demonstrating that this MAE can increase to 27 nm when the scaffold-splitting strategy is used [271]. Additionally, Greenman et al. showed that the open-source Chemprop D-MPNN models with the multifidelity approach, coined ChempropMultiFidelity, can realise lower MAE values

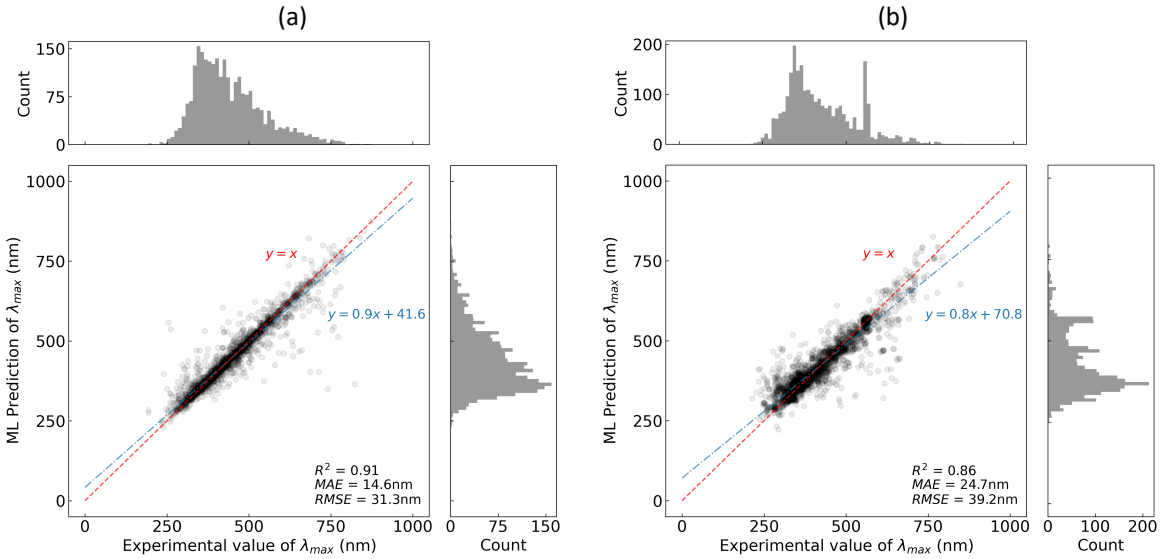


Fig. 5.6 Regression analysis of the multifidelity ML model prediction of the optical absorption peak wavelength ($\lambda_{\max}$) values of unseen chemical dyes against their experimental measurements. The two separate predictions were made by two Bayesian-optimised gradient-boosting models with a training set that was afforded using (a) random and (b) scaffold-splitting strategies, respectively. The solid blue line is a linear fit between the experimental measurements and ML-based predictions, generated using ordinary least-squares refinement. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the experimental measurements.

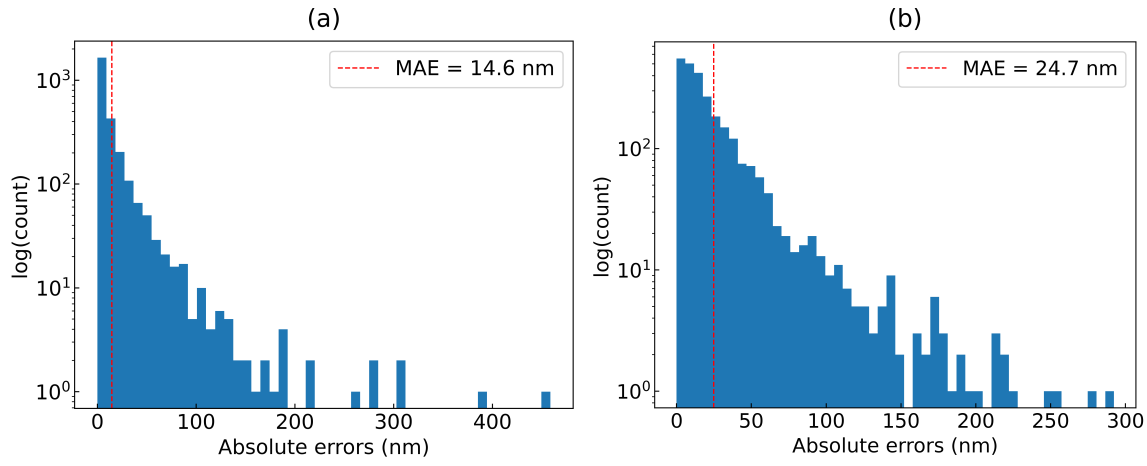


Fig. 5.7 Distributions of the absolute errors of the ML-based prediction of the optical absorption peak wavelength ($\lambda_{\max}$) against the experimental measurements. (a) Error distribution obtained using the random splitting strategy and (b) error distribution obtained using the scaffold splitting approach. The dashed red line indicates the MAE.

Table 5.9 Summary of the Model Performance categorised by the Choice of Workflow, Data Type, Splitting Strategy, and the Statistical figures of Merit, which Includes MAE, MSE, and the Coefficient of Determinant (R^2).

Data	Workflow	Split Strategy	MAE (nm)	RMSE (nm)	R^2
sTDA-DFT	GBFS	Random	14.1	23.3	0.88
	DRCNN	Random	22.8	31.2	0.71
TD-DFT	GBFS	Random	14.5	23.1	0.89
	DRCNN	Random	22.7	34.9	0.75
Experimental	GBFS	Random	16.2	32.2	0.91
	DRCNN	Random	23.3	44.4	0.82
	GBFS	Scaffold	27.0	39.3	0.85
	DRCNN	Scaffold	35.9	49.5	0.77
	Multi-Fidelity	Random	14.6	31.3	0.91
		Scaffold	24.7	39.2	0.86

Table 5.10 List of Features Identified to Have the Most Relevance in the Prediction of λ_{max} Values against Their Experimental Measurements of Optical Absorption Peak Wavelength[a].

No.	Feature Abbreviation	Feature Description
1	DR-CNN TD-DFT DFR-345	DR-CNN DFR of TD-DFT (index 345)
2	DR-CNN TD-DFT Peak	DR-CNN-based TD-DFT λ_{max}
3	GBFS sTDA-DFT Peak	GBFS-based sTDA-DFT λ_{max}
4	MACCSkeys-49	MACCS keys 49 (i.e. C=C(C)C)
5	NumAliphaticRings	No. of aliphatic rings
6	MaxAbsEStateIndex	Maximum absolute E-state index
7	MorganFeature3Counts	Count-based Morgan fingerprint features of radius 3 (index 31)
8	RDKitFPBits-1285	RDKit fingerprint bits (index 1285)
9	Fr_Allylic_Oxid	No. of allylic oxidation sites (excl. steroid dienone)
10	MaxEStateIndex	Maximum E-state index
11	SlogP_VSA1	Log of the octanol-water partition coefficient & Van-der-Walls surface area approximation (index 1)
12	LabuteASA	Labute accessible surface area value
13	BCUT2D_MWLOW	2D Burden eigenvalue (low mass eigenvalue)
14	PEOE_VSA8	Exposed functionality for MOE-like approximate molecular surface area descriptors (index 8)
15	VSA_EState2	Hybrid E-state-Van-der-Walls surface area approximation descriptor (index 2)

[a]See Figure 5.8 for the corresponding total loss reduction realised by the features.

on all split types compared to GBRTs on test sets derived from the Deep4Chem data set. Furthermore, Joung et al. realised an RMSE of 31.6 nm using GCNNs on a test set of the Deep4Chem data set with a random splitting strategy [269, 286]. Greenman et al. demonstrated that ChempropMultiFidelity can improve the realised RMSE to 27.47 nm even with scaffold splitting using such a data set. These results suggest that ChempropMultiFidelity has a higher degree of generalisability and predictability when compared to both GBRT and GCNN modelling approaches, as adopted by Ju et al. and Joung et al. This makes ChempropMultiFidelity a suitable benchmark for our proposed method.

As previously mentioned, we trained our DR-CNN and GBFS models and computed the corresponding performance metrics (Table 5.9) using data sets identical to those used by Greenman et al., where the ChemDataExtractor-generated data set was excluded from the combined experimental data set (Table 5.7), and the scaffold-splitting strategy was used to create the training and test sets. This ensured a rigorous evaluation of our method and enabled a like-for-like comparison. With 5-fold cross-validation, Greenman et al. achieved an MAE of 27.78 $\pm$ 5.07 nm, an RMSE of 47.13 $\pm$ 11.10 nm, and an R^2 of 0.8 $\pm$ 0.07 using the ChempropMultiFidelity model on the test set [271], while our proposed multifidelity method realised errors in energy which, once converted into wavelength, afforded an MAE of 24.91 $\pm$ 0.64 nm, an RMSE of 39.33 $\pm$ 0.85 nm, and an R^2 of 0.85 $\pm$ 0.01 using the same test set. Our results are therefore comparable to those obtained by Greenman et al. when accounting for their uncertainties, whose method is reported to outperform both the GBRTs and GCNNs. In terms of average performance, however, our modelling approach demonstrates superiority, achieving a lower mean across the three performance metrics. Our proposed method exhibits significantly lower standard errors across these metrics, indicating a more stable model with reduced variance in the predictions. For instance, considering one standard deviation, approximately 68% of the predictions are expected to fall within the defined uncertainty range. While the performance of both methods is statistically comparable, our advantage of having a more stable prediction is evident.

Moreover, the use of scaffold splitting meant that the ML models were less reliant on the combination of dye molecules resided in the training set and they were encouraged to incorporate learning from the effects of chemical composition, molecular structure, and solvent. These results demonstrate the generalisability of our proposed method in predicting the molecular absorption maxima in the optical spectrum. It is important to note that the combined experimental data sets widened the chemical landscape compared to the original, individual data sets by introducing a different chemical space

that was represented by each data set. Consequently, we expect the magnitude of the errors in these predictions to be greater than those reported in the literature, which utilises a subset of the combined experimental data set in Table 5.7.

We now take a closer examination of the performance of our model on dyes in which metals are present. In total, there are 308 dye molecules that contain metals, 281 in the training set and 27 in the test set with scaffold splitting. Further categorisation shows that there are 267 post-transition metals, 39 transition metals, and 2 alkali metals. The analysis of the absolute errors in energy for the ML-based predictions of λ_{max} against the experimental measurements on these metal-containing dyes, once converted into wavelength, yielded an MAE of 58.0 nm. The minimum and maximum absolute errors of 4.3 and 167.5 nm were observed, respectively. Upon segregating the error analysis by the types of metals, we observed an MAE of 61.7 nm for dyes containing post-transition metals, while an MAE of 11.2 nm was observed for those containing transition metals. At a high level, the model appears to encounter difficulties in accurately predicting λ_{max} for dye molecules in the presence of metals, with larger errors being predominant in those containing post-transition metals. This stands to reason by using chemical intuition. However, it is crucial to emphasise that this observation does not constitute a definitive conclusion, given that the subset of metal-containing dyes constitutes a minority within the data set, comprising 281 out of 23,754 in the training set. We anticipate an improvement in the model accuracy with an increasing number of metal-containing dye molecules.

We also examined the performance of our model by considering solvent effects. Thereby, the distribution of absolute errors in the ML-based predictions of λ_{max} against the experimental measurements was partitioned based on the eight most frequently occurring solvents in the training set using scaffold splitting (for details, see Appendix B.3). The MAE ranges from the lowest value of 15.1 nm (for dimethyl sulfoxide, i.e., SMILES string, CS(C)=O) and the highest value of 32.8 nm (for ethanol, i.e., SMILES string, CCO). This does not imply that there is a more pronounced adverse impact on the predictive model for chemicals solvated in ethanol compared with dimethyl sulfoxide. Assessing the effect of these solvents on the prediction is challenging due to substantial variations in the number of observations among solvents and the absence of dye molecules solvated in a variety of solvents within the test set. A comprehensive exploration of the effects of solvents, encompassing their role as dielectric media or their role in more intricate phenomena such as the modification of a conformational ensemble, constitutes a multifaceted and nuanced subject. This topic is suggested for further exploration as a potential avenue for subsequent research.

We attributed the generalisability demonstrated by the proposed method to three components of our operational workflow. First, in contrast to a naive approach to utilising feature descriptors, our GBFS workflow design identifies the subset of features that maximises the relevance to the target variable while minimising feature redundancy. This allows the exploration of a wide range of descriptors, where only a subset of features may be relevant to the prediction of the optical absorption peaks; one can avoid training an ML model with a convoluted list of features, thereby preventing the model from overfitting by achieving a better bias-variance trade-off. In other words, our GBFS workflow is able to capture the relationship between the exploratory features and the target variable while controlling the sensitivity of our method to small fluctuations, or noises, within the data set.

Second, the multifidelity modelling approach, where auxiliary ML models (both DR-CNN and GBFS) are trained on DFT data sets, appears to have a substantial impact on the prediction of λ_{max} values against their experimental measurements. For example, the DFR of the dyes and solvents learned via the DR-CNN workflow appears to have made the most significant contribution in the model prediction. Figure 5.8 shows that the largest total loss reduction is realised by the λ_{max} values that have been predicted using the DFR from the DR-CNN workflow, which was afforded by its training against the TD-DFT peak vertical excitation energy using random splitting. Meanwhile, the DFR itself that was learned from the TD-DFT calculations ranked first when scaffold splitting was used. Moreover, the second largest total loss reduction was realised by the prediction of λ_{max} values via the GBFS workflow, which was afforded by its training against sTDA-DFT absorption maxima [266]. Both examples involve either the prediction or a learned representation of the DFT data set via one of the subworkflows. This demonstrates the effectiveness of the multifidelity modelling approach. These results suggest that the inclusion of readily available DFT data sets from the literature appears to help the ML models navigate the chemical space in search of experimental values of the optical absorption peaks. Alternatively, one can consider the multifidelity modelling method as creating a nested chemical space such that the search space is contained within a wider landscape defined by the computational data, which cannot be explored solely by relying on limited number of experimental data.

Third, we highlight the important contribution that was made by the DR-CNN workflow in the final prediction of λ_{max} values against their experimental measurements. DR-CNN is able to process and learn from the feature matrices through its deep architecture and identify complex relationships between the features derived exclusively from SMILES strings and λ_{max} values. It should also be noted that the ML models

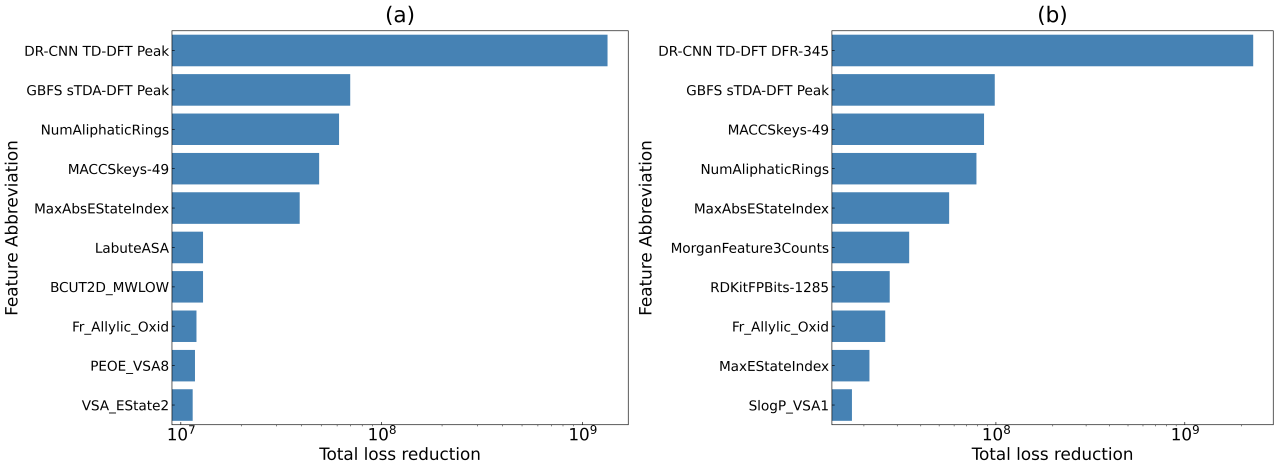


Fig. 5.8 Bar chart depicting the total loss reduction realised by the top 10 features identified to have the most relevance to the experimental measurements of the optical absorption peak wavelength, using the (a) random and (b) scaffold-splitting strategies. The corresponding feature descriptions can be found in Table 5.10.

trained on the DFT data sets can be leveraged to make separate predictions, as opposed to carrying out more costly DFT calculations for novel chemical molecules.

We note that our proposed method does not directly compare between vertical transition energies, which are calculated using quantum chemical methods, and experimental absorption maxima. It is well established that these two quantities differ in their definitions. Large discrepancies can arise due to a shift between equilibrium positions of potential energy surfaces, for instance, when a harmonic model predicts the vertical transition energy as the centroid of a convolution of a Poisson distribution in the absence of Duschinsky rotation [287–289]. Furthermore, spectral shapes are determined by the vibrational envelope and can experience broadening effects, among other factors, making direct comparisons or benchmarking between the two quantities not necessarily valid. Our objective is to explore whether there is additional information or underpinning relationships that can be captured by considering both pieces of information. The results indicate a considerable information gain (or reduction in loss) when vertical transition energies are taken into account by the predictive model, as illustrated in Figure 5.8. This suggests that despite the known disparities or deviations between the two quantities, the ML algorithm can capture valuable information or relationships between them, resulting in lower loss during model training.

Furthermore, it is necessary to be conscious of conformational effects in the context of spectroscopy and molecular chemistry [76]. These effects manifest as variations in the atomic arrangement within a molecule, stemming from different spatial orientations

or conformations. While spectroscopic data can offer insights into these conformational effects by representing a Boltzmann average of all conformers present in a solution, SMILES strings lack the capacity to consider such variations. Given that this study does not distinguish between cases involving single or multiple conformations in the experimental measurements, nor does it address potential modifications to the conformational ensemble due to interactions with the environment (i.e., the solvent), it is imperative to consider these factors when citing or referencing the methodology and the findings presented herein. This is pertinent not only to this study but also to other research endeavours employing SMILES representation.

Application of ML Models Across Chemical Space

With promising statistical figures of merit on both accounts, the results in Table 5.9 present predictions of λ_{max} values against experimental measurements and DFT calculations. Nonetheless, it is important to validate these results by considering how these predictions fare across the diverse range of optically active chemicals rather than simply demonstrating their collective statistical quality in an anonymised form.

Indeed, there are well-known relationships between optical absorption wavelengths and the molecular structure of organic compound [290, 291]. Thus, it would be natural to test how well these predictions of λ_{max} values fare against illustrative types of organic chemicals. For example, one might consider whether predictions are more or less successful when a molecule has fused rings or not or if it contains many rings or rings that feature certain heteroatoms. Does the quality of the predictions vary as a function of the number of π-conjugated bonds in the molecule, the overall size of the molecule, or the length of the π-bridge in a donor–π–acceptor molecule?

Such chemical classifications for optically active molecules on the ChemDataExtractor-generated data set of experimentally measured λ_{max} values have been made by Flanagan and Cole [291]. They assembled a list encompassing some of the most common chemical groups identified in optically absorbing organic dyes across multiple domains. The presence or absence of a functional group in a molecule is ascertained through SMARTS (SMILES Arbitrary Target Specification) pattern matching. Subsequently, a hierarchical fingerprint scheme is used to categorise similar functional molecular fragments along with illustrative examples of molecules spanning this diversity of chemical space. Thus, it seemed quite natural to apply our multifidelity ML model to the SMILES strings of the molecules that they illustrated. This analysis enabled us to test the accuracy of our λ_{max} predictions across a rich set of optically active chemicals.

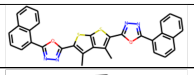
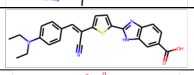
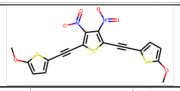
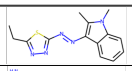
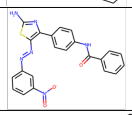
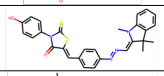
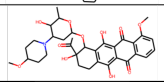
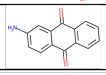
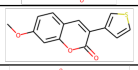
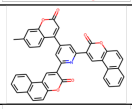
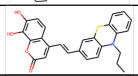
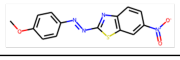
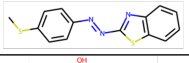
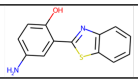
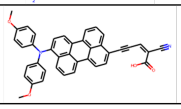
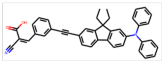
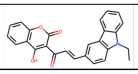
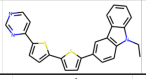
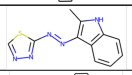
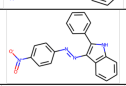
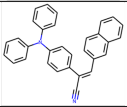
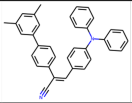
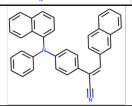
	Prediction Input			Prediction Output	
Group	Molecule || Solvent	Drawing of 2D Molecules	Expt λ_{max} (eV)	Pred λ_{max} (eV)	Pred - Expt $\|\Delta\lambda_{max}\|$ (meV)
Thiophene	Cc1c(sc2c1c(C)c(s2)c1nnc(o1)c1cccc2c1cccc2)c1nnc(o1)c1cccc2c1cccc2 ||Ethanol OCC||		3.434	3.422	12 (~1 nm)
Thiophene	CCN(c1ccc(cc1)C=C(c1ccc(s1)c1nc2c([nH]1)cc(cc2)C(=O)O)C#N)CC ||Dimethylformamide CN(C)C=O||		2.695	2.682	13 (~2 nm)
Thiophene	COc1ccc(s1)C#Cc1sc(c(c1[N+](=O)[O-])[N+](=O)[O-])C#Cc1ccc(s1)OC ||Dichloromethane ClCCl||		2.695	2.722	27 (~5 nm)
Azo	CCc1nnc(s1)N=Nc1c2ccccc2n(c1C)C ||Dimethyl sulfoxide CS(=O)C||		2.805	2.811	6 (~1 nm)
Azo	Nc1sc(c(n1)c1ccc(cc1)NC(=O)c1ccccc1)N=Nc1cccc(c1)[N+](=O)[O-] ||Dimethyl sulfoxide CS(=O)C||		2.515	2.542	27 (~5 nm)
Azo	Oc1ccc(cc1)N1C(=S)SC(=Cc2ccc(cc2)N=NC=C2N(C)c3c(C2(C)C)cccc3)C1=O ||Tetrahydrofuran C1CCOC1||		2.362	2.335	27 (~6 nm)
Anthraquinone	COC1CCN(CC1)C1CC(OC(C1O)C)OC1c2c(CCC1(O)C(=O)C)c(O)c1c(c2O)C(=O)c2c(C1=O)cccc2OC ||1,4-Dioxane O1CCOCC1||		2.931	2.869	62 (~9 nm)
Anthraquinone	Nc1ccc2c(c1)C(=O)c1c(C2=O)cccc1 ||Methanol CO||		2.644	2.841	197 (~33 nm)
Coumarin	COc1ccc2c(c1)oc(=O)c(c2)c1cscc1 ||Chloroform ClC(Cl)Cl||		3.333	3.345	12 (~1 nm)
Coumarin	Cc1ccc2c(c1)oc(=O)cc2c1cc(nc(c1)c1cc2c(oc1=O)ccc1c2cccc1)c1cc2c(oc1=O)ccc1c2cccc1 ||Chlorobenzene c1ccc(cc1)Cl||		3.220	3.206	14 (~2 nm)
Coumarin	CCCN1c2ccc(cc2Sc2c1cccc2)C=Cc1cc(=O)oc2c1ccc(c2O)O ||Chloroform ClC(Cl)Cl||		2.904	2.869	35 (~5 nm)
Benzothiazole	COc1ccc(cc1)N=Nc1nc2c(s1)cc(cc2)[N+](=O)[O-] ||Methanol CO||		3.046	3.024	22 (~3 nm)
Benzothiazole	CSc1ccc(cc1)N=Nc1nc2c(s1)cccc2 ||Methanol CO||		3.077	3.054	22 (~3 nm)
Benzothiazole	Nc1ccc(c(c1)c1nc2c(s1)cccc2)O ||Benzene c1ccccc1||		3.100	3.055	45 (~6 nm)
Cyanoacrylate	COc1ccc(cc1)N(c1ccc2c3c1cccc3c1c3c2cccc3c(cc1)C#CC=C(C(=O)O)C#N)c1ccc(cc1)OC ||Toluene Cc1ccccc1||		2.515	2.531	16 (~3 nm)
Cyanoacrylate	N#CC(=Cc1cccc(c1)C#Cc1ccc2c(c1)C(CC)(CC)c1c2ccc(c1)N(c1ccccc1)c1ccccc1)C(=O)O ||Tetrahydrofuran C1CCOC1||		2.965	2.872	93 (~14 nm)
Carbazole	CCn1c2ccccc2c2c1ccc(c2)C=CC(=O)c1c(=O)oc2c(c1O)cccc2 ||Methanol CO||		2.655	2.722	67 (~12 nm)
Carbazole	CCn1c2ccccc2c2c1ccc(c2)c1ccc(s1)c1ccc(s1)c1ccncn1 ||Acetonitrile CC#N||		2.988	3.058	70 (~10 nm)
Indole	Cc1[nH]c2c(c1N=Nc1nncs1)cccc2 ||Dimethyl sulfoxide CS(=O)C||		2.644	2.692	49 (~8 nm)
Indole	[O-][N+](=O)c1ccc(cc1)N=Nc1c([nH]c2c1cccc2)c1ccccc1 ||Methanol CO||		2.786	2.694	92 (~15 nm)
Triarylamine	N#CC(=Cc1ccc2c(c1)cccc2)c1ccc(cc1)N(c1ccccc1)c1ccccc1 ||Tetrahydrofuran C1CCOC1||		3.009	2.998	11 (~2 nm)
Triarylamine	N#CC(=Cc1ccc(cc1)N(c1ccccc1)c1ccccc1)c1ccc(cc1)c1cc(C)cc(c1)C ||Cyclohexane C1CCCCC1||		3.002	2.980	22 (~3 nm)
Triarylamine	N#CC(=Cc1ccc2c(c1)cccc2)c1ccc(cc1)N(c1cccc2c1cccc2)c1ccccc1 ||Tetrahydrofuran C1CCOC1||		3.069	3.044	25 (~3 nm)

Fig. 5.9 Examples of input molecules and the corresponding prediction of $\lambda_{\max}$ values against their experimental measurements. The input to the prediction is exclusively the SMILES string of the subject molecule unless the user wishes to include solvent information for which its SMILES string will then be given. The output is the predicted $\lambda_{\max}$ value of the subject compound. We considered unseen chemical molecules from a list of nine chemical classes commonly found in optically absorbing organic molecules. The absolute difference between the predicted values and the experimental measurements ($|\Delta\lambda_{max}|$) are colour-coded in green ($0 - 45\,\text{meV}$), amber ($46 - 90\,\text{meV}$) or red ($\geq 91\,\text{meV}$).

Among the list of chemical classes that they examined, we chose unseen candidates with associated solvent information for analysis using our proposed methodology. The results are shown in Figure 5.9, where the analysis considered nine classes commonly found in optically absorbing organic molecules. The absolute difference of each predicted $\lambda_{\max}$ value from its experimental measurement is colour-coded according to the classifications green ($0 - 45\,\text{meV}$), amber ($46 - 90\,\text{meV}$) or red ($\geq 91\,\text{meV}$).

We observe relatively higher discrepancies in three classes of chemicals listed in Figure 5.9, namely, anthraquinones, carbazoles, and indolines. Poor predictions are noted for both case studies of anthraquinone molecules. The smaller one of these molecules exhibits pseudosymmetry, i.e., it would have D_{2h} point group symmetry if its NH_2 group was absent; this will compromise any "push–pull" π-conjugated character typically associated with optically absorbing molecules. The larger anthraquinone molecule displays a minimal level of π-conjugation beyond its parent anthraquinoidal moiety; indeed, its other moiety is predominantly saturated, which will preclude intramolecular charge transfer (ICT) and thus deter the optical absorption properties of the overarching molecule. In both cases, these π-conjugated perturbations deviate from the typical π-conjugated characteristics observed in the other molecules shown in Figure 5.9; such perturbations hinder the prediction accuracy of the model for these molecules.

Another type of π-conjugated perturbation may also be responsible for the poor prediction accuracy of the second cyanoacrylate listed in Figure 5.9. The π-conjugated backbone of this molecule contains an alkynyl group; while this features π-bonding, the presence of two mutually perpendicular p-orbitals within its triple bond will complicate its ICT characteristics and, thus, the ability to predict the optically absorbing properties of this molecule. While there exist two alkynyl groups in one of the other molecules in Figure 5.9 (the third thiophene in the list), they do not lie on the only π-conjugated

backbone of the molecule in contrast to this cyanoacrylate molecule, so their impact on the predicted optical absorption properties will likely be of less significance.

The poor predictive performance of the indole- and carbazole-based molecules appears to be systematic. On the one hand, this chemical trend is not surprising since carbazoles are based on the indole structure: the second benzene ring of the carbazole has typically been created through chemical substitution at the 2,3 positions of an indole structure. On the other hand, the rationale for large discrepancies between predicted and experimental values of λ_{max} for these molecules is less clear. For example, the molecular structures of the first azo and first indole chemicals listed in Figure 5.9 are almost identical. Yet, the prediction for this azo compound matches well with its optical absorption from the experiment, while those of the indole do not. Both compounds have been solvated in the same solvent, dimethyl sulfoxide (DMSO). One key chemical difference between these two compounds is that the nitrogen atom on the one classed as an indole exhibits a terminal hydrogen atom, while the hydrogen atom of the N on the indole ring of its comparative azo molecule has already been substituted by a methyl group. The indole will be inclined to deprotonate when solvated in DMSO, and the predictions will not capture this solvent effect or model well this tendency toward protonation, while the experimental λ_{max} value will implicitly capture this effect. The N atom of the ring in the other indole listed in Figure 5.9 also carries a terminal hydrogen, although its experimental λ_{max} value was captured by using methanol as the solvent. The rationale behind the significant λ_{max} differences between the prediction and experiment for the two carbazoles listed in Figure 5.9 is less clear.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis presents the formulation and deployment of a gradient-boosted statistical feature selection (GBFS) workflow specifically designed for materials informatics. It elucidates the efficacy of this methodology through its application in predicting diverse material properties within several key research domains, such as solid-state physics, condensed matter physics, materials science and engineering, and spectroscopic characterisation. The overarching objective of this work is to develop a comprehensive, universally applicable materials discovery workflow that significantly reduces computational demands and diminishes potential human biases during the model development phase. This general-purpose workflow efficiently processes both organic and inorganic materials, thus providing a robust and adaptable tool suitable for a wide array of material inputs.

Chapter 1 reviews the evolution of the materials discovery process and underscores the critical influence of machine learning on these developments. It highlights existing limitations in current predictive methodologies within the field of materials informatics. The chapter elaborates on fundamental machine learning principles pertinent to the discussions in subsequent chapters. Particular emphasis is given to the role of regularisation techniques that are often used during the model training phase, as they offer feature selection capabilities and improve model generalisation. This section emphasises the necessity for a sophisticated computational modelling approach capable of efficiently processing complex, multidimensional data sets without incurring substantial computational costs. Moreover, the potential of gradient-boosting algorithms as a powerful tool in statistical learning is examined. This analysis sets the stage for

their strategic integration into materials informatics, illustrating how they can enhance predictive accuracy and efficiency in materials science research.

Chapter 2 details the theoretical foundation and practical construction of the GBFS workflow. It includes a comprehensive discussion on the choice of algorithms, the architecture of the workflow, and the rationale behind its design. This chapter also delves into how gradient boosting is specifically adapted to enhance feature selection, thereby improving the accuracy and efficiency of materials property predictions. This chapter sets the foundational concepts and techniques that will be explored and implemented throughout the thesis, establishing a thorough framework for the integration of machine learning in materials discovery.

The developed workflow is applied across various research domains, demonstrating its effectiveness in accurately predicting material properties when compared to high-throughput calculations and experimental measurements. This capability forms a central theme that is explored and elaborated upon.

Chapter 3 focuses on inorganic chemistry, showcasing the application of the GBFS workflow for predicting Curie temperature from the chemical compositions of ferromagnetic materials and to automate the construction of magnetic phase diagrams. The chapter goes on to explore the capabilities of these machine learning models beyond conventional benchmarking against state-of-the-art methodologies. Specifically, it includes predictions of Curie temperature for rare-earth intermetallic compounds and the generation of magnetic phase diagrams for compounds within binary systems. These case studies demonstrate the models' ability to make accurate out-of-distribution predictions that extend beyond the initial distribution of the training set. This process involves extrapolating relationships previously identified within a comprehensive materials database. The findings underscore the potential of machine learning models to provide insights into unexplored domains of materials informatics, highlighting their utility in predicting properties and behaviors of materials under conditions not previously documented within the existing data frameworks.

We extend the application of our GBFS workflow to additional research domains. For instance, in **Chapter 3**, the multifidelity modelling approach is employed to explore the prediction of band gaps in inorganic materials. This chapter details the development of two distinct models: a classification model that differentiates between conductors and semiconductors, and a regression model designed to estimate band gap values based solely on chemical compositions. The integration of high-throughput calculations with experimental measurements plays a crucial role in this study, notably reducing the number of input features necessary for maintaining robust model performance.

This reduction is a key advantage of the multifidelity modelling approach, as it allows for the extrapolation of learned patterns from high-throughput calculations to predict band gaps against their experimental counterparts. The chapter also discusses the challenges associated with this approach, particularly the inherent errors that can limit model accuracy. These limitations are analysed, providing a balanced view of the benefits and potential pitfalls of using multifidelity modelling techniques in the prediction of material properties.

Chapter 4 presents the results of the predictive modelling for high-entropy alloys (HEAs) and amorphous metallic alloys (AMAs). The methodology efficiently isolated relevant features from complex data sets, primarily based on the chemical compositions of materials, to train predictive models with minimal redundancy and maximal relevance. The effectiveness of this approach was validated through two case studies: a regression analysis on the bulk modulus of HEAs and a classification analysis on the glass-forming ability of AMAs. These results outperform many existing studies, highlighting the benefits of precise feature selection and analysis over mere reliance on algorithmic complexity. This approach has proven successful in predicting a broad range of properties for alloys, leveraging chemical data from diverse literature sources.

We conclude **Chapter 4** with a focus on predicting the critical temperature of superconductors based solely on their chemical composition. Enhanced by Bayesian optimisation to fine-tune gradient-boosting trees, this approach demonstrated robust predictive capabilities. A detailed analysis of approximately 16,400 chemical compounds facilitated the development of a classification model capable of identifying chemical compositions likely to exhibit a critical temperature above 10 K. In parallel, a regression model was constructed to predict critical temperature values, utilising features exclusively derived from the chemical composition of the materials. This dual-model approach underscores the potent predictive power inherent in targeted machine learning methodologies applied to materials science. Extending beyond standard benchmarks, our machine learning model precisely predicted the critical temperature for 44 out of 50 compounds, aligning closely with experimental measurements. Moreover, the model performed out-of-distribution predictions on $(Ca_{1-x}La_x)FeAs_2$ with varying lanthanum content, affirming the strong extrapolative capabilities of the model. These results highlight the importance of meticulous feature selection in enhancing model accuracy, demonstrating a clear advantage over models that primarily depend on algorithmic complexity.

Chapter 5 underscores the versatility and applicability of our GBFS workflow within the domain of organic chemistry while showcasing its seamless integration with

other machine-learning frameworks. The chapter begins by applying the workflow to over 20 complex predictive tasks, including the prediction of quantum mechanical properties and drug-related characteristics, using SMILES strings as the sole input representation. Substructure vector embeddings are incorporated into a semi-supervised learning framework to enhance predictive accuracy. This approach highlights the capacity of our workflow to leverage advanced techniques, solidifying its efficacy in addressing a wide range of intricate challenges in organic chemistry.

Chapter 5 then transitions to a study demonstrating the integration of the proposed workflow with convolutional neural networks for predicting optical absorption peaks in organic materials. This integration not only emphasizes the effectiveness of our GBFS workflow but also highlights its compatibility with sophisticated machine-learning methodologies. The study employs a multifidelity modelling approach, synergistically combining high-throughput computational predictions with experimental measurements to achieve improved model accuracy. Such a strategic enhancement of performance exemplifies the adaptability of the workflow and its potential for broader applications in materials informatics, establishing it as a robust tool for advancing predictive modelling in diverse scientific domains.

Overall, this work represents a paradigm shift in materials discovery methodologies, facilitating the function of model regularisation prior to the training and optimisation of predictive models. Our proposed method, which seamlessly integrates with other machine-learning techniques, enables the discovery of new materials across diverse research fields. It does so while minimising computational resources and surpassing the capability of most state-of-the-art methods, thereby accelerating the discovery pipeline compared to traditional approaches in materials research. This enhanced efficiency not only optimises the use of computational assets but also hastens the pace at which new materials can be explored and developed.

6.2 Future Outlook

This section outlines potential areas for future research that could further enhance the application of machine learning in materials informatics. Before achieving a truly universal materials-discovery pipeline, several key research domains stand to benefit significantly from continued advancements. These enhancements will contribute to refining the integration of machine-learning strategies within the broader framework of materials science.

6.2.1 Hybrid Modelling Approaches in Materials Science: Integrating Physics-Based and Data-Driven Models

To advance the field of materials informatics using machine learning for predicting materials properties, several promising avenues for future research can be explored. Hybrid modelling approaches represent a promising yet underexplored frontier in materials science, particularly in enhancing the predictive capabilities and interpretability of models. Here, we aim to outline a structured approach to hybrid modelling by integrating physics-based models with advanced data-driven techniques.

The core concept of hybrid modelling lies in the synthesis of two distinct modelling paradigms: physics-based (or physics-inspired) models and purely data-driven approaches. This synthesis not only aims to leverage the strengths of each approach but also to mitigate their individual weaknesses. Physics-inspired models, such as physics-informed neural networks (PINNs) [292], provide a robust framework grounded in the fundamental laws of physics, typically represented through nonlinear partial differential equations. These models excel in environments where physical laws are well-understood and can guide the learning process, enhancing model interpretability and reliability. On the other hand, purely data-driven models, particularly those based on machine-learning algorithms like the Gradient Boosted Feature Selection (GBFS) workflow, excel in scenarios with large data sets. They adaptively learn from data patterns to make predictions, which makes them highly efficient but sometimes opaque and limited by data quality and quantity.

The integration of these models can be approached by first employing the GBFS workflow to identify and select relevant features from large, complex data sets. This data-driven phase capitalises on the ability of machine learning to process and model large volumes of data efficiently. Following this, the selected features can be fed into a PINN framework, where they are used to train a neural network that not only predicts outcomes based on input data but also conforms to the underlying physical laws. Alternatively, both approaches can be simultaneously trained to develop a meta-learner that synthesises insights from each model to make the final prediction; an approach used for predicting the optical absorption peak of chemical molecules as described in Chapter 5.2. This dual approach allows the hybrid model to benefit from the adaptability and efficiency of data-driven models in handling diverse and large data sets while grounding the predictions within the reliable and interpretable framework provided by physics-based models.

More specifically, the physics-based component of the hybrid model acts as a natural regulator in training phases, particularly beneficial in scenarios characterised by small

data regimes. This regularisation helps prevent overfitting, a common challenge in purely data-driven models. By combining data-driven methods with theoretical or theory-inspired models, hybrid models can facilitate the discovery of new scientific insights, which might not be possible through purely data-driven approaches. They can effectively handle incomplete data and models, making them suitable for exploratory research in materials science. In particular, hybrid models can make accurate predictions even with incomplete data sets by using physical laws to fill in gaps that data cannot cover. This is especially critical in materials science, where experimental data may be costly or difficult to obtain. Moreover, integrating a physics-inspired layer helps make the decision of models more interpretable, which is crucial for scientific validation.

Further research into hybrid modelling should focus on refining the integration techniques between data-driven and physics-informed components to ensure seamless interaction and mutual enhancement. Additionally, expanding the application domains to test these models in various sub-fields of materials science could provide deeper insights and validation of their effectiveness. Hybrid modelling approaches hold significant potential to transform materials science by bridging the gap between theoretical or theory-inspired models and modern methodologies, enhancing both the accuracy and applicability of predictive models in the field.

6.2.2 Expanding the Scope of Machine Learning in Materials Science through Semi-Supervised Learning Frameworks

In materials informatics, a novel but fundamentally distinct approach from hybrid modeling is the adoption of a semi-supervised learning framework. This strategy blends unsupervised learning models with data-driven approaches, such as the Gradient Boosted Feature Selection (GBFS) workflow, to optimise the use of both labelled and unlabelled data. The core of semi-supervised learning involves enhancing the learning process by integrating the strengths of supervised and unsupervised approaches, which is particularly valuable in fields where acquiring extensively labelled data sets is impractical due to high costs or logistical constraints.

Recent advancements in natural language processing (NLP) have seen the successful application of transformer-based language models, pre-trained on vast amounts of unlabelled data. These models, such as the Molecular Language Transformer [37], demonstrate how molecular embeddings can be generated from a corpus of 1.1 billion unlabelled molecules using SMILES sequences. By training a transformer encoder model equipped with an attention mechanism, these unsupervised models have shown signif-

icant capability in capturing essential chemical and structural information—termed molecular representation learning. These molecular language models prove instrumental when fine-tuned on smaller, task-specific data sets, enabling the prediction of diverse molecular properties. In the context of materials informatics, these pre-trained embeddings can be integrated with supervised learning frameworks like the GBFS workflow. The GBFS can utilise these embeddings to enhance feature selection processes and improve prediction accuracy. This integration effectively bridges the gap between the vast unlabelled molecular data and the precise, targeted requirements of supervised learning tasks.

The utility of semi-supervised learning becomes particularly evident in scenarios where the cost or time required to label data sets comprehensively is prohibitive. By leveraging both labelled and unlabelled data, semi-supervised learning can improve learning outcomes without the need for extensive labelled data sets. This makes it exceptionally suitable for applications like NLP and materials property prediction, where labelled data may be scarce, yet there is an abundance of unlabeled data that can provide valuable insights. Therefore, adopting semi-supervised learning frameworks in materials science allows for the effective utilisation of all available data, reducing reliance on expensive and time-consuming data labelling processes. Future research could focus on refining these integrative approaches, enhancing the efficiency and accuracy of materials property predictions. By expanding on these semi-supervised methodologies, researchers could advance the field of materials informatics, paving the way for more efficient and cost-effective materials discovery and analysis.

References

[1] Jain, A.; Hautier, G.; Ong, S. P.; Persson, K. New Opportunities for Materials Informatics: Resources and Data Mining Techniques for Uncovering Hidden Relationships. *Mater. Res.* **2016**, *31*, 977–994.

[2] Rajan, K. Materials Informatics. *Mater. Today* **2005**, *8*, 38–45.

[3] Kohn, W. Nobel Lecture: Electronic Structure of Matter-Wave Functions and Density Functionals. *Rev. Mod. Phys.* **1999**, *71*, 1253.

[4] Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191–201.

[5] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

[6] Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on REpresentational State Transfer (REST) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209–215.

[7] Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *J. Miner. Met. Mater. Soc.* **2013**, *65*, 1501–1509.

[8] Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction Model of Band Gap for Inorganic Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques. *Phys. Rev. B* **2016**, *93*, 115104.

[9] Pilania, G.; Gubernatis, J.; Lookman, T. Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.

[10] Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 1668–1673.

[11] Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including Crystal Structure Attributes in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Phys. Rev. B* **2017**, *96*, 024104.

[12] You, D.; Ganorkar, S.; Kim, S.; Kang, K.; Shin, W.-Y.; Lee, D. Machine Learning-Based Prediction Models for Formation Energies of Interstitial Atoms in HCP Crystals. *Scr. Mater* **2020**, *183*, 1–5.

[13] Kang, P.; Liu, Z.; Abou-Rachid, H.; Guo, H. Machine-Learning Assisted Screening of Energetic Materials. *J. Phys. Chem. A* **2020**, *124*, 5341–5351.

[14] Shandiz, M. A.; Gauvin, R. Application of Machine Learning Methods for the Prediction of Crystal System of Cathode Materials in Lithium-Ion Batteries. *Comput. Mater. Sci.* **2016**, *117*, 270–278.

[15] Jalem, R.; Kanamori, K.; Takeuchi, I.; Nakayama, M.; Yamasaki, H.; Saito, T. Bayesian-Driven First-Principles Calculations for Accelerating Exploration of Fast Ion Conductors for Rechargeable Battery Application. *Sci. Rep.* **2018**, *8*, 1–10.

[16] Butler, K. T.; Frost, J. M.; Skelton, J. M.; Svane, K. L.; Walsh, A. Computational Materials Design of Crystalline Solids. *Chem. Soc. Rev.* **2016**, *45*, 6138–6146.

[17] Li, X.; Dan, Y.; Dong, R.; Cao, Z.; Niu, C.; Song, Y.; Li, S.; Hu, J. Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning. *Appl. Sci.* **2019**, *9*, 5510.

[18] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, 54.

[19] Jose, R.; Ramakrishna, S. Materials 4.0: Materials Big Data Enabled Materials Discovery. *Appl. Mater. Today* **2018**, *10*, 127–132.

[20] Ramakrishna, S.; Zhang, T.-Y.; Lu, W.-C.; Qian, Q.; Low, J. S. C.; Yune, J. H. R.; Tan, D. Z. L.; Bressan, S.; Sanvito, S.; Kalidindi, S. R. Materials Informatics. *J. Intell. Manuf.* **2019**, *30*, 2307–2326.

[21] Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information From the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904.

[22] Mavracic, J.; Court, C. J.; Isazawa, T.; Elliott, S. R.; Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *J. Chem. Inf. Model.* **2021**, *61*, 4280–4289.

[23] Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; Lei, M. Machine Learning in Materials Science. *InfoMat* **2019**, *1*, 338–358.

[24] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Materiomics* **2017**, *3*, 159–177.

[25] Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5*, 83.

[26] Jung, G.; Jung, S. G.; Cole, J. M. Automatic Materials Characterization From Infrared Spectra Using Convolutional Neural Networks. *Chem. Sci.* **2023**, *14*, 3600–3609.

[27] Bart'ok, A. P.; Kondor, R.; Cs'anyi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

[28] Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

[29] Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101.

[30] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

[31] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem* **2015**, *6*, 2326–2331.

[32] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Advances in Neural Information Processing Systems. 2015.

[33] Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

[34] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572.

[35] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. Proceedings of the 34th International Conference on Machine Learning. 2017; pp 1263–1272.

[36] Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv preprint arXiv:2010.09885* **2020**,

[37] Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties. *Nat. Mach. Intell.* **2022**, *4*, 1256–1264.

[38] Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular Contrastive Learning of Representations via Graph Neural Networks. *Nat. Mach. Intell.* **2022**, *4*, 279–287.

[39] Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[40] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, 1189–1232.

[41] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.

[42] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017; p 3149–3157.

[43] Doncaster, C. P.; Davey, A. J. *Analysis of Variance and Covariance: How to Choose and Construct Models for the Life Sciences*; Cambridge University Press, 2007.

[44] Anderson, J. 7 Logistic Discrimination. *Handb. Stat.* **1982**, *2*, 169–191.

[45] Timm, N. H. *Applied Multivariate Analysis*; Springer, 2002.

[46] Kraskov, A.; St"ogbauer, H.; Grassberger, P. Estimating Mutual Information. *Phys. Rev. E* **2004**, *69*, 066138.

[47] Ross, B. C. Mutual Information Between Discrete and Continuous Data Sets. *PloS ONE* **2014**, *9*, e87357.

[48] Kozachenko, L.; Leonenko, N. N. Sample Estimate of the Entropy of a Random Vector. *Probl. Inf. Transm.* **1987**, *23*, 9–16.

[49] Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT press, 2012.

[50] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, *58*, 267–288.

[51] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks From Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

[52] Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2. Red Hook, NY, USA, 2012; p 2951–2959.

[53] Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; De Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2015**, *104*, 148–175.

[54] Rasmussen, C. E. Gaussian Processes in Machine Learning. Proceedings of the Summer School on Machine Learning. 2003; pp 63–71.

[55] Williams, C. K.; Rasmussen, C. E. *Gaussian Processes for Machine Learning*; MIT Press Cambridge, MA, 2006; Vol. 2.

[56] Tipping, M. E. *Advanced Lectures on Machine Learning*; Springer, 2003; pp 41–62.

[57] Kushner, H. J. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *J. Basic Eng.* **1964**, *86*, 97–106.

[58] Mockus, J.; Tiesis, V.; Zilinskas, A. The Application of Bayesian Methods for Seeking the Extremum. *Towards Glob. Optim.* **1978**, *2*, 117–129.

[59] Srinivas, N.; Krause, A.; Kakade, S.; Seeger, M. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA, 2010; p 1015–1022.

[60] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics; Springer, 2009.

[61] Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.

[62] Ng, A. Y. Feature Selection, L1 Vs. L2 Regularization, and Rotational Invariance. Proceedings of the 21st International Conference on Machine Learning. 2004; p 78.

[63] McDonald, G. C. Ridge Regression. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1*, 93–100.

[64] McDonald, G. C.; Schwing, R. C. Instabilities of Regression Estimates Relating Air Pollution to Mortality. *Technometrics* **1973**, *15*, 463–481.

[65] Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045017.

[66] Choudhary, K.; DeCost, B.; Tavazza, F. Machine Learning with Force-Field-Inspired Descriptors for Materials: Fast Screening and Mapping Energy Landscape. *Phys. Rev. Mater.* **2018**, *2*, 083801.

[67] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

[68] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 1–7.

[69] Deml, A. M.; O'Hayre, R.; Wolverton, C.; Stevanović, V. Predicting Density Functional Theory Total Energies and Enthalpies of Formation of Metal-Nonmetal Compounds by Linear Regression. *Phys. Rev. B* **2016**, *93*, 085142.

[70] Ward, L. et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.

[71] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

[72] Landrum, G. RDKit: Open-Source Cheminformatics. 2016; https://www.rdkit.org, (Accessed: 01 June 2022).

[73] Jha, D.; Ward, L.; Paul, A.; Liao, W.-K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. reports* **2018**, *8*, 1–13.

[74] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

[75] Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

[76] Zefirov, N. S. The Problem of Conformational Effects. *Tetrahedron* **1977**, *33*, 3192.

[77] Urban, A.; Seo, D. H.; Ceder, G. Computational Understanding of Li-Ion Batteries. *npj Comput. Mater.* **2016**, *2*, 1–13.

[78] Jónsson, E.; Johansson, P. Modern Battery Electrolytes: Ion–Ion Interactions in Li+/Na+ Conductors From DFT Calculations. *Phys. Chem. Chem. Phys.* **2012**, *14*, 10774–10779.

[79] Dobrota, A. S.; Pašti, I. A.; Mentus, S. V.; Johansson, B.; Skorodumova, N. V. Functionalized Graphene for Sodium Battery Applications: The DFT Insights. *Electrochim. Acta* **2017**, *250*, 185–195.

[80] Aykol, M.; Kim, S.; Hegde, V. I.; Snydacker, D.; Lu, Z.; Hao, S.; Kirklin, S.; Morgan, D.; Wolverton, C. High-Throughput Computational Design of Cathode Coatings for Li-Ion Batteries. *Nat. Commun.* **2016**, *7*, 1–12.

[81] Shi, S.; Gao, J.; Liu, Y.; Zhao, Y.; Wu, Q.; Ju, W.; Ouyang, C.; Xiao, R. Multi-Scale Computation Methods: Their Applications in Lithium-Ion Battery Research and Development. *Chin. Phys. B* **2015**, *25*, 018212.

[82] Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational Screening of Perovskite Metal Oxides for Optimal Solar Light Capture. *Energy Environ. Sci.* **2012**, *5*, 5814–5819.

[83] Wu, Y.; Lazic, P.; Hautier, G.; Persson, K.; Ceder, G. First Principles High Throughput Screening of Oxynitrides for Water-Splitting Photocatalysts. *Energy Environ. Sci.* **2013**, *6*, 157–168.

[84] Kuhar, K.; Crovetto, A.; Pandey, M.; Thygesen, K. S.; Seger, B.; Vesborg, P. C.; Hansen, O.; Chorkendorff, I.; Jacobsen, K. W. Sulfide Perovskites for Solar Energy Conversion Applications: Computational Screening and Synthesis of the Selected Compound $LaYS_3$. *Energy Environ. Sci.* **2017**, *10*, 2579–2593.

[85] Chen, Y.-H.; Lin, L.-Y.; Lu, C.-W.; Lin, F.; Huang, Z.-Y.; Lin, H.-W.; Wang, P.-H.; Liu, Y.-H.; Wong, K.-T.; Wen, J. Vacuum-Deposited Small-Molecule Organic Solar Cells with High Power Conversion Efficiencies by Judicious Molecular Design and Device Optimization. *J. Am. Chem. Soc.* **2012**, *134*, 13616–13623.

[86] Shan, Q.; Song, J.; Zou, Y.; Li, J.; Xu, L.; Xue, J.; Dong, Y.; Han, B.; Chen, J.; Zeng, H. High Performance Metal Halide Perovskite Light-Emitting Diode: From Material Design to Device Optimization. *Small* **2017**, *13*, 1701770.

[87] Zhang, K.; Hu, Z.; Sun, C.; Wu, Z.; Huang, F.; Cao, Y. Toward Solution-Processed High-Performance Polymer Solar Cells: From Material Design to Device Engineering. *Chem. Mater.* **2017**, *29*, 141–148.

[88] Wan, J.; Xu, X.; Zhang, G.; Li, Y.; Feng, K.; Peng, Q. Highly Efficient Halogen-Free Solvent Processed Small-Molecule Organic Solar Cells Enabled by Material Design and Device Engineering. *Energy Environ. Sci.* **2017**, *10*, 1739–1745.

[89] Bellman, R. E. *Adaptive Control Processes: A Guided Tour*; Princeton University Press, 2015.

[90] Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput. Mater.* **2020**, *6*, 138.

[91] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

[92] Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Proceedings of the International Conference on Intelligent Computing. 2005; pp 878–887.

[93] He, H.; Bai, Y.; Garcia, E. A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008; pp 1322–1328.

[94] Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Artif. Intell. Res.* **2017**, *18*, 559–563.

[95] Menardi, G.; Torelli, N. Training and Assessing Classification Rules with Imbalanced Data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122.

[96] Mehta, M.; Agrawal, R.; Rissanen, J. SLIQ: A Fast Scalable Classifier for Data Mining. Proceedings of the International Conference on Extending Database Technology. 1996; pp 18–32.

[97] Shafer, J.; Agrawal, R.; Mehta, M. SPRINT: A Scalable Parallel Classifier for Data Mining. Proceedings of the 22nd International Conference on Very Large Data Bases. 1996; pp 544–555.

[98] Ranka, S.; Singh, V. CLOUDS: A Decision Tree Classifier for Large Datasets. Proceedings of the 4th Knowledge Discovery and Data Mining Conference. 1998.

[99] Jin, R.; Agrawal, G. Communication and Memory Efficient Parallel Decision Tree Construction. Proceedings of the 2003 SIAM International Conference on Data Mining. 2003; pp 119–129.

[100] Li, P.; Wu, Q.; Burges, C. McRank: Learning to Rank Using Multiple Classification and Gradient Boosting. *Adv. Neural Inf. Process. Syst.* **2007**, *20*, 897–904.

[101] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

[102] Ruxton, G. D. The Unequal Variance T-Test is an Underused Alternative to Student's T-Test and the Mann–Whitney U Test. *Behav. Ecol.* **2006**, *17*, 688–690.

[103] Kim, T. K. T-Test as a Parametric Statistic. *Korean J. Anesthesiol.* **2015**, *68*, 540.

[104] Satterthwaite, F. E. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* **1946**, *2*, 110–114.

[105] Welch, B. L. The Generalization of 'Student's' Problem When Several Different Population Variances Are Involved. *Biometrika* **1947**, *34*, 28–35.

[106] Paternoster, R.; Brame, R.; Mazerolle, P.; Piquero, A. Using the Correct Statistical Test for the Equality of Regression Coefficients. *Criminology* **1998**, *36*, 859–866.

[107] McHugh, M. L. The Chi-Square Test of Independence. *Biochem. Med.* **2013**, *23*, 143–149.

[108] Boslaugh, S. *Statistics in a Nutshell: A Desktop Quick Reference*; O'Reilly Media, Inc., 2012.

[109] Cram'er, H. *Mathematical Methods of Statistics (PMS-9)*; Princeton University Press, 2016; Vol. 9.

[110] Batina, L.; Gierlichs, B.; Prouff, E.; Rivain, M.; Standaert, F.-X.; Veyrat-Charvillon, N. Mutual Information Analysis: A Comprehensive Study. *J. Cryptol.* **2011**, *24*, 269–291.

[111] Hosmer Jr, D. W.; Lemeshow, S.; Sturdivant, R. X. *Applied Logistic Regression*; John Wiley & Sons, 2013; Vol. 398.

[112] Ward Jr, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

[113] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422.

[114] Coey, J. M. *Magnetism and Magnetic Materials*; Cambridge University Press, 2010.

[115] Slonczewski, J. Current-Driven Excitation of Magnetic Multilayers. *J. Magn. Magn. Mater.* **1996**, *159*, L1–L7.

[116] Nowak, J. J.; Robertazzi, R. P.; Sun, J. Z.; Hu, G.; Park, J.-H.; Lee, J.; Annunziata, A. J.; Lauer, G. P.; Kothandaraman, R.; O'Sullivan, E. J.; Trouilloud, P. L.; Kim, Y.; Worledge, D. C. Dependence of Voltage and Size on Write Error Rates in Spin-Transfer Torque Magnetic Random-Access Memory. *IEEE Magn. Lett.* **2016**, *7*, 1–4.

[117] Parkin, S. S. P.; Hayashi, M.; Thomas, L. Magnetic Domain-Wall Racetrack Memory. *Science* **2008**, *320*, 190–194.

[118] Gartside, J. C.; Jung, S. G.; Yoo, S. Y.; Arroo, D. M.; Vanstone, A.; Dion, T.; Stenning, K. D.; Branford, W. R. Current-Controlled Nanomagnetic Writing for Reconfigurable Magnonic Crystals. *Commun. Phys.* **2020**, *3*, 219.

[119] Kikuchi, R. On the Minimum of Magnetization Reversal Time. *J. Appl. Phys.* **2004**, *27*, 1352–1357.

[120] Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094–1101.

[121] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.

[122] Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. Recent Advances and Applications of Deep Learning Methods in Materials Science. *npj Comput. Mater.* **2022**, *8*, 59.

[123] Jung, S. G.; Jung, G.; Cole, J. M. Gradient Boosted and Statistical Feature Selection Workflow for Materials Property Predictions. *J. Chem. Phy.* **2023**, *159*, 194106.

[124] Jung, S. G.; Jung, G.; Cole, J. M. Predictive Modeling of High-Entropy Alloys and Amorphous Metallic Alloys Using Machine Learning. *J. Chem. Inf. Model.* **2024**, *64*, 7313–7336.

[125] Jung, S. G.; Jung, G.; Cole, J. M. Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow. *J. Chem. Inf. Model.* **2024**, *64*, 1187–1200.

[126] Jung, S. G.; Jung, G.; Cole, J. M. Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks. *J. Chem. Inf. Model.* **2024**, *64*, 1486–1501.

[127] Court, C. J.; Jain, A.; Cole, J. M. Inverse Design of Materials That Exhibit the Magnetocaloric Effect by Text-Mining of the Scientific Literature and Generative Deep Learning. *Chem. Mater.* **2021**, *33*, 7217–7231.

[128] Ucar, H.; Paudyal, D.; Choudhary, K. Machine Learning Predicted Magnetic Entropy Change Using Chemical Descriptors Across a Large Compositional Landscape. *Comput. Mater. Sci.* **2022**, *209*, 111414.

[129] Long, T.; Fortunato, N. M.; Zhang, Y.; Gutfleisch, O.; Zhang, H. An Accelerating Approach of Designing Ferromagnetic Materials via Machine Learning Modeling of Magnetic Ground State and Curie Temperature. *Mater. Res. Lett.* **2021**, *9*, 169–174.

[130] Dam, H. C.; Nguyen, V. C.; Pham, T. L.; Nguyen, A. T.; Kino, H.; Terakura, K.; Miyake, T. A Regression-Based Feature Selection Study of the Curie Temperature of Transition-Metal Rare-Earth Compounds: Prediction and Understanding. *arXiv preprint arXiv:1705.00978* **2017**,

[131] Nelson, J.; Sanvito, S. Predicting the Curie Temperature of Ferromagnets Using Machine Learning. *Phys. Rev. Mater.* **2019**, *3*, 104405.

[132] Belot, J. F.; Taufour, V.; Sanvito, S.; Hart, G. L. W. Machine Learning Predictions of High-Curie-Temperature Materials. *APL* **2023**, *123*, 042405.

[133] Singh, P.; Del Rose, T.; Palasyuk, A.; Mudryk, Y. Physics-Informed Machine-Learning Prediction of Curie Temperatures and Its Promise for Guiding the Discovery of Functional Magnetic Materials. *Chem. Mater.* **2023**, *35*, 6304–6312.

[134] Sanvito, S.; Oses, C.; Xue, J.; Tiwari, A.; Zic, M.; Archer, T.; Tozman, P.; Venkatesan, M.; Coey, M.; Curtarolo, S. Accelerated Discovery of New Magnets in the Heusler Alloy Family. *Sci. Adv.* **2017**, *3*, e1602241.

[135] Crangle, J.; Parsons, D. The Magnetization of Ferromagnetic Binary Alloys of Cobalt or Nickel with Elements of the Palladium and Platinum Groups. *Proc. R. Soc. Lond. A* **1960**, *255*, 509–519.

[136] Albert, H. J.; Rubin, L. R. *Platinum Group Metals and Compounds*; Chapter 1, pp 1–16.

[137] Franco, V.; Blázquez, J.; Ipus, J.; Law, J.; Moreno-Ramírez, L.; Conde, A. Magnetocaloric Effect: From Materials Research to Refrigeration Devices. *Prog. Mater. Sci.* **2018**, *93*, 112–232.

[138] Lu, S.; Zhou, Q.; Guo, Y.; Wang, J. On-The-Fly Interpretable Machine Learning for Rapid Discovery of Two-Dimensional Ferromagnets with High Curie Temperature. *Chem* **2022**, *8*, 769–783.

[139] Byland, J. K.; Shi, Y.; Parker, D. S.; Zhao, J.; Ding, S.; Mata, R.; Magliari, H. E.; Palasyuk, A.; Bud'ko, S. L.; Canfield, P. C.; Klavins, P.; Taufour, V. Statistics on Magnetic Properties of Co Compounds: A Database-Driven Method for Discovering Co-Based Ferromagnets. *Phys. Rev. Mater.* **2022**, *6*, 063803.

[140] Buschow, K. H. J. Intermetallic Compounds of Rare-Earth and 3d Transition Metals. *Reports on Progress in Physics* **1977**, *40*, 1179.

[141] Connolly, T. F.; Copenhaver, E. D. *Bibliography of Magnetic Materials and Tabulation of Magnetic Transition Temperatures*; Springer US: Boston, MA, 1972; pp 1–29.

[142] Xu, Y.; Yamazaki, M.; Villars, P. Inorganic Materials Database for Exploring the Nature of Material. *Jpn. J. Appl. Phys.* **2011**, *50*, 11RH02.

[143] Fecher, G. H.; Kandpal, H. C.; Wurmehl, S.; Felser, C.; Schönhense, G. Slater-Pauling Rule and Curie Temperature of Co_2-Based Heusler Compounds. *J. Appl. Phys.* **2006**, *99*.

[144] Slater, J. C. The Ferromagnetism of Nickel. II. Temperature Effects. *Phys. Rev.* **1936**, *49*, 931.

[145] Pauling, L. The Nature of the Interatomic Forces in Metals. *Phys. Rev.* **1938**, *54*, 899.

[146] Takahashi, C.; Ogura, M.; Akai, H. First-Principles Calculation of the Curie Temperature Slater–Pauling Curve. *J. Phys. Condens. Matter* **2007**, *19*, 365233.

[147] Galanakis, I.; Dederichs, P.; Papanikolaou, N. Slater-Pauling Behavior and Origin of the Half-Metallicity of the Full-Heusler Alloys. *Phys. Rev. B* **2002**, *66*, 174429.

[148] Wurmehl, S.; Fecher, G. H.; Kandpal, H. C.; Ksenofontov, V.; Felser, C.; Lin, H.-J.; Morais, J. Geometric, Electronic, and Magnetic Structure of Co_2 FeSi: Curie Temperature and Magnetic Moment Measurements and Calculations. *Phys. Rev. B* **2005**, *72*, 184434.

[149] Balke, B.; Wurmehl, S.; Fecher, G. H.; Felser, C.; Kübler, J. Rational Design of New Materials for Spintronics: Co_2FeZ (Z = Al, Ga, Si, Ge). *Sci. Technol. Adv. Mat.* **2008**, *9*, 014102.

[150] Kundu, A.; Ghosh, S.; Banerjee, R.; Ghosh, S.; Sanyal, B. New Quaternary Half-Metallic Ferromagnets with Large Curie Temperatures. *Sci. Rep.* **2017**, *7*, 1803.

[151] Castelliz, L. Contribution to the Ferromagnetism of Alloys of Ütransition Metals with Elements of the B Group. *Int. J. Mater. Res.* **1955**, *46*, 198–203.

[152] Kanomata, T.; Shirakawa, K.; Kaneko, T. Effect of Hydrostatic Pressure on the Curie Temperature of the Heusler Alloys Ni_2MnZ (Z = Al, Ga, In, Sn and Sb). *J. Magn. Magn. Mater.* **1987**, *65*, 76–82.

[153] Sanvito, S.; Oses, C.; Xue, J.; Tiwari, A.; Zic, M.; Archer, T.; Tozman, P.; Venkatesan, M.; Coey, M.; Curtarolo, S. Accelerated Discovery of New Magnets in the Heusler Alloy Family. *Sci. Adv.* **2017**, *3*, e1602241.

[154] Adachi, Y.; Morita, H.; Kanomata, T.; Sato, A.; Yoshida, H.; Kaneko, T.; Nishihara, H. Pressure Effect on the Curie Temperature of the Heusler Alloys Rh_2MnZ (Z = Sn, Ge). *J. Alloys Compd.* **2004**, *383*, 37–39.

[155] Coey, J.; Venkatesan, M.; Fitzgerald, C. Donor Impurity Band Exchange in Dilute Ferromagnetic Oxides. *Nature materials* **2005**, *4*, 173–179.

[156] Pan, F.; Song, C.; Liu, X.; Yang, Y.; Zeng, F. Ferromagnetism and Possible Application in Spintronics of Transition-Metal-Doped ZnO Films. *Materials Science and Engineering: R: Reports* **2008**, *62*, 1–35.

[157] Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017; p 4768–4777.

[158] Ishida, K.; Nishizawa, T. The Co-Mn (Cobalt-Manganese) System. *Bull. Alloy Phase Diagr.* **1990**, *11*, 125–137.

[159] Menshikov, A.; Takzei, G.; Dorofeev, Y. A.; Kazantsev, V.; Kostyshin, A.; Sych, I. Magnetic Phase Diagram of Cobalt–Manganese Alloys. *Zh. Eksp. Teor. Fiz.* **1985**, *89*, 1269–1279.

[160] Pickles, A.; Sucksmith, W. A Magnetic Study of the Two-Phase Iron-Nickel Alloys. *Proc. R. Soc. Lond. A* **1940**, *175*, 331–344.

[161] Swartzendruber, L.; Itkin, V.; Alcock, C. The Fe-Ni (Iron-Nickel) System. *J. Phase Equilib.* **1991**, *12*, 288–312.

[162] Xiong, W.; Zhang, H.; Vitos, L.; Selleby, M. Magnetic Phase Diagram of the Fe–Ni System. *Acta Materialia* **2011**, *59*, 521–530.

[163] Cacciamani, G.; Dinsdale, A.; Palumbo, M.; Pasturel, A. The Fe–Ni System: Thermodynamic Modelling Assisted by Atomistic Calculations. *Intermetallics* **2010**, *18*, 1148–1162.

[164] Sucksmith, W. The Measurement of Magnetic Saturation Intensities at Different Temperatures. *Proc. R. Soc. Lond. A* **1939**, *170*, 551–560.

[165] Hoselitz, K.; Sucksmith, W. A Magnetic Study of the Two-Phase Iron-Nickel Alloys. II. *Proc. R. Soc. Lond. A* **1943**, *181*, 303–313.

[166] Kaufman, L.; Ringwood, A. High Pressure Equilibria in the Iron-Nickel System and the Structure of Metallic Meteorites. *Acta Metall.* **1961**, *9*, 941–944.

[167] Shishkin, M.; Kresse, G. Self-Consistent *GW* Calculations for Semiconductors and Insulators. *Phys. Rev. B* **2007**, *75*, 235102.

[168] Skelton, J. M.; Parker, S. C.; Togo, A.; Tanaka, I.; Walsh, A. Thermal Physics of the Lead Chalcogenides PbS, PbSe, and PbTe From First Principles. *Phys. Rev. B* **2014**, *89*, 205203.

[169] Yu, L.; Kokenyesi, R. S.; Keszler, D. A.; Zunger, A. Inverse Design of High Absorption Thin-Film Photovoltaic Materials. *Adv. Energy Mater.* **2013**, *3*, 43–48.

[170] Tran, F.; Blaha, P. Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential. *Phys. Rev. Lett.* **2009**, *102*, 226401.

[171] Gerosa, M.; Bottani, C. E.; Caramella, L.; Onida, G.; Di Valentin, C.; Pacchioni, G. Electronic Structure and Phase Stability of Oxide Semiconductors: Performance of Dielectric-Dependent Hybrid Functional DFT, Benchmarked Against GW Band Structure Calculations and Experiments. *Phys. Rev. B* **2015**, *91*, 155201.

[172] Schimka, L.; Harl, J.; Kresse, G. Improved Hybrid Functional for Solids: The HSEsol Functional. *J. Chem. Phys.* **2011**, *134*, 024116.

[173] Kiselyova, N. N.; Dudarev, V. A.; Korzhuyev, M. A. Database on the Bandgap of Inorganic Substances and Materials. *Inorg. Mater. Appl. Res.* **2016**, *7*, 34–39.

[174] Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; Smola, A. J. Deep Sets. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

[175] Keller, J. M.; Gray, M. R.; Givens, J. A. A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Trans. Syst. Man. Cybern.* **1985**, 580–585.

[176] Weisberg, S. *Applied Linear Regression*; John Wiley & Sons, 2005; Vol. 528.

[177] Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

[178] Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

[179] Hearst, M. A.; Dumais, S. T.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst.* **1998**, *13*, 18–28.

[180] Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.

[181] Stevanović, V.; Lany, S.; Zhang, X.; Zunger, A. Correcting Density Functional Theory for Accurate Predictions of Compound Enthalpies of Formation: Fitted Elemental-Phase Reference Energies. *Phys. Rev. B* **2012**, *85*, 115104.

[182] Hill, R. The Elastic Behaviour of a Crystalline Aggregate. *Proc. Phys. Soc. A* **1952**, *65*, 349.

[183] Vitos, L.; Abrikosov, I.; Johansson, B. Anisotropic Lattice Distortions in Random Alloys From First-Principles Theory. *Phys. Rev. Lett.* **2001**, *87*, 156401.

[184] Vitos, L. *Computational Quantum Mechanics for Materials Engineers: The EMTO Method and Applications*; Springer Science & Business Media, 2007.

[185] Zhang, J.; Cai, C.; Kim, G.; Wang, Y.; Chen, W. Composition Design of High-Entropy Alloys with Deep Sets Learning. *npj Comput. Mater.* **2022**, *8*, 89.

[186] De Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Krishna Ande, C.; Van Der Zwaag, S.; Plata, J. J.; Toher, C.; Curtarolo, S.; Ceder, G.; Persson, K.; Asta, M. Charting the Complete Elastic Properties of Inorganic Crystalline Compounds. *Sci.* **2015**, *2*, 1–13.

[187] Tian, F.; Delczeg, L.; Chen, N.; Varga, L. K.; Shen, J.; Vitos, L. Structural Stability of $NiCoFeCrAl_x$ High-Entropy Alloy From Ab Initio Theory. *Phys. Rev. B* **2013**, *88*, 085128.

[188] Xiong, J.; Shi, S.-Q.; Zhang, T.-Y. A Machine-Learning Approach to Predicting and Understanding the Properties of Amorphous Metallic Alloys. *Mater. Des.* **2020**, *187*, 108378.

[189] Miedema, A.; De Chatel, P.; De Boer, F. Cohesion in Alloys—Fundamentals of a Semi-Empirical Model. *Physica B+C* **1980**, *100*, 1–28.

[190] Inoue, A. Stabilization of Metallic Supercooled Liquid and Bulk Amorphous Alloys. *Acta Mater.* **2000**, *48*, 279–306.

[191] Lu, Z.; Li, Y.; Ng, S. Reduced Glass Transition Temperature and Glass Forming Ability of Bulk Glass Forming Alloys. *J. Non-Cryst. Solids* **2000**, *270*, 103–114.

[192] Lu, Z.; Tan, H.; Ng, S.; Li, Y. The Correlation Between Reduced Glass Transition Temperature and Glass Forming Ability of Bulk Metallic Glasses. *Scr. Mater.* **2000**, *42*, 667–673.

[193] Wang, W. H. The Elastic Properties, Elastic Models and Elastic Perspectives of Metallic Glasses. *Prog. Mater. Sci.* **2012**, *57*, 487–656.

[194] Stanev, V.; Oses, C.; Kusne, A. G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine Learning Modeling of Superconducting Critical Temperature. *npj Comput. Mater.* **2018**, *4*, 29.

[195] Villars, P.; Phillips, J. C. Quantum Structural Diagrams and High-T_c Superconductivity. *Phys. Rev. B* **1988**, *37*, 2345.

[196] Rabe, K.; Phillips, J.; Villars, P.; Brown, I. Global Multinary Structural Chemistry of Stable Quasicrystals, High-T_c Ferroelectrics, and High-T_c Superconductors. *Phys. Rev. B* **1992**, *45*, 7650.

[197] Maxwell, E. Isotope Effect in the Superconductivity of Mercury. *Phys. Rev.* **1950**, *78*, 477.

[198] Reynolds, C.; Serin, B.; Wright, W.; Nesbitt, L. Superconductivity of Isotopes of Mercury. *Phys. Rev.* **1950**, *78*, 487.

[199] Reynolds, C. A.; Serin, B.; Nesbitt, L. B. The Isotope Effect in Superconductivity. I. Mercury. *Phys. Rev.* **1951**, *84*, 691–694.

[200] Bardeen, J. In *Cooperative Phenomena*; Haken, H., Wagner, M., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 1973; pp 63–78.

[201] Bardeen, J.; Cooper, L. N.; Schrieffer, J. R. Theory of Superconductivity. *Phys. Rev.* **1957**, *108*, 1175.

[202] Lorenz, B.; Chu, C. In *Frontiers in Superconducting Materials*; Narlikar, A. V., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2005; pp 459–497.

[203] Hosono, H.; Tanabe, K.; Takayama-Muromachi, E.; Kageyama, H.; Yamanaka, S.; Kumakura, H.; Nohara, M.; Hiramatsu, H.; Fujitsu, S. Exploration of New Superconductors and Functional Materials, and Fabrication of Superconducting Tapes and Wires of Iron Pnictides. *Sci. Technol. Adv. Mat.* **2015**, *16*, 033503.

[204] Konno, T.; Kurokawa, H.; Nabeshima, F.; Sakishita, Y.; Ogawa, R.; Hosako, I.; Maeda, A. Deep Learning Model for Finding New Superconductors. *Phys. Rev. B* **2021**, *103*, 014509.

[205] Pereti, C.; Bernot, K.; Guizouarn, T.; Laufek, F.; Vymazalová, A.; Bindi, L.; Sessoli, R.; Fanelli, D. From Individual Elements to Macroscopic Materials: In Search of New Superconductors via Machine Learning. *npj Comput. Mater.* **2023**, *9*, 71.

[206] Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R. R.; Smola, A. J. Deep Sets. *Adv. Neural Inf. Process Syst.* **2017**, *30*.

[207] Pauling, L. C. A Resonating-Valence-Bond Theory of Metals and Intermetallic Compounds. *Proc. R. Soc. A: Math. Phys. Sci.* **1949**, *196*, 343–362.

[208] Anderson, P. W. New Approach to the Theory of Superexchange Interactions. *Phys. Rev.* **1959**, *115*, 2.

[209] Chao, K.; Spałek, J.; Oleś, A. Canonical Perturbation Expansion of the Hubbard Model. *Phys. Rev. B* **1978**, *18*, 3453.

[210] Anderson, P. W.; Baskaran, G.; Zou, Z.; Hsu, T. Resonating–Valence-Bond Theory of Phase Transitions and Superconductivity in La_2CuO_4-Based Compounds. *Phys. Rev. Lett.* **1987**, *58*, 2790.

[211] Shastry, B. S. $t-j$ Model and Nuclear Magnetic Relaxation in High-T_c Materials. *Phys. Rev. Lett.* **1989**, *63*, 1288.

[212] Emery, V.; Kivelson, S.; Lin, H. Phase Separation in the $t-j$ Model. *Phys. Rev. Lett.* **1990**, *64*, 475.

[213] Izyumov, Y. A. Strongly Correlated Electrons: The $t-j$ Model. *Physics-Uspekhi* **1997**, *40*, 445.

[214] Matthias, B. T. Empirical Relation Between Superconductivity and the Number of Valence Electrons Per Atom. *Phys. Rev.* **1955**, *97*, 74.

[215] Maeda, H.; Tanaka, Y.; Fukutomi, M.; Asano, T. A New High-T_c Oxide Superconductor Without a Rare Earth Element. *Jpn. J. Appl. Phys.* **1988**, *27*, L209.

[216] Schilling, A.; Cantoni, M.; Guo, J.; Ott, H. Superconductivity Above 130 K in the Hg–Ba–Ca–Cu–O System. *Nature* **1993**, *363*, 56–58.

[217] Gao, L.; Xue, Y.; Chen, F.; Xiong, Q.; Meng, R.; Ramirez, D.; Chu, C.; Eggert, J.; Mao, H. Superconductivity up to 164 K in $HgBa_2Ca_{m-1}Cu_mO_{2m+2+\delta}$ (m = 1, 2, and 3) Under Quasihydrostatic Pressures. *Phys. Rev. B* **1994**, *50*, 4260.

[218] Eggert, J.; Hu, J.; Mao, H.; Beauvais, L.; Meng, R.; Chu, C. Compressibility of the $HgBa_2Ca_{n-1}Cu_nO_{2n+2+\delta}$ (n = 1, 2, 3) High-Temperature Superconductors. *Phys. Rev. B* **1994**, *49*, 15299.

[219] Weller, T. E.; Ellerby, M.; Saxena, S. S.; Smith, R. P.; Skipper, N. T. Superconductivity in the Intercalated Graphite Compounds C_6Yb and C_6Ca. *Nature Phys.* **2005**, *1*, 39–41.

[220] Li, Z.; He, X.; Zhang, C.; Wang, X.; Zhang, S.; Jia, Y.; Feng, S.; Lu, K.; Zhao, J.; Zhang, J. Superconductivity Above 200 K Discovered in Superhydrides of Calcium. *Nat. Commun.* **2022**, *13*, 2863.

[221] Kortan, A.; Kopylov, N.; Glarum, S.; Gyorgy, E.; Ramirez, A.; Fleming, R.; Thiel, F.; Haddon, R. Superconductivity at 8.4 K in Calcium-Doped C_{60}. *Nature* **1992**, *355*, 529–532.

[222] Wang, H.; Tse, J. S.; Tanaka, K.; Iitaka, T.; Ma, Y. Superconductive Sodalite-Like Clathrate Calcium Hydride at High Pressures. *Proc. Nat. Acad. Sci.* **2012**, *109*, 6463–6466.

[223] Ichinokura, S.; Sugawara, K.; Takayama, A.; Takahashi, T.; Hasegawa, S. Superconducting Calcium-Intercalated Bilayer Graphene. *ACS Nano* **2016**, *10*, 2761–2765.

[224] Markiewicz, R. Van Hove Singularities and High-T_c Superconductivity: A Review. *Int. J. Mod. Phys. B* **1991**, *5*, 2037–2071.

[225] Hirsch, J.; Scalapino, D. Enhanced Superconductivity in Quasi Two-Dimensional Systems. *Phys. Rev. Lett.* **1986**, *56*, 2732.

[226] Kudo, K.; Kitahama, Y.; Fujimura, K.; Mizukami, T.; Ota, H.; Nohara, M. Superconducting Transition Temperatures of up to 47 K From Simultaneous Rare-Earth Element and Antimony Doping of 112-Type $CaFeAs_2$. *J. Phys. Soc. Japan* **2014**, *83*, 093705.

[227] Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

[228] Gao, W.; Fu, T.; Sun, J.; Coley, C. Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. Advances in Neural Information Processing Systems. 2022; pp 21342–21357.

[229] Jaeger, S.; Fulle, S.; Turk, S. Mol2Vec: Unsupervised Machine Learning Approach with Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27–35, PMID: 29268609.

[230] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

[231] Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model* **2012**, *52*, 2864–2875.

[232] Delaney, J. S. Esol: Estimating Aqueous Solubility Directly From Molecular Structure. *J. Chem. Inf. Comput.* **2004**, *44*, 1000–1005.

[233] Mobley, D. L.; Guthrie, J. P. Freesolv: A Database of Experimental and Calculated Hydration Free Energies, with Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711–720.

[234] Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

[235] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

[236] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

[237] Aids Antiviral, Screen Data Http://wiki.nci.nih.gov/display/ncidtpdata/aids+Antiviral+screen+data (Accessed 2024-01-03).

[238] Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936–1949.

[239] Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian Approach to in Silico Blood-Brain Barrier Penetration Modeling. *J. Chem. Inf. Model.* **2012**, *52*, 1686–1697.

[240] Huang, R.; Xia, M.; Nguyen, D.-T.; Zhao, T.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Rossoshek, A.; Simeonov, A. Tox21Challenge to Build Predictive Models of Nuclear Receptor and Stress Response Pathways as Mediated by Exposure to Environmental Chemicals and Drugs. *Front. Environ. Sci.* **2016**, *3*, 85.

[241] Tox21 Challenge, Http://tripod.nih.gov/tox21/challenge/ (Accessed 2024-01-03).

[242] Kuhn, M.; Letunic, I.; Jensen, L. J.; Bork, P. The Sider Database of Drugs and Side Effects. *Nucleic Acids Res.* **2016**, *44*, D1075–D1079.

[243] Gayvert, K. M.; Madhukar, N. S.; Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem. Biol.* **2016**, *23*, 1294–1301.

[244] Artemov, A. V.; Putin, E.; Vanhaelen, Q.; Aliper, A.; Ozerov, I. V.; Zhavoronkov, A. Integrated Deep Learned Transcriptomic and Structure-Based Predictor of Clinical Trials Outcomes. *BioRxiv* **2016**, 095653.

[245] Novick, P. A.; Ortiz, O. F.; Poelman, J.; Abdulhay, A. Y.; Pande, V. S. SWEETLEAD: An In Silico Database of Approved Drugs, Regulated Chemicals, and Herbal Isolates for Computer-Aided Drug Discovery. *PloS ONE* **2013**, *8*, e79568.

[246] Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, Inc., 2019.

[247] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

[248] Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2019**, *63*, 8749–8760.

[249] Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283–293.

[250] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-Chemical Insights From Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, 13890.

[251] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2015**, *44*, D1202–D1213.

[252] Irwin, J. J.; Shoichet, B. K. Zinc - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

[253] Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019**, *33*, 4602–4609.

[254] Vignac, C.; Loukas, A.; Frossard, P. Building Powerful and Equivariant Graph Neural Networks with Structural Message-Passing. *Adv. Neural Inf. Process Syst.* **2020**, *33*, 14143–14155.

[255] Gasteiger, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. International Conference on Learning Representations. 2020.

[256] Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; Tang, J. Pre-Training Molecular Graph Representation with 3D Geometry. International Conference on Learning Representations. 2022.

[257] Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.* **2022**, *4*, 127–134.

[258] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

[259] Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure–Property Relationships Through Smiles Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914–923.

[260] Maron, H.; Ben-Hamu, H.; Serviansky, H.; Lipman, Y. Provably Powerful Graph Networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

[261] Hearst, M.; Dumais, S.; Osuna, E.; Platt, J.; Scholkopf, B. Support Vector Machines. *IEEE Intell. Syst.* **1998**, *13*, 18–28.

[262] Lu, C.; Liu, Q.; Wang, C.; Huang, Z.; Lin, P.; He, L. Molecular Property Prediction: A Multilevel Quantum Interactions Modeling Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence* **2019**, *33*, 1052–1060.

[263] Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model* **2019**, *59*, 3370–3388.

[264] Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-Training Graph Neural Networks. International Conference on Learning Representations. 2020.

[265] Liu, S.; Demirel, M. F.; Liang, Y. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019.

[266] Beard, E. J.; Sivaraman, G.; Vázquez-Mayagoitia, Á.; Vishwanath, V.; Cole, J. M. Comparative Dataset of Experimental and Computational Attributes of UV/Vis Absorption Spectra. *Sci. Data* **2019**, *6*, 307.

[267] Venkatraman, V.; Kallidanthiyil Chellappan, L. An Open Access Data Set Highlighting Aggregation of Dyes on Metal Oxides. *Data* **2020**, *5*, 45.

[268] Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *J. Chem. Inf. Model* **2021**, *61*, 1053–1065.

[269] Joung, J. F.; Han, M.; Jeong, M.; Park, S. Experimental Database of Optical Properties of Organic Compounds. *Sci. Data* **2020**, *7*, 295.

[270] Venkatraman, V.; Raju, R.; Oikonomopoulos, S. P.; Alsberg, B. K. The Dye-Sensitized Solar Cell Database. *J. Cheminformatics* **2018**, *10*, 1–9.

[271] Greenman, K. P.; Green, W. H.; Gómez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci* **2022**, *13*, 1152–1162.

[272] Hirata, S.; Head-Gordon, M. Time-Dependent Density Functional Theory Within the Tamm-Dancoff Approximation. *Chem. Phys. Lett.* **1999**, *314*, 291–299.

[273] Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

[274] Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA Quantum Chemistry Program Package. *J. Chem. Phys.* **2020**, *152*, 224108.

[275] Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB — An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

[276] Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phy. Rev. A* **1988**, *38*, 3098.

[277] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[278] Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[279] Frisch, M. J. et al. Gaussian 09, Revision B.01. *Gaussian Inc., Wallingford CT* **2009**,

[280] Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide Toward Best Practices. *Chem. Mater.* **2020**, *32*, 4954–4965.

[281] Artrith, N.; Butler, K. T.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508.

[282] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016; pp 770–778.

[283] Hochreiter, S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertain. Fuzziness Knowlege-Based Syst.* **1998**, *6*, 107–116.

[284] Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy, 2010; pp 249–256.

[285] Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015; pp 448–456.

[286] Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *J. Am. Chem. Soc. Au* **2021**, *1*, 427–438.

[287] Lax, M. The Franck-condon Principle and Its Application to Crystals. *J. Chem. Phys.* **2004**, *20*, 1752–1760.

[288] Cederbaum, L. S.; Domcke, W. A Many-body Approach to the Vibrational Structure in Molecular Electronic Spectra. I. Theory. *J. Chem. Phys.* **2008**, *64*, 603–611.

[289] Davidson, E. R.; Jarzęcki, A. A. Zero Point Corrections to Vertical Excitation Energies. *Chem. Phys. Lett.* **1998**, *285*, 155–159.

[290] Liu, X.; Xu, Z.; Cole, J. M. Molecular Design of UV–Vis Absorption and Emission Properties in Organic Fluorophores: Toward Larger Bathochromic Shifts, Enhanced Molar Extinction Coefficients, and Greater Stokes Shifts. *J. Phys. Chem. C* **2013**, *117*, 16584–16595.

[291] Flanagan, P. J.; Cole, J. M. Clustering a Database of Optically Absorbing Organic Molecules via a Hierarchical Fingerprint Scheme That Categorizes Similar Functional Molecular Fragments. *J. Chem. Phys.* **2022**, *156*.

[292] Raissi, M.; Perdikaris, P.; Karniadakis, G. E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.* **2019**, *378*, 686–707.

Appendix A

Supporting Information: Chapter 3

A.1 Class-Probability Distributions

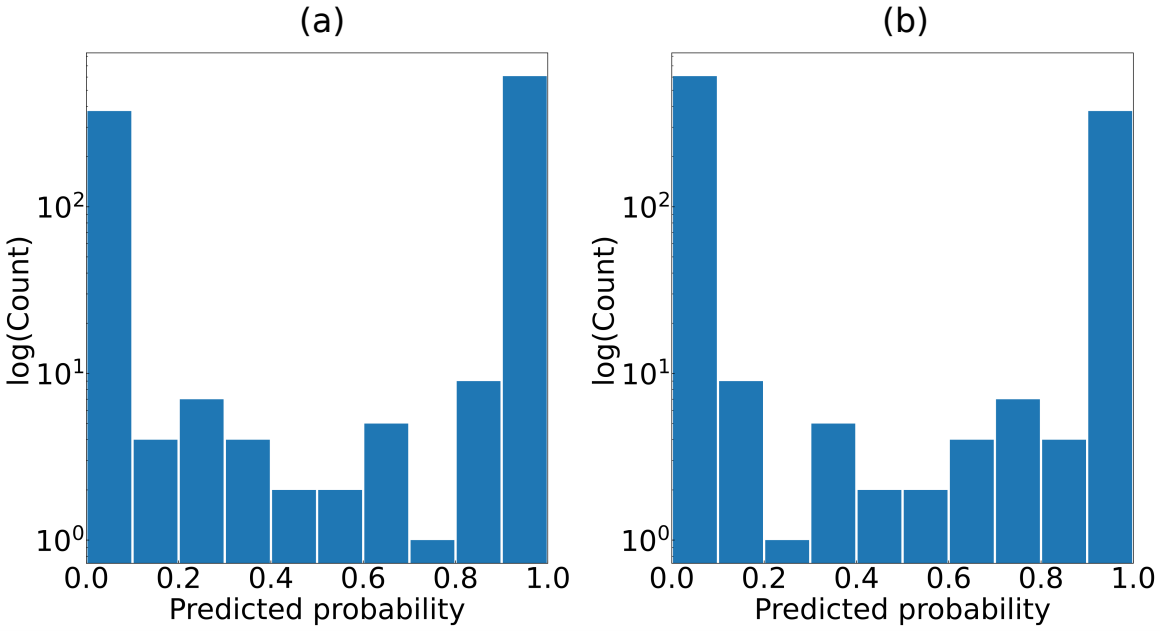


Fig. A.1 The predicted probability distribution of the final ML classifier for (a) non-metal and (b) metal in logarithmic scale. The output class-probabilities of the classifier are predominantly around zero and one, illustrating almost a binary outcome and demonstrates that the model is highly discriminative towards the target classes.

A.2 Predictions Against DFT Calculations

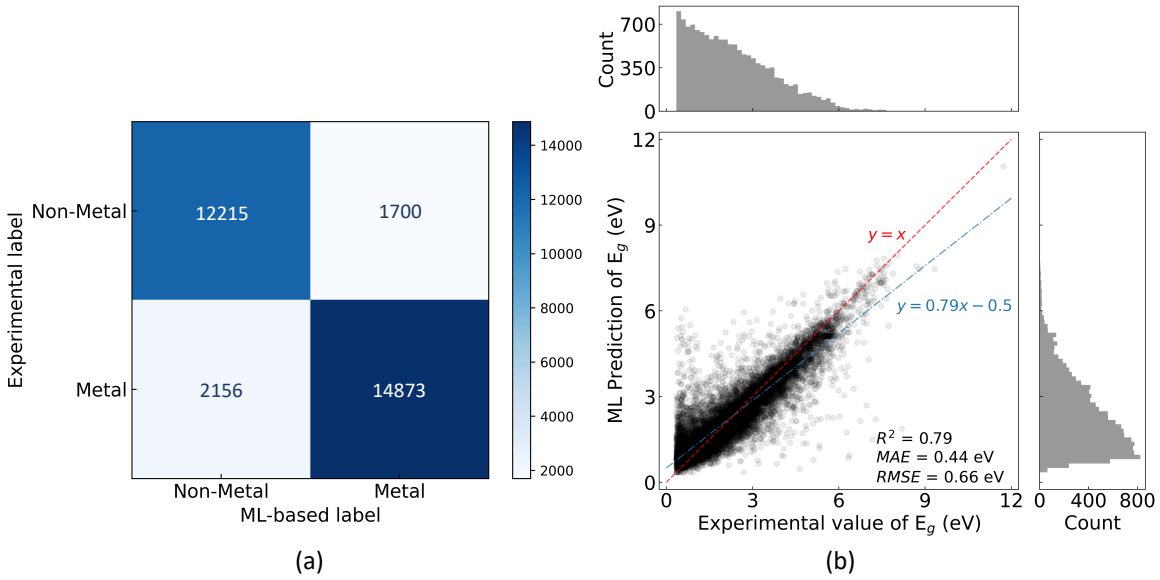


Fig. A.2 Bayesian-optimised model performance against DFT calculations (Materials Project) for the (a) classification of materials by metallicity and (b) regression analysis of E_g the test set, where the classification model was trained with 35 features and the regression model was trained on 44 features, both selected via the GBFS workflow. For the classification, a macro-weighted ROC-AUC of 0.949, AP of 0.962, F1-score of 0.874 and an balanced accuracy of 0.876 were achieved.

Appendix B

Supporting Information: Chapter 4

B.1 DFT-Aware Auxiliary Model Performance

Figure is presented on the following page.

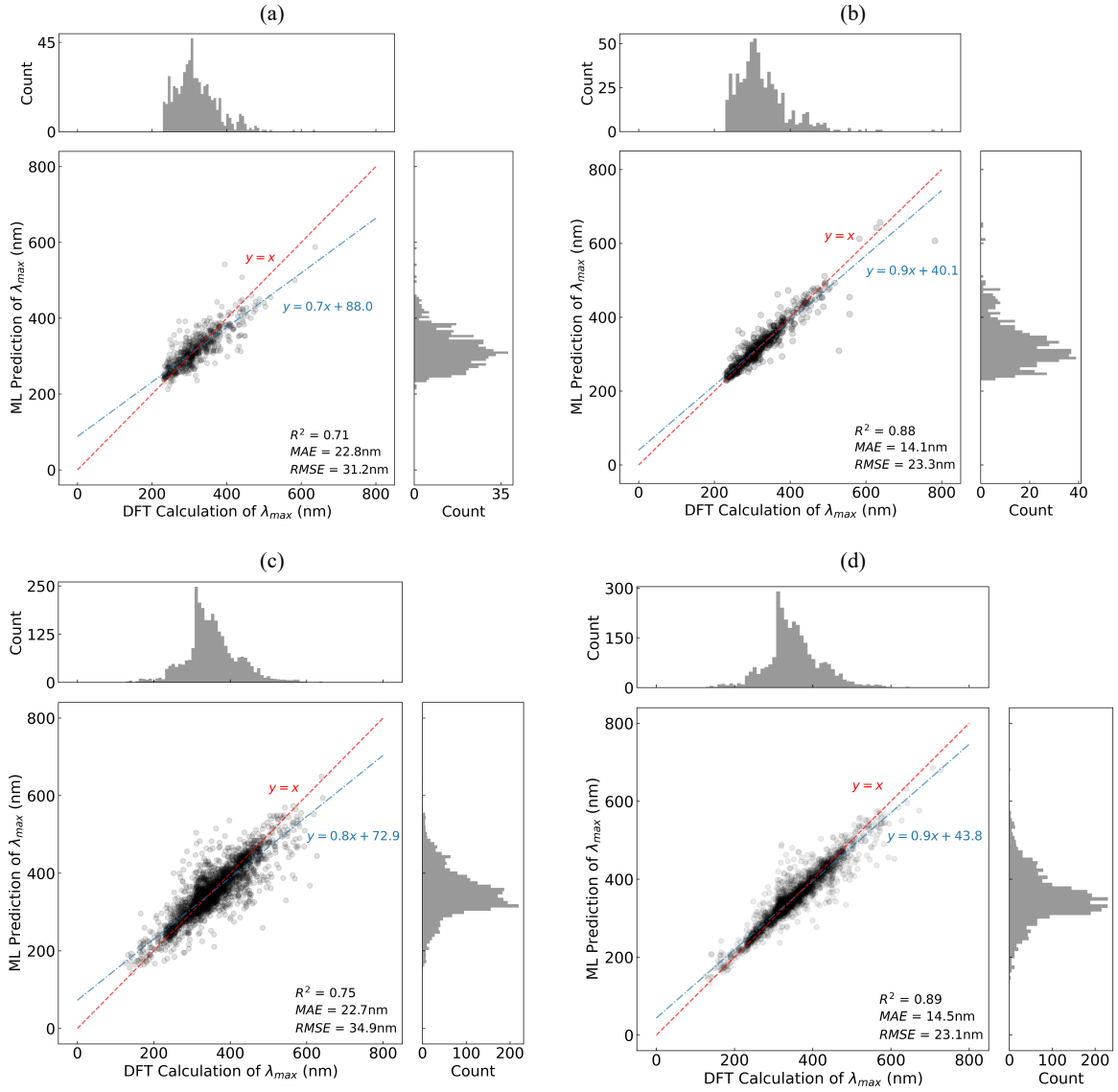


Fig. B.1 Regression analysis of the ML-based prediction of the vertical excitation energies, and their corresponding peak wavelengths (λ_{max}), against the DFT-based calculations. The predictions were made by Bayesian-optimised gradient-boosting models trained on the sTDA-DFT data set via (a) the DR-CNN and (b) the GBFS sub-workflows, and on the TD-DFT data set via (c) the the DR-CNN and (d) the GBFS sub-workflows. The solid blue line is a linear fit between the DFT and ML-based predictions that was generated using ordinary least squares refinement. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the DFT-based calculations.

B.2 Experimental-Aware ML Model Performance

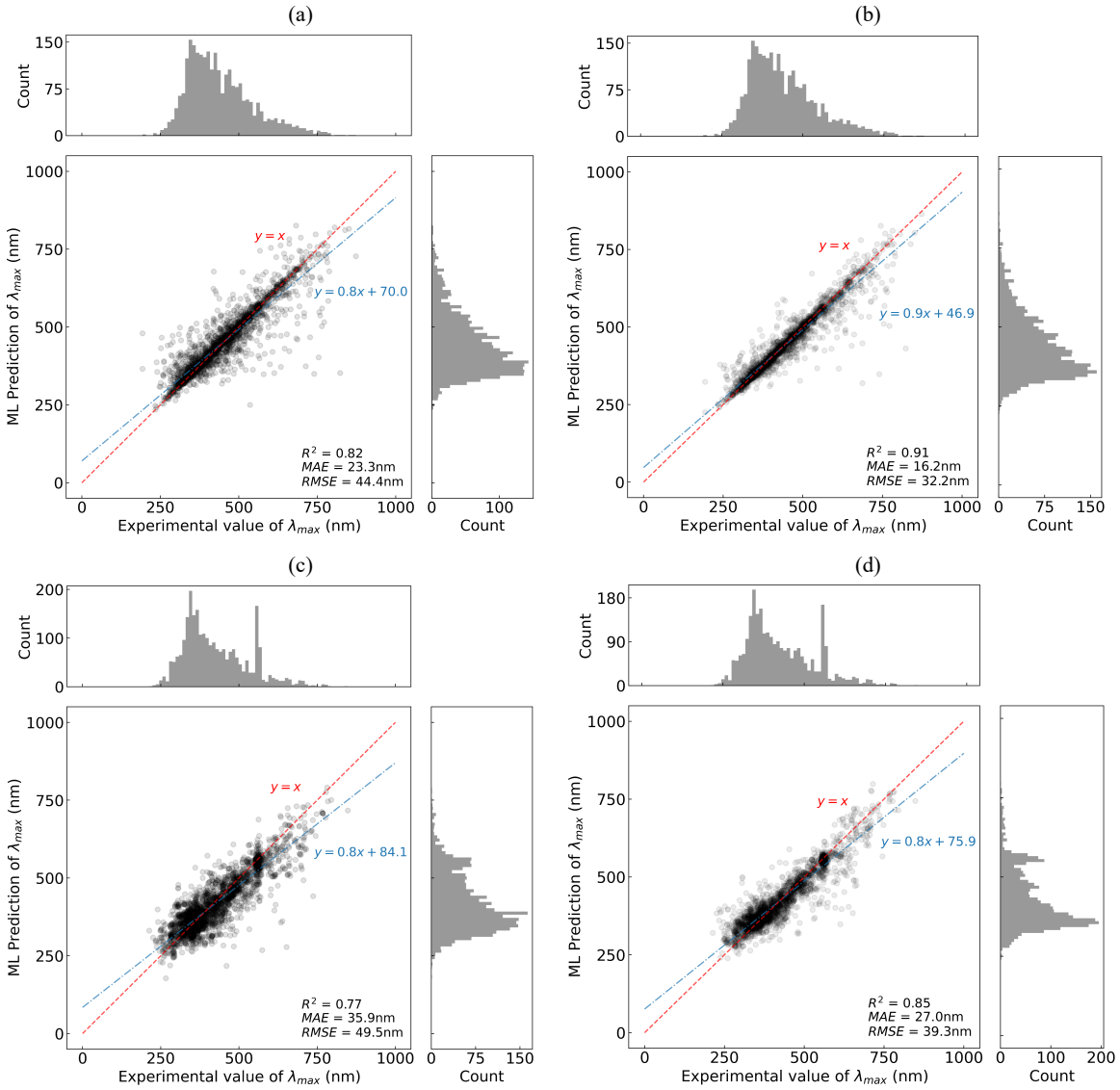


Fig. B.2 Regression analysis of the ML-based prediction of the experimental optical absorption peak (λ_{max}) of unseen chemical dyes against their experimental measurements. The predictions were made by Bayesian-optimised gradient-boosting models that were trained using the random splitting strategy via (a) the DR-CNN and (b) the GBFS sub-workflows, and using the scaffold splitting strategy via (c) the DR-CNN and (d) the GBFS sub-workflows. The solid blue line is a linear fit between the experimental measurements and ML-based predictions that was generated using ordinary least squares refinement. The dashed red line represents the hypothetical case, where the ML-based prediction would equal the experimental measurements.

B.3 Error Distribution by Type of Solvents

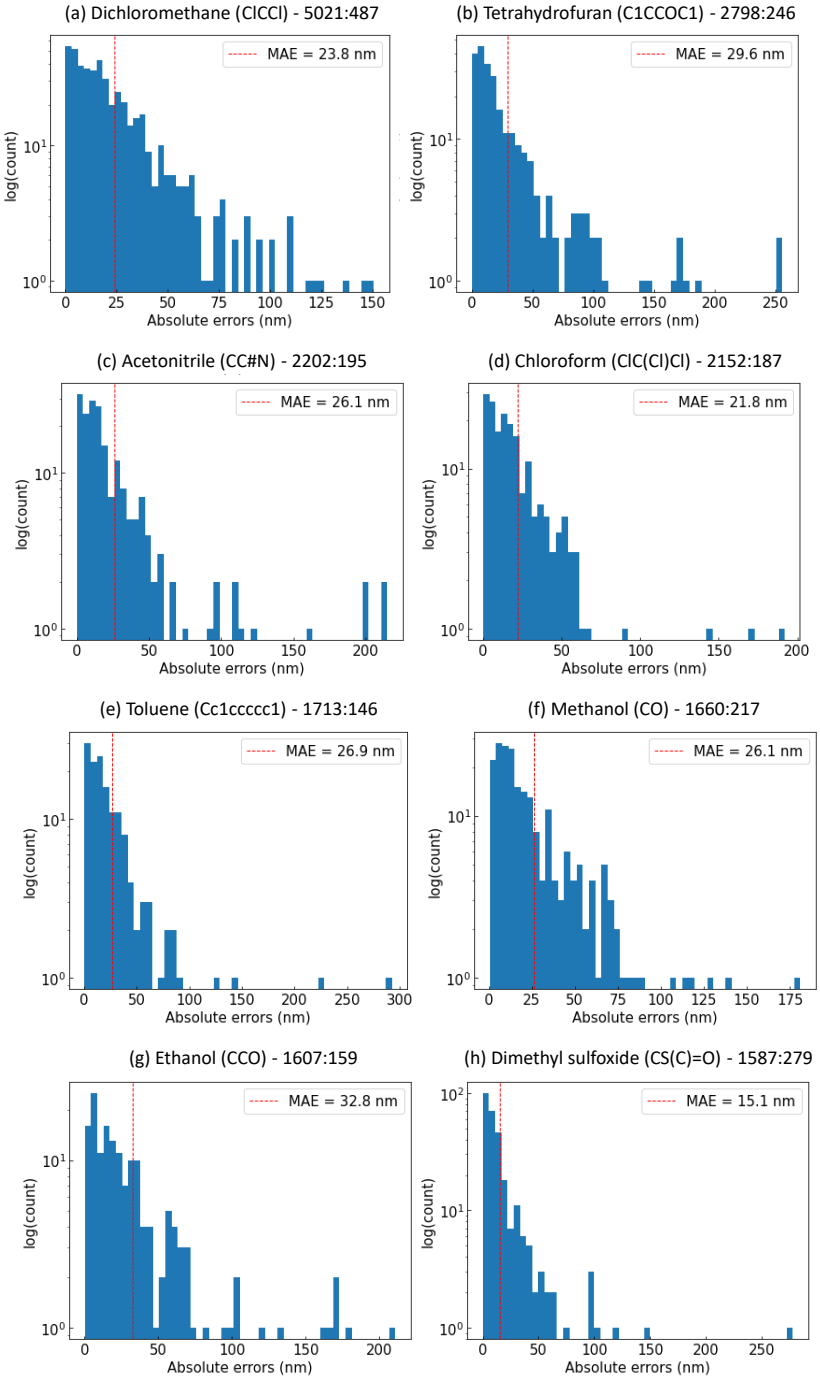


Fig. B.3 Distributions of the absolute errors of the ML-based prediction of λ_{max} against the experimental measurements, partitioned by solvent types. The ratio signifies the occurrence frequency of the solvent in the training and test sets with scaffold splitting, respectively.