UNIVERSITY OF CALIFORNIA SAN DIEGO

Domain-Knowledge-Guided Machine Learning Towards Accurate Materials Property Prediction and Materials Discovery

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Chemistry

by

Weike Ye

Committee in charge:

Professor Shyue Ping Ong, Chair
Professor Francesco Paesani, Co-Chair
Professor William Trogler
Professor Wei Xiong
Professor Kesong Yang

2021

The dissertation of Weike Ye is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my beloved family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is a genuine pleasure to express my deep gratitude to my advisor, Prof.Shyue Ping Ong, for his guidance and support throughout my graduate studies. His timely advice, meticulous scrutiny, scholarly advice, and scientific vision have helped me tremendously to accomplish my work. His keen interest in science has inspired me, and his high standards for scientific rigor have prepared me for my future career. I consider myself fortunate to have had the opportunity to learn so much from him.

I must also pay tribute to the high quality of mentor-ship given by Dr. Chi Chen. There are numerous times when I felt discouraged and unsure of the intrinsic worth of my work, and it was he who offered precious advice and encouragement.

I am grateful to my collaborators in our MAVRL group, Dr. Zhenbin Wang, Dr. Iek-Heng Chu, Mahdi Amachraa, Yunxing Zuo, and Hui Zheng, for lending your expertise and intuition to my scientific and technical problems. I also owe my gratitude to our experimental collaborators, including Professor Joanna McKittrick and Professor Olivia Graeve from the University of California San Diego, and Professor Jakoah Brgoch from the University of Houston. Thanks for having faith in our predictions and offering to provide experimental verification.

My sincere gratitude also goes to the team MAVRL. It is my pleasure to have worked with such a talented group.

There is no way to express how much it meant to me to have been supported by my beloved family and friends. It is the unconditional love from my parents that raised me up again and again in this long and uneasy journey. I am forever indebted to them for giving me the opportunities and experiences that have made me who I am. I would also like to give special thanks to Dr. Fengqin Lian. Thank you for being there through my ups and downs.

Chapter 2, in full, is a reprint of the material ”Deep neural networks for accurate predictions of crystal stability” as it appears on Nature Communications, Weike Ye, Chi Chen, Zhenbin Wang, Iek-Heng Chu, Shyue Ping Ong, 2018, 9 (1), 1-6. The dissertation author was the primary

investigator and author of this paper.

Chapter 3, in full, is under preparation for publication of the material ”High-throughput screening of Eu^{2+}-doped red-emission garnet phosphors using density functional theory and machine learning”, Weike Ye, Chi Chen, Mahdi Amachraa, Yunxing Zuo, Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is under preparation for publication of the material ”A universal machine learning model for elemental grain boundary energies”, Weike Ye, Hui Zheng, Chi Chen, Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

VITA

2014	B. S. in Chemistry, Nanjing University
2016	M. S. in Chemistry, University of California San Diego
2021	Ph. D. in Chemistry, University of California San Diego

PUBLICATIONS

1. **Weike Ye**, Chi Chen, Zhenbin Wang, Iek-Heng Chu, Shyue Ping Ong "Deep neural networks for accurate predictions of crystal stability", Nature Communications, 2018, 9 (1), 1-6.
2. **Weike Ye**, Chi Chen, Mahdi Amachraa, Yunxing Zuo, Shyue Ping Ong "High-throughput screening of Eu^{2+}-doped red-emission garnet phosphors using density functional theory and machine learning", under preparation.
3. **Weike Ye**, Hui Zheng, Chi Chen, Shyue Ping Ong "A universal machine learning model for elemental grain boundary energies", under preparation.

ABSTRACT OF THE DISSERTATION

Domain-Knowledge-Guided Machine Learning Towards Accurate Materials Property Prediction and Materials Discovery

by

Weike Ye

Doctor of Philosophy in Chemistry

University of California San Diego, 2021

Professor Shyue Ping Ong, Chair
Professor Francesco Paesani, Co-Chair

In the past few decades, the first principles modeling algorithms, especially density functional theory (DFT), have been important complements to experiments in studying properties and materials design. Thanks to the success of DFT and the fast development of computational capabilities, we have witnessed the exploration of a huge amount of materials data. The logical next step is the introduction of tools capable of making use of the generated data. Machine learning (ML) techniques are such tools to extract knowledge from data and make predictions at a sub-second speed, which are currently steering materials science into a new data-driven paradigm.

In this thesis, following the close guidance of domain knowledge in materials science, we strive to develop accurate, interpretable ML models that could potentially serve as the surrogate of DFT in property prediction and the design of new materials. A unifying theme that differentiates the models in this thesis from their counterparts in other existing ML works is the practice of the principle of parsimony, where we aspire to develop and explain the models with minimum features.

The thesis is divided into three topics. In the first topic (Chapter 2), we aimed at predicting the phase stability of the inorganic crystals, which is often the first step in any materials discovery. Inspired by Pauling's rules, we show that deep neural networks utilizing just the Pauling electronegativity and ionic radii of the species of the symmetrically distinct sites can predict the DFT formation energies of garnets and perovskites within the low mean absolute errors (MAEs) of 7-34 meV atom^{-1}. The models can be easily extended to mixed garnets and perovskites with little loss in accuracy by using a binary encoding scheme, extending the applicability of ML models to the infinite universe of mixed-species crystals.

In the second topic (Chapter 3), we targeted predicting the bandgap. By machine learning on 1823 data, we show that the eXtreme gradient boosting(XGBoost) model reaches the state-of-the-art MAE of 0.13 eV at predicting the DFT bandgap (using generalized gradient approximation functional) of garnets. Interpreting the model's behavior reveals that the bandgap is affected mainly by the atomic number of the species occupying the tetrahedron sites in a garnet crystal. Integrating the models from both Chapter 2 and Chapter 3, we devised a high-throughput screening (HTS) workflow to screen for Eu^{2+}-doped red emission phosphors in the garnet crystal family. Two candidates, $Ca(Er,Tb)_2Mg_2Si_3O_{12}$, were identified by rapidly transversing 5554 candidate compositions, which is computationally prohibitive for pure DFT-based HTS workflows due to the large cell size of the garnet structures.

In the last topic (Chapter 4), we investigated the 2D defect, grain boundary (GB), in polycrystalline systems. We show that the energy of a grain boundary, normalized by the bulk

cohesive energy, can be described purely by four geometric features. By machine learning on a large computed database of 369 low-Σ ($\Sigma < 10$) GBs of more than 50 metals, we developed a model that can predict the grain boundary energies to within 0.12 J m^{-2}. This universal GB energy model can be extrapolated to the energies of higher sigma GBs with a modest increase in prediction error.

Chapter 1

Introduction

1.1 Background

The prediction of materials properties and the discovery of new materials are among the most important subjects in materials science. For the past decades, the growing computational resources and the well-established quantum mechanical approximations to the Schrödinger's equation, in particular density functional theory (DFT)[2,3], have enabled the researchers to predict the physical and chemical properties of materials and virtually guide the experimental efforts. The algorithm development in electronic structure codes such as DFT and the computing capabilities have advanced to the degree that first-principles calculations can be performed in a high-throughput fashion. High-throughput DFT calculations have greatly accelerated the discovery of numerous materials such as alkali-ion batteries[4–6], catalysts[7], organic semiconductors[8], and phosphors[9,10]. It also fueled the development of such large, high-quality open databases of computed materials as the Materials Project[11], Open Quantum Materials Database[12], the AFLOW repository[13], etc[14–17].

However, despite the advances in theoretical methodologies, DFT is known for its poor scalability and high cost. On the one hand, there is a finite limit on the system size of ~ 1000 atoms because the scaling of DFT to the number of electrons is typically $O(n_e^3)$ or higher[18–20]. On the other hand, when the number of candidates reaches a medium level of thousands, the high computational cost of DFT calculations becomes the bottleneck in high-throughput screening (HTS) workflows.

Machine learning (ML) is the branch of artificial intelligence that focuses on developing algorithms to extract patterns from data. Important advances of ML have been made across a variety of tasks such as playing the Go[21], natural language processing[22], autonomous driving[23] and etc. The growing accessibility of the large number of high-quality data in materials science has nourished the application of ML to make rapid property predictions in the vast unexplored structure space without performing first-principles calculations. The accuracy and efficiency of

ML models make them promising solution to the scaling problem embedded in DFT.

There are three key steps of any ML tasks: (1) collecting data with sufficient quantity and quality, and the curation of the data, (2) design of the task, including the scheme to map the input data to a numerical representation (descriptor/feature) that is relevant to the target, and the choice of proper target, and (3) the fitting of the model. In this chapter, we first discuss in details on these steps for ML applications in material science. Then, we review the current ML applications in materials' property prediction. Lastly, we conclude with objectives and an overview of this thesis.

1.2 Machine learning model development

1.2.1 Data

Obtaining large and diverse data sets is the prerequisite for developing rational ML models, and it has been one of the major limitations to applying ML in materials science. Quantity-wise, 50 data points are often considered the lower limit to build a descent ML model. Quality-wise, the data should display a good coverage of both the chemistry space and the property space, and maintain consistency. In principle, fit-to-experiment predictions are more exciting but existing experimental data repositories[24–26] are still limited by the scarcity of property data and suffer from data inconsistency as a result of uncontrolled experimental conditions. Therefore, using computed data sources is still the more prevailing choice. Take the Materials Project as an example, it currently hosts $\sim$ 133000 crystals structures with properties such as DFT-relaxed energies and bandgaps available, which breeds an extensive amount of high-impact ML works in this field[27–29]. In addition to using the ready data from databases, generating data from scratch sometimes is necessary. For one, it should be noted that most large computed materials databases are still constructed using the Perdew-Berke-Ernzerhof (PBE)[2] generalized gradient approximation (GGA) functional, which is efficient in computing but could fail for systems with strong electron correlation and van der waals interactions[30–32]. For another, general databases could lose resolution in more constrained chemistry spaces. The data generation is often carried out via high-throughput ab initio calculations, facilitated by open-source materials analysis and HT workflow management softwares such as Pymatgen[33], Fireworks[34], Atomate scientific workflow packages[35], and etc[13,14].

1.2.2 Task definition

To define an ML task, there are two key components, i.e., the mapping of the input data to a numerical representation (descriptor/feature), and the choice of a learnable metric for the target. The choice of descriptors is critical for the model performance. Basic requirements for descriptors are informative and discriminating. Being informative requires the descriptors to reflect the underlying physics behind the predicting target, whereas being discriminating challenges the descriptors to have sufficient distance for instances that have small statistical distance. In materials science, descriptors are typically two types, i.e., compositional and structural. Compositional descriptors are numerical values that represent physical aspects of the constituent elements such as the atomic number, electronegativity, atomic radii, electronic structure, etc. These descriptors have been shown to have reasonably good performance for predicting as varied materials properties as thermoelectric figures of merit[36], thermal conductivity[37], solute diffusion barriers in face-centered-cubic metals[38], elastic properties[39], glass-forming ability[40], and bandgaps[40]. While compositions-based descriptors are usually highly informative, the limitation is obvious as they are intrinsically unable to distinguish between polymorphs. For most problems, a feature set that describes the full materials' structure is desired. Graph-based representation[41] of crystals and molecules has gained substantial interest in recent years[27,29,42]. Neural networks based on such representation (GNN) have achieved state-of-the-art performances in predicting the formation energies, bandgaps, and other common materials properties[27,29]. However, we should be aware that training of such models requires a large number of data; hence only a limited number of properties can afford the training of such delicate models[43]. Furthermore, current available GNNs are trained on general-purpose databases like Materials Project. There is no guarantee that the performance of such models is also optimal in more contained chemistry or structure spaces.

In additional to the basic requirements, the compactness of the feature set is also critical to the performance and generalizability of the ML model. The selection of features can be knowledge-driven or data-driven. The former relies on applying physical and chemical intuition to select appropriate features for the ML problem. The knowledge-driven approach often leads to more efficient features and thus more interpretable models. However, there is no guarantee of the optimal performance. On the other hand, the data-driven approach starts from a large initial set of candidate features and down-selects an optimal subset. There are numerous available statistical tools to automate this down-selection process, such as using L_0 or L_1 regularization (least absolute shrinkage and selection operator, LASSO)[44–46], feature importance[47,48], principal component analysis (PCA)[42,49,50] and etc. Features chosen by this approach can often achieve global-optimal performance while accompanied by possible sacrifice in the interpretability.

The definition or engineering of the target is arguably the most important but often underestimated step. The target should have clear-defined uncertainties and errors and ideally display a normal-like distribution. Choosing the wrong target could be detrimental to the model performance and generalizability. For example, phase stability is one of the central problems in materials science. The common metrics to measure the stability are the 0 K DFT formation energy E_f or the energy above convex hull E_{hull}[51]. The latter is a much more difficult target than the former since the errors of the E_{hull} are inconsistent across chemical spaces, and there is a lower bound at zero of E_{hull} by definition.

1.2.3 Model fitting

Eventually, we enter the last step of model fitting. The machine learning algorithms are normally categorized into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is by far the most common for ML in material science, where the data is structured as composition/structure-property pair. There are numerous regression algorithms from linear regression to graph networks in the ascending order of complexity. They can be easily constructed and tuned for optimized performance with the aid from open-source ML software libraries such as scikit-learn[52], Tensorflow[53], and Pytorch[54].

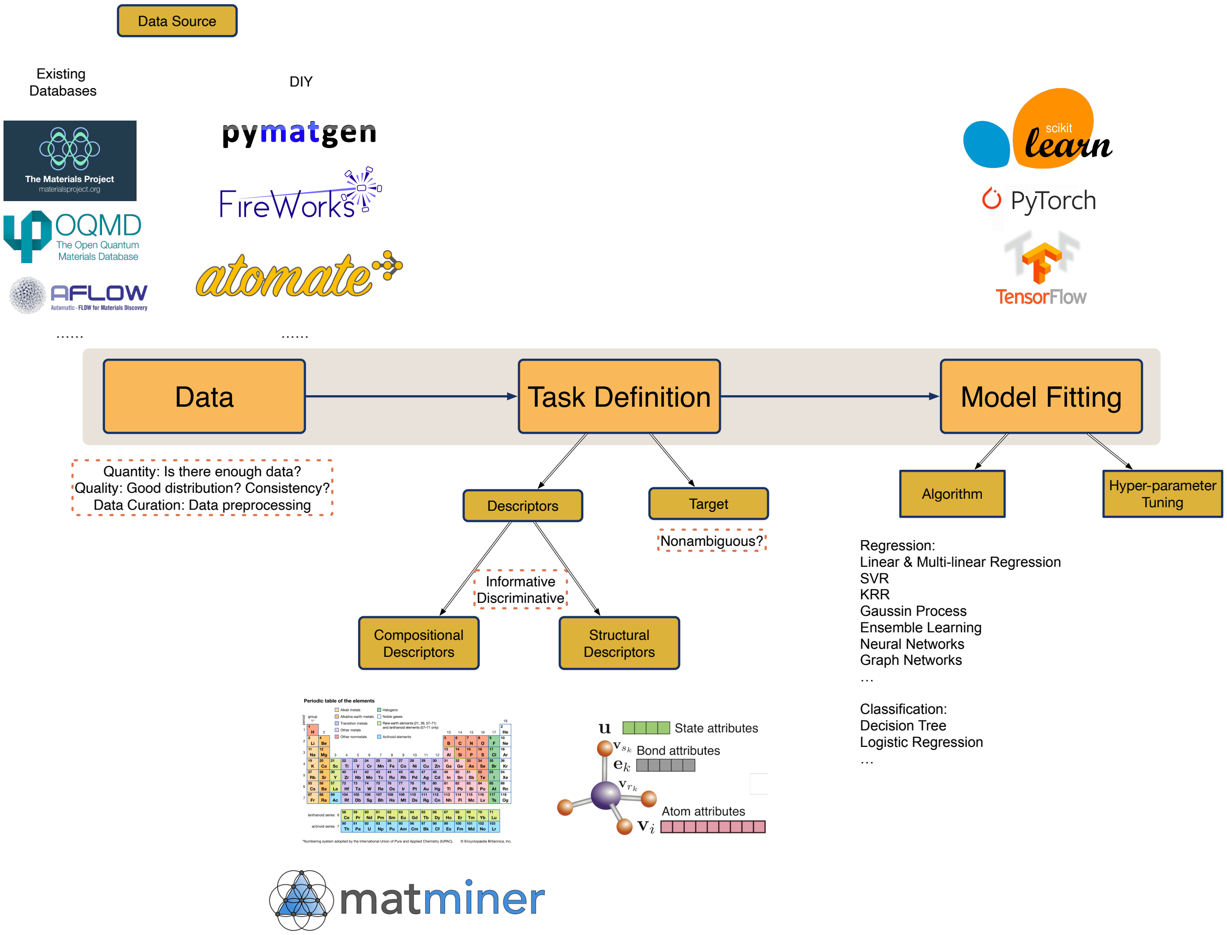


Figure 1.1: Three key steps for constructing machine learning models, starting from collecting enough high quality data, to defining the task by coming up with the descriptor scheme and selecting the learnable metric of the target, and eventually fitting the model.

1.3 Current application of machine learning in materials' property prediction

Property prediction is one of the significant applications of ML in materials science. In this section, we summarize existing works categorized by the properties predicted. Phase stability is a property of ubiquitous interest in materials science. ML works that predict phase stability can be loosely categorized based on the applicable scope, i.e., the general models and structure-type-specific models. The general models that are based on compositional-based features typically have the higher error between 50-88 meV atom^{-1}[40,55], whereas their counterparts that are based on graph-neural-networks have the state-of-the-art accuracy within the error of only 28 meV atom^{-1}[27]. For structure-specific models, there are multiple ML works that predicts the phase stability of the perovskites[48,56,57] and Heusler compounds[58–60]. The typical mean absolute error of these works is at the level of 21-121 meV atom^{-1}.

The bandgap is another important material property commonly estimated via first-principles calculations. Similar to the phase stability, there are general models[27,29,40,61,62] and more specific models[48,63–65]. The typical mean absolute error of GGA band gap for non-metal crystals is $\sim$ 0.24 eV for general models and $\sim$ 0.2 eV for specific models. It should be noted that the majority of the works are based on GGA funcitonal which is known to underestimate the bandgap due to the approximation in exchange-correlation functionals, the self-interaction error, and the missing derivative discontinuity. There are more advanced but expensive algorithms that provide more close-to-experiment bandgaps, such as the modified Becke–Johnson (mBJ) functional[66], the delta self-consistent-field (ΔSCF) method[67], hybrid functionals (HSE06)[68], and GW calculations based on many body perturbation theory[69]. Lee et al. developed a support vector regression (SVR) model to predict the G_0W_0 bandgap, of which the root-means-squared error is 0.24 eV[28].

In addition to bulk crystal properties, modeling of more complex defect systems are possible. Grain boundary (GB), the interface between two grains in a polycrystalline material, is a 2D defect in the crystal structure. The energy of GB strongly affects polycrystalline materials' mechanical properties such as strength, toughness, and corrosion resistance[70,71]. Multiple works have developed ML models for such restricted chemistry and structure types as face-centered cubic (fcc) Cu[72], fcc Ni[73,74], or fcc Al systems[75] with the typical mean absolute error at the level of below 0.1 J m^{-2}. Note that all of the ML works are based on data calculated from embedded atom model(EAM) potentials since the more accurate and general ab initio database of GB energy is only available recently[76].

1.4 Objectives and overview

Despite the increasingly important role ML plays in property prediction, there is ample room for improvements on multiple fronts. To begin with, the infusion of domain knowledge in feature and target engineering can lead to more efficient and interpretable models. The feature selection process can be a hybrid of both knowledge-driven and data-driven to ensure both interpretability and optimal performance. Furthermore, carefully designed targets under the guidance of domain knowledge offer advantages in the model's predictive ability. Secondly, property prediction often the times is not the ultimate goal in materials science, but rather a intermediate step towards the discovery of novel materials with desired properties. The integration of ML models into HTS workflow to surrogate pure-DFT counterparts enables more efficient screening.

In this thesis, we showcase that under the guidance of domain-knowledge, a series of high accuracy, interpretable ML models are developed and are integrated into HTS workflows to accelerate the discover of promising phosphor materials. The thesis can be divided into three topics. In the first topic, we demonstrate the neural networks that can accurately predict the phase stability of bulk crystal. In developing the phase stability model, the intuitive chemical hypothesis that the ionic crystal stability should be quantitatively related to the electronegtivity and ionic radii of the species occupying symmetrically distinctive sites guided our choice of features. The knowledge of the well-known limitations of DFT calculations in handling redox reaction energies[77] directed us to choose the formation energy from the binary oxides as the appropriate target instead of more widely used formation energy from the elements and the energy above the hull. Starting from the carefully designed features and target, we developed neural network models that are able to predict the DFT formation energies of garnets and perovskites to within 7-34 meV atom^{-1}, and extended the models to mixed compositions with little loss in accuracy.

In the second topic, We continued to develop an eXtreme gradient boosting (XGBoost) model to predict the GGA bandgap of garnets to within the error of 0.13 eV, a substantial improvement compared to a common mean absolute error of 0.2 eV for structure-specific ML models. The feature selection was performed utilizing both the domain knowledge and data techniques, where starting from elemental attributes related to crystal electronic structures, we down-selected the optimal feature set by evaluating the whole feature space. The highly interpretable model reveals that the atomic number of the species occupying the tetrahedron sites of the garnets has the most strong negative correlation with the bandgap of the garnets. We further integrated both models for predicting the phase stability and the GGA bandgap of garnets into an ML-DFT hybrid workflow to screen for the Eu^{2+}-doped red-emission phosphors. Two candidates ($Ca(Er,Tb)_2Mg_2Si_3O_{12}$) were identified from more than 5000 candidates compositions, the screening of which is computational prohibitive by pure DFT-based workflow.

In the last topic, we considered more complex structures of 2D defects in polycrystalline metals. By normalizing the grain boundary energy over the bulk cohesive energy, we show that a universal and extrapolatable model can predict the grain energies to within 0.12 J m^{-2} by machine learning on 369 low-sigma GBs of more than 50 metals using only four pure geometric features.

A brief description for each subsequent chapter is provided as follows:

- Chapter 2 presents the development of deep neural networks utilizing just two attributes—the

Pauling electronegativity and ionic radii to predict the DFT formation energies of $C_3A_2D_3O_{12}$ garnets and ABO_3 perovskites within the error of 7–34 meV atom^{-1}, well within the limits of DFT accuracy. A further extension to mixed garnets and perovskites with little loss in accuracy can be achieved using a binary encoding scheme, addressing a critical gap in the extension of machine-learning models to the vast combinatorial chemical spaces. Finally, we demonstrate that the potential of these models to rapidly transverse vast chemical spaces to accurately identify stable compositions, accelerating the discovery of novel materials with potentially superior properties.

- Chapter 3 presents a study on developing an ML model that predicts the PBE bandgap of garnets to within the error of 0.13 eV using only six features per structure. We integrated the models from Chapter 2 and this work into an HTS workflow to screen for Eu^{2+}-doped red-emission phosphor. Two superior candidates, $Ca(Er,Tb)_2Mg_2Si_3O_{12}$ were identified from more than 5000 compositions.

- Chapter 4 presents a study showcasing that the energy of a grain boundary, normalized by the bulk cohesive energy, can be described purely by four geometric features. By machine learning on a large computed database of 369 low-sigma (sigma $<$ 10) GBs of more than 50 metals, we developed an interpretable and extrapolatable model that can predict the grain energies within 0.12 J m^{-2}.

Chapter 2

Deep neural networks for accurate predictions of crystal stability

2.1 Introduction

The formation energy of a crystal is a key metric of its stability and synthesizability. It is typically defined relative to constituent unary/binary phases (E_f) or the stable linear combination of competing phases in the phase diagram (E_{hull}, or energy above convex hull)[51]. In recent years, machine learning (ML) models trained on DFT[2] calculations have garnered widespread interest as a means to scale quantitative predictions of materials properties[28,57,64,78,79], including energies of crystals. However, most previous efforts at predicting E_f or E_{hull} of crystals[40,57,80–83] using ML models have yielded mean absolute errors (MAEs) of 70-100 meV atom^{-1}, falling far short of the necessary accuracy for useful crystal stability predictions. This is because approximately 90% of the crystals in the Inorganic Crystal Structure Database (ICSD) have $E_{hull} < 70$ meV atom^{-1} [84], and the errors of DFT-calculated formation energies of ternary oxides from binary oxides relative to experiments are ~ 24 meV atom^{-1} [85].

We propose to approach the crystal stability prediction problem by using artificial neural networks (ANNs)[86], i.e., algorithms that are loosely modeled on the animal brain, to quantify well-established chemical intuition. The Pauling electronegativity and ionic radii guide much of our understanding about the bonding and stability of crystals today, for example, in the form of Pauling's five rules[87] and the Goldschmidt tolerance factor for perovskites[88]. Though these rules are qualitative in nature, their great success points to the potential existence of a direct relationship between crystal stability and these descriptors.

To probe these relationships, we choose, as our initial model system, the garnets, a large family of crystals with widespread technological applications such as luminescent materials for solid-state lighting[89] and lithium superionic conductors for rechargeable lithium-ion batteries[90,91]. Garnets have the general formula $C_3A_2D_3O_{12}$, where C, A and D denote the three cation sites with Wyckoff symbols 24*c* (dodecahedron), 16*a* (octahedron) and 24*d* (tetrahedron), respectively, in the prototypical cubic $Ia\bar{3}d$ garnet crystal shown in Fig. 2.1a. The distinct coordination environments of the three sites result in different minimum ionic radii ratios (and hence, species preference) according to Pauling's first rule. We further demonstrate the generalizability of our approach to the ABO_3 perovskites (Fig. 2.1b), another broad class of technologically important crystals[92–96].

In this work, we show that ANNs using only the Pauling electronegativity[97] and ionic radii[98] of the constituent species as the input descriptors can achieve extremely low MAEs of 7–10 meV atom^{-1} and 20-34 meV atom^{-1} in predicting the formation energies of garnets and perovskites, respectively. We also introduce two alternative approaches to extend such ANN models beyond simple unmixed crystals to the much larger universe of mixed cation crystals – a rigorously defined averaging scheme for the electronegativity and ionic radii for modeling complete cation disorder, and a novel binary encoding scheme to account for the effect of cation orderings with minimal increase in feature dimension. Finally, we demonstrate the application of the NN models in accurately and efficiently identifying stable compositions out of thousands of garnet and perovskite candidates, greatly expanding the space for the discovery of materials with potentially superior properties.

2.2 Results

2.2.1 Model construction and definitions

We start with the hypothesis that the formation energy E_f of a $C_3A_2D_3O_{12}$ garnet is some unknown function f of the Pauling electronegativities (χ) and Shannon ionic radii (r) of the species in the C, A and D sites, i.e.,

$$E_f = f(\chi_C, r_C, \chi_A, r_A, \chi_D, r_D) \tag{2.1}$$

Here, we define E_f as the change in energy in forming the garnet from binary oxides with

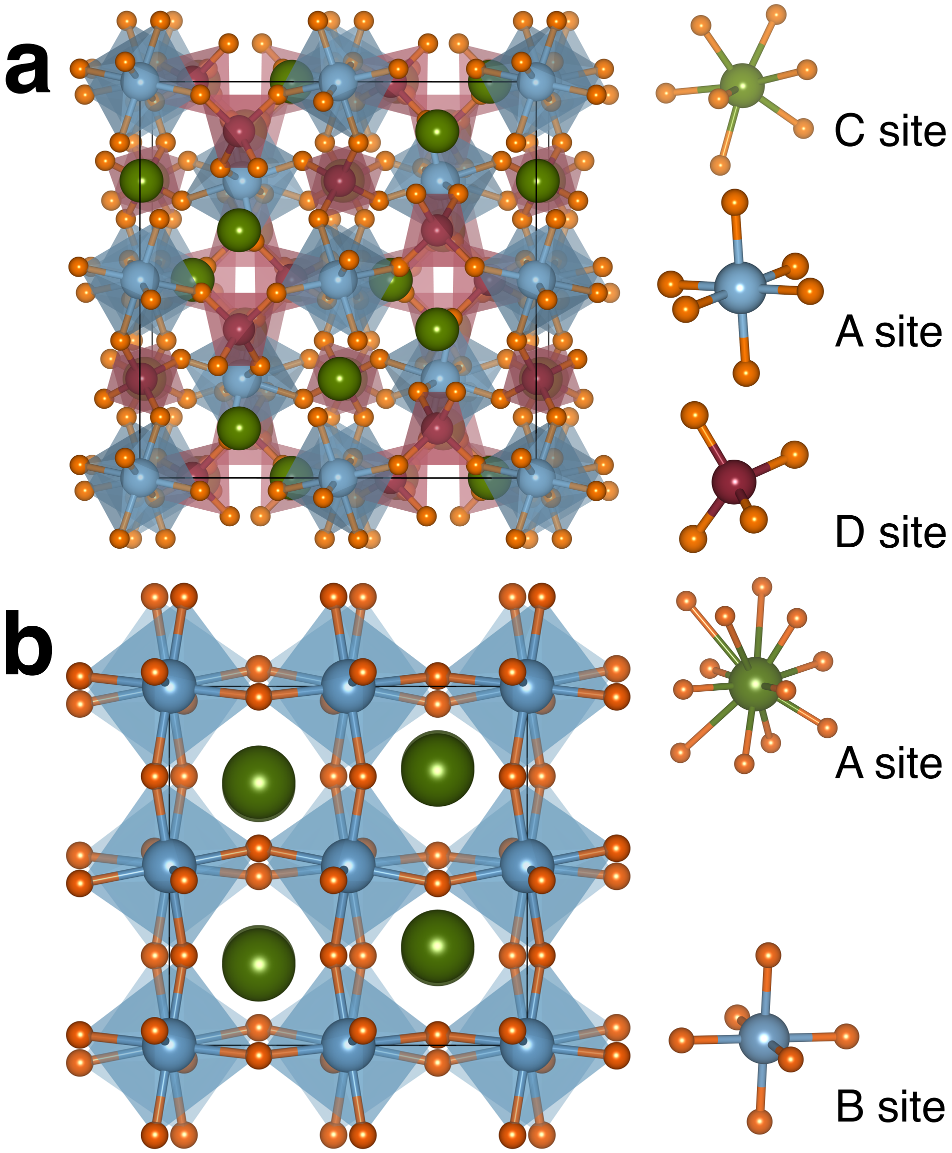


Figure 2.1: Crystal structures of garnet and perovskite prototypes. a. Crystal structure of $Ia\bar{3}d$ $C_3A_2D_3O_{12}$ garnet prototype. Green (C), blue (A) and red (D) spheres are atoms in the 24*c* (dodecahedron), 16*a* (octahedron) and 24*d* (tetrahedron) sites, respectively. The orange spheres are oxygen atoms. b. Crystal structure of *Pnma* ABO_3 perovskite prototype. Green (A) and blue (B) spheres are atoms in the 4*c* (cuboctahedron) and 4*d* (octahedron) sites, respectively. The orange spheres are oxygen atoms.

elements in the same oxidation states, i.e., E_f^{oxide} as opposed to the more commonly used formation energy from the elements $E_f^{element}$ used in previous works[40,80–82]. Using the $Ca_3Al_2Si_3O_{12}$ garnet (grossular) as an example, E_f^{oxide} is given by the energy of the reaction: $3\,CaO + Al_2O_3 + 3\,SiO_2 \longrightarrow Ca_3Al_2Si_3O_{12}$. This choice of definition of E_f is motivated by two reasons. First, binary oxides are frequently used as synthesis precursors. Second, our definition ensures that garnets that share elements in the same oxidation states have E_f that are referenced to the same binary oxides, minimizing well-known DFT errors. In contrast, $E_f^{element}$ and E_{hull} are both poor target metrics for a ML model. $E_f^{element}$ suffers from non-systematic DFT errors associated with the incomplete cancellation of the self-interaction error in redox reactions[77], while E_{hull} is defined with respect to the linear combination of stable phases at the $C_3A_2D_3O_{12}$ composition in the C-A-D-O phase diagram, which can vary unpredictably even for highly similar chemistries. Henceforth, the notation E_f in this work refers to E_f^{oxide} unless otherwise stated. The binary oxides used to calculate the E_f for garnets and perovskites are listed in Supplementary Table A.2 and A.3, respectively.

Based on the universal approximation theorem[99], we may model the unknown function $f(\chi_C, r_C, \chi_A, r_A, \chi_D, r_D)$, which is clearly non-linear (see Supplementary Fig. A.1), using a feed-forward artificial neural network (ANN), as depicted in Fig. 2.2. The loss function and metric are chosen to be the mean squared error (MSE) and MAE, respectively. We will denote the architecture of the ANN using n^i-$n^{[1]}$-$n^{[2]}$-$\cdots$-1, where n^i and $n^{[l]}$ are the number of neurons in the input and l_{th} hidden layer, respectively.

2.2.2 Neural network model for unmixed garnets

We developed an initial ANN model for unmixed garnets, i.e., garnets with only one type of species each in C, A and D. A data set comprising 635 unmixed garnets was generated by performing full DFT relaxation and energy calculations (see Methods) on all charge-neural combinations of allowed species (Supplementary Table A.2) on the C, A and D sites[1]. This

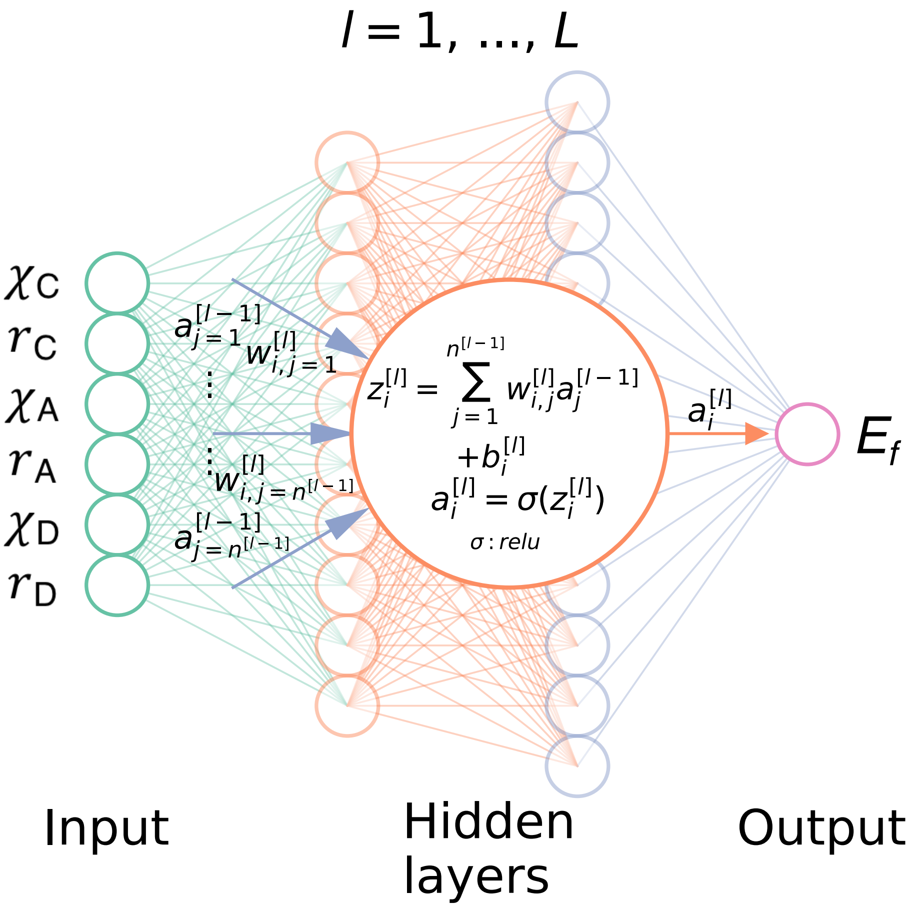


Figure 2.2: General schematic of the artificial neural network. The artificial neural network (ANN) comprises an input layer of descriptors (the Pauling electronegativity and ionic radii on each site), followed by a number of hidden layers, and finally an output layer (E_f). The large circle in the centre shows how the output of the i_{th} neuron in l_{th} layer, $a_i^{[l]}$, is related to the received inputs from $(l-1)_{th}$ layer $a_j^{[l-1]}$. $w_{(i,j)}^{[l]}$ and $b_i^{[l]}$ denote the weight and bias between the j_{th} neuron in $(l-1)_{th}$ layer and i_{th} neuron in l_{th} layer. σ is the activation function (rectified linear unit in this work). The ANN models were implemented using Keras[100] deep learning library with the Tensorflow[53] backend.

dataset was randomly divided into training, validation and test data in the ratio of 64:16:20. Using 50 repeated random sub-sampling cross validation, we find that a 6-24-1 ANN architecture yields a small root mean square error (RMSE) of 12 meV atom^{-1}, as well as the smallest standard deviation in the RMSE among the 50 sub-samples (Supplementary Fig. A.2a). The training, validation and test MAEs for the optimized 6-24-1 model are $\sim$ 7–10 meV atom^{-1} (Fig. 2.3a), an order of magnitude lower than the $\sim$ 100 meV atom^{-1} achieved in previous ML models[40,57,80,81]. For comparison, the error in the DFT E_f of garnets relative to experimental values is around 14 meV atom^{-1} (Supplementary Table A.4). Similar RMSEs are obtained for deep neural network (DNN) architectures containing two hidden layers (Supplementary Fig. A.2b), indicating that a single-hidden-layer architecture is sufficient to model the relationship between E_f and the descriptors.

2.2.3 Averaged neural network models for mixed garnets

To extend our model to mixed garnets, i.e., garnets with more than one type of species in the C, A, and D sites, we explored two alternative approaches — one based on averaging of descriptors, and another based on expanding the number of descriptors to account for the effect of species ordering. The data set for mixed garnets were created using the same species pool, but allowing two species to occupy one of the sites. Mixing on the A sites was set at a 1:1 ratio, and that on the C and D sites was set at a 2:1 ratio, generating garnets of the form $C_3A'A''D_3O_{12}$ (211 compositions), $C'C''A_2D_3O_{12}$ (445 compositions) and $C_3A_2D'D''_2O_{12}$ (116 compositions). For each composition, we calculated the energies of all symmetrically distinct orderings within a single primitive unit cell of the garnet. All orderings must belong to a subgroup of the $Ia\bar{3}d$ garnet space group.

In the first approach, we characterized each C, A, or D site using weighted averages of the ionic radii and electronegativities of the species present in each site, given by the following

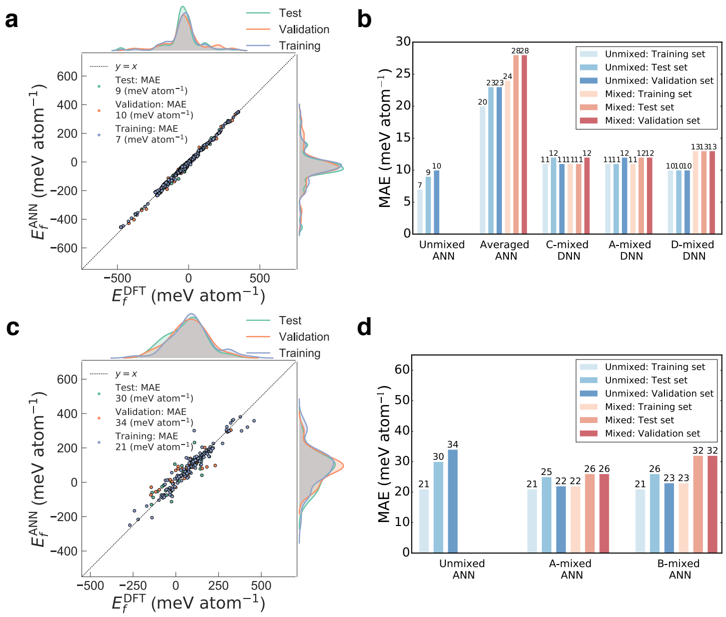


Figure 2.3: Performance of artificial neural network (ANN) models. a. Plot of E_f^{ANN} against E_f^{DFT} of unmixed garnets for optimized 6-24-1 ANN model. The histograms at the top and right show that the training, validation and test sets contain a good spread of data across the entire energy range of interest with standard deviations of 122-134 meV atom^{-1}. Low mean absolute errors (MAEs) in E_f of 7, 10 and 9 meV atom^{-1} are observed for the training, validation and test sets respectively. b. MAEs in E_f of unmixed and mixed samples in training, validation and test sets of all garnet models. The C-, A- and D-mixed DNNs have similar MAEs as the unmixed ANN model, indicating that the neural network has learned the effect of orderings on E_f. Each C-, A- and D-mixed composition has 20, 18, and 7 distinct orderings, respectively, which are encoded using 5-bit, 5-bit and 3-bit binary arrays, respectively. c. MAEs in E_f of unmixed and mixed samples for training, validation and test sets of unmixed perovskites for 4-12-1 ANN model. The E_f^{DFT} of training, validation and test sets similarly contain a good spread of data across the entire energy range of interest with standard deviations of 104-122 meV atom^{-1}. Low mean absolute errors (MAEs) in E_f of 21, 34 and 30 meV atom^{-1} are observed for the training, validation and test sets, respectively. d. MAEs in E_f for training, validation and test sets of all perovskite models. Each A- and B- mixed perovskite compositions has ten distinct orderings, which are both encoded using 4-bit binary arrays. The black lines (dashed) in a. and c. are the identity lines serving as references.

expressions (see Methods):

$$r_{avg} = xr_{\mathrm{X}} + (1-x)r_{\mathrm{Y}} \tag{2.2}$$

$$\chi_{avg} = \chi_{\mathrm{O}} - \sqrt{x(\chi_{\mathrm{X}} - \chi_{\mathrm{O}})^2 + (1-x)(\chi_{\mathrm{Y}} - \chi_{\mathrm{O}})^2} \tag{2.3}$$

where X and Y are the species present in a site with fraction *x* and (1-*x*), respectively, and O refers to the element oxygen. The implicit assumption in this "averaged" ANN model is that species X and Y are completely disordered, i.e., different orderings of X and Y result in negligible DFT energy differences.

Using the same 6-24-1 ANN architecture, we fitted an "averaged" model using the energy of the ground state ordering of the 635 unmixed and 772 mixed garnets. We find that the training, validation, and test MAEs of the optimized model are 22, 26, and 26 meV atom^{-1} , respectively (Supplementary Fig. A.3a). These MAEs are about double that of the unmixed ANN model, but still comparable to the error of the DFT E_f relative to experiments. The larger MAEs may be attributed to the fact that the effect of species orderings on the crystal energy is not accounted for in this "averaged" model.

2.2.4 Ordered neural network model for mixed garnets

In the second approach, we undertook a more ambitious effort to account for the effect of species orderings on crystal energy. Here, we discuss the results for species mixing on the C site only, for which the largest number of computed compositions and orderings is available. For 2:1 mixing, there are 20 symmetrically distinct orderings within the primitive garnet cell, which can be encoded using a 5-bit binary array [b_0, b_1, b_2, b_3, b_4]. This binary encoding scheme is significantly more compact that the commonly used one-hot encoding scheme, and hence, minimizes the increase in the descriptor dimensionality. We may then modify Eqn. 2.1 as follows:

$$E_f = f(\chi_{C'}, r_{C'}, \chi_{C''}, r_{C''}, \chi_A, r_A, \chi_D, r_D, b_0, b_1, b_2, b_3, b_4) \quad (2.4)$$

where the electronegativities and ionic radii of both species on the C sites are explicitly represented. In contrast to the "averaged" model, we now treat the 20 ordering-E_f pairs at each composition as distinct data points. Each unmixed composition was also included as 20 data points with the same descriptor values and E_f, but different binary encodings.

We find that a two-hidden-layer DNN is necessary to model this more complex composition-ordering-energy relationship. The final optimized 13-22-8-1 model exhibits overall training, validation and test MAEs of $\sim$ 11-12 meV atom^{-1} on the entire unmixed and mixed dataset (Supplementary Fig. A.3b). The comparable MAEs between this extended DNN model and the unmixed ANN model is clear evidence that the DNN model has successfully captured the additional effect of orderings on E_f. We note that the average standard deviation of the predicted E_f of different orderings of unmixed compositions using this extended DNN model is only 2.8 meV atom^{-1}, indicating that the DNN has also learned the fact that orderings of the same species on a particular site have little effect on the energy. Finally, similar MAEs can be achieved for A and D site mixing (Supplementary Fig. A.3c and A.3d) using the same approach.

2.2.5 Stability classification of garnets using ANN models

While E_f is a good target metric for a predictive ANN model, the stability of a crystal is ultimately characterized by its E_{hull}. Using the predicted E_f from our DNN models and pre-calculated DFT data from the Materials Project[11], we have computed E_{hull} by constructing the 0 K C-A-D-O phase diagrams. From Fig. 2.4a, we may observe that the extended C-mixed DNN model can achieve a $>$ 90% accuracy in classifying stable/unstable unmixed garnets at a strict E_{hull} threshold of 0 meV atom^{-1} and rises rapidly with increasing threshold. Similarly, high classification accuracies of greater than 90% are achieved for all three types of mixed garnets.

Given the great flexibility of the garnet prototype in accommodating different species, there are potentially millions of undiscovered compositions. Even using our restrictive protocol of single-site mixing in specified ratios, 8,427 mixed garnet compositions can be generated, of which 2,307 are predicted to have E_{hull} of 0 meV atom^{-1}, i.e., potentially synthesizable (Supplementary Fig. A.4a). A web application that computes E_f and E_{hull} for any garnet composition using the optimized DNNs has been made publicly available for researchers at http://crystals.ai.

2.2.6 Neural network models for unmixed and mixed perovskites

To demonstrate that our proposed approach is generalizable and not specific to the garnet crystal prototype, we have constructed similar neural network models using a dataset of 240 unmixed, 222 A-mixed and 80 B-mixed ABO_3 perovskites generated using the species in Supplementary Table A.3. We find that a 4-12-1 single-hidden-layer neural network is able to achieve MAEs of 21-34 meV atom^{-1} in the predicted E_f for unmixed perovskites (Fig. 2.3c), while two 10-24-1 neural networks are able to achieve MAEs of 22-39 meV atom^{-1} in the E_f of the mixed perovskites (Supplementary Fig. A.5). These MAEs are far lower than those of prior ML models of unmixed perovskites, which generally have MAEs of close to 100 meV atom^{-1} or higher[57,81]. As shown in Fig. 2.3, the accuracy of classifying stable versus unstable perovskites exceeds 80% at a strict E_{hull} threshold of 0 meV atom^{-1} and maintains at above 70% at a loosened E_{hull} threshold of 30 meV atom^{-1}. During the review of this work, a new work by Li et al.[56] reported achieving comparable MAEs of $\sim$ 28 meV atom^{-1} in predicting the E_{hull} of perovskites using a kernel ridge regression model. However, this performance was achieved using a set of 70 descriptors, with model performance sharply dropping with less than 70 descriptors. Furthermore, Li et al.'s model is restricted to perovskites with $E_{hull} < 400$ meV atom^{-1} and only a single ordering for each mixed perovskite, while in this work, the highest E_{hull} is 747 meV atom^{-1} for the perovskite dataset and all symmetrically distinct orderings on the A and B sites within a $\sqrt{2} \times \sqrt{2} \times 1$ orthorhombic conventional perovskite unit cell (ten structures each) are considered.

2.3 Discussion

To summarize, we have shown that NN models can quantify the relationship between traditionally chemically intuitive descriptors, such as the Pauling electronegativity and ionic radii, and the energy of a given crystal prototype. A key advantage of our proposed NN models is that they rely only on an extremely small number (two) of site-based descriptors, i.e., no structural degrees of freedom are considered beyond the ionic radii of a particular species in a site and the ordering of the cations in the mixed oxides. This is in stark contrast to most machine-learning models in the literature utilizing a large number of correlated descriptors, which render such models highly susceptible to overfitting, or machine-learning force-fields, which can incorporate structural and atomic degrees of freedom but at a significant loss of transferability to different compositions. Most importantly, we derive two alternative approaches — a rigorously defined averaging scheme to model complete cation disorder and a binary encoding scheme to account for the effect of orderings—to extend high-performing unmixed deep learning models to mixed cation crystals with little/no loss in error performance and minimal increase in descriptor dimensionality. It should be noted that our NN models are still restricted to the garnet and perovskite compositions (with or without cation mixing) with no vacancies, though further extensions to other common crystal structure prototypes and to account for vacancies should in principle be possible. Finally, we show how predictive models of E_f can be combined with existing large public databases of DFT computed energies to predict E_{hull} and hence, phase stability. These capabilities can be used to efficiently traverse large chemical spaces of unmixed and mixed crystals to identify stable compositions and orderings, greatly accelerating the potential for novel materials discovery.

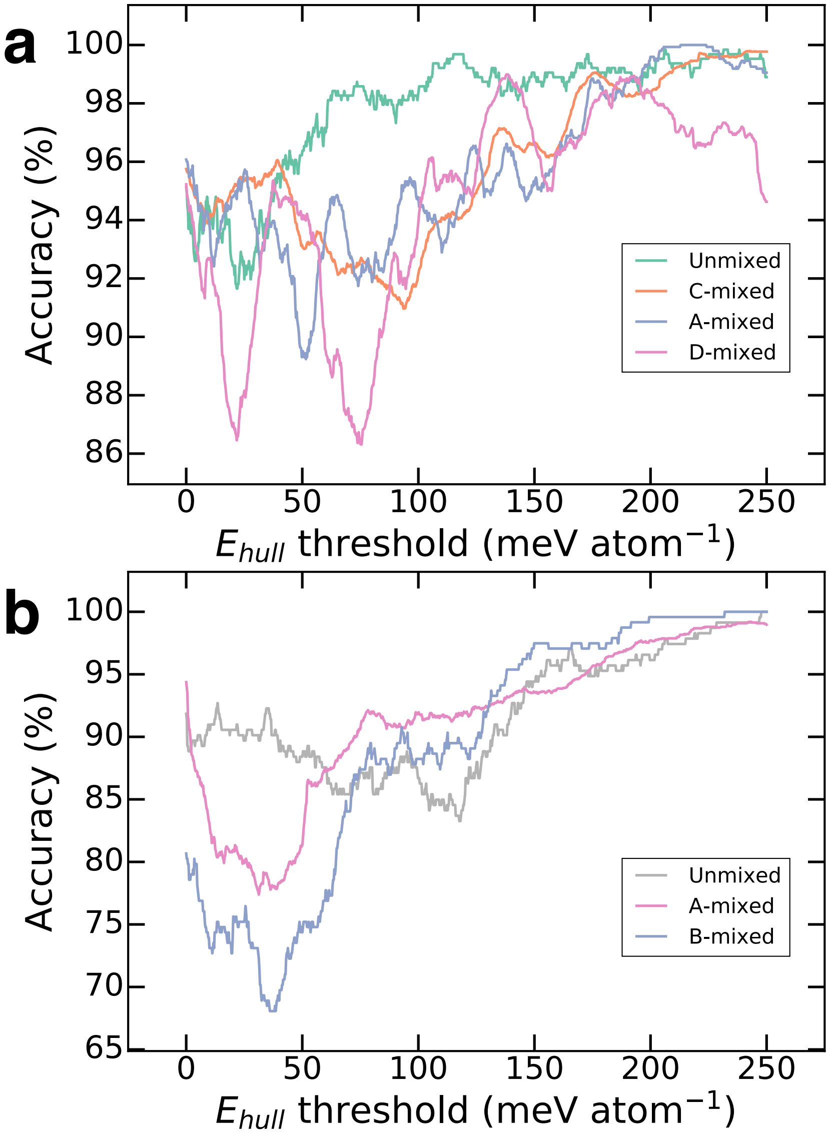


Figure 2.4: Accuracy of stability classification. Plots of the accuracy of stability classification of the ANN models compared to DFT as a function of the E_{hull} threshold for a. garnets, and b. perovskites. The accuracy is defined as the sum of the true positive and true negative classification rates. A true positive (negative) means that the E_{hull} for a particular composition predicted from the optimized artificial neural network model and DFT are both below (above) the threshold. For the mixed compositions, an E_{hull} is calculated for all orderings (20, 7 and 18 orderings per composition for C-, A- and D-mixed garnets, respectively, and 10 orderings per composition for both A- and B-mixed perovskites).

2.4 Methods

2.4.1 DFT calculations

All DFT calculations were performed using Vienna ab initio simulation package (VASP) within the projector augmented wave approach[101,102]. Calculation parameters were chosen to be consistent with those used in the Materials Project, an open database of pre-computed energies for all known inorganic materials[11]. The Perdew-Burke-Ernzehof generalized gradient approximation exchange-correlation functional[103] and a plane-wave energy cut-off of 520 eV were used. Energies were converged to within 5×10^{-5} eV atom^{-1}, and all structures were fully relaxed. For mixed compositions, symmetrically distinct orderings within the 80-atom primitive garnet unit cell and the 40-atom $\sqrt{2} \times \sqrt{2} \times 1$ orthorhombic perovskite supercell were generated using the enumlib library[104] via the Python Materials Genomics package[33].

2.4.2 Training of ANNs

Training of the artificial neural networks (ANNs) was carried out using the Adam optimizer[105] at a learning rate of 0.2, with the mean square error of E_f as the loss metric. For each architecture, we ran with a random 64:16:20 split of training, validation and test data, i.e., random sub-sampling cross validation.

2.4.3 Electronegativity averaging

Pauling's definition of electronegativity is based on an "additional stabilization" of a heteronuclear bond X-O compared to average of X-X and O-O bonds, as follows.

$$(\chi_{\mathrm{X}} - \chi_{\mathrm{O}})^2 = E_d(\mathrm{XO}) - \frac{E_d(\mathrm{XX}) + E_d(\mathrm{OO})}{2} \tag{2.5}$$

where χ_X and χ_O are the electronegativities of species X and O, respectively, and E_d is the dissociation energy of the bond in parentheses. Here, O refers to oxygen.

For a disordered site containing species X and Y in the fractions x and (1-x), respectively, we obtain the following:

$$\begin{aligned}(\chi_{\mathrm{X}_x\mathrm{Y}_{1-x}} - \chi_{\mathrm{O}})^2 &= xE_d(\mathrm{XO}) + (1-x)E_d(\mathrm{YO}) - \frac{xE_d(\mathrm{XX}) + (1-x)E_d(\mathrm{YY}) + E_d(\mathrm{OO})}{2} \\ &= x(\chi_{\mathrm{X}} - \chi_{\mathrm{O}})^2 + (1-x)(\chi_{\mathrm{Y}} - \chi_{\mathrm{O}})^2\end{aligned} \tag{2.6}$$

We then obtain the effective electronegativity for the disordered site as follows:

$$\chi_{\mathrm{X}_x\mathrm{Y}_{1-x}} = \chi_{\mathrm{O}} - \sqrt{x(\chi_{\mathrm{X}} - \chi_{\mathrm{O}})^2 + (1-x)(\chi_{\mathrm{Y}} - \chi_{\mathrm{O}})^2} \tag{2.7}$$

2.4.4 Data availability

The datasets generated during and/or analysed during the current study are available in the GitHub repository https://github.com/materialsvirtuallab/garnetdnn as well as the Dryad Digital Repository (doi: 10.5061/dryad.760r5b6). A web application that estimates E_f and E_{hull} for any given garnet or perovskite composition using the optimized DNNs is available at http://crystals.ai/.

Chapter 2, in full, is a reprint of the material “Deep neural networks for accurate predictions of crystal stability” as it appears on Nature Communications, Weike Ye, Chi Chen, Zhenbin Wang, Iek-Heng Chu, Shyue Ping Ong, 2018, 9 (1), 1-6. The dissertation author was the primary investigator and author of this paper.

Chapter 3

High-throughput screening of Eu^{2+}-doped red-emission garnet phosphors using density functional theory and machine learning

3.1 Introduction

Solid-state white-light-emitting diodes (wLEDs) are energy efficient, robust, durable and environment-friendly solid state lighting devices[106,107]. Nowadays, a blue diode chip combined with a yellow phosphor such as $Y_3Al_5O_{12}$:Ce^{3+} is still the most mature method for fabricating commercial wLEDs. However, they suffer from poor color rendering effects due to the lack of red components[108]. To this end, great efforts have been made to explore novel red phosphors. So far, the popular choices of red-emitting activators are Mn^{4+}[109,110], Eu^{3+}[111,112], and Eu^{2+}[113–116]. Mn^{4+}-doped phosphors are mainly fluorides[109,117], which are notoriously known for poor chemical stability and great synthesis difficulty. On the other hand, the Eu^{3+}-doped phosphors suffer from poor absorption efficiency under blue light excitation as the maximum absorption peaks are often the results of the charge transfers taking place in the UV and near-UV region. This problem can be circumvented by doping Eu^{2+} instead. However, existing high-profile Eu^{2+}-doped red phosphors are mainly nitrides[113–116,118–120], which require harsh synthesis conditions. Therefore, the developments of Eu^{2+}-doped red oxide phosphors are necessary to complement current phosphor materials. Garnets are known as superior hosts for high efficiency and thermal stability[121–124], however, there is only one Eu^{2+}-doped red-emission phosphor reported hitherto[110] that adopts the garnet structure, leaving many opportunities for new materials discovery.

Computational high throughput screening (HTS) is an effective approach that down-selects a large pool of candidates based on successive property evaluations and is often adopted to search for new phosphors[10,125]. Screening of phosphor usually considers cost, safety, phase stability, emission color, thermal quenching (percentage loss of emission at elevated temperatures during operation), and other more refined assessments[9]. The cost and safety factors are often assured by the constraints on the candidates' constituent elements, and the rest of the properties are often assessed by the density functional theory (DFT)[2]. Despite the advances in theoretical methodologies and computational power, the major bottleneck in HTS is still the high computational cost of DFT calculations when the number of candidates reaches a medium level of thousands. An emerging bypass is to develop surrogate machine learning (ML) models for DFT, which accurately map the structures to the properties at sub-second high speed. In fact, there have already been several successful cases of utilizing ML-DFT hybrid HTS to discover novel energy materials, such as quaternary Heusler compounds[60], photovoltaic materials[126], nitrogen fixation catalysts[127], etc.

A critical property to evaluate in the phosphor HTS is the emission color. Previous works have shown that the bandgap (E_{bg}) of the host material is inversely related to the emission wavelength[10]. Predicting bandgap using ML models has been investigated extensively[27,62–65,128–130]. Due to the diversity of the models, the accuracy of them varies. In general, two categories of structure-related ML models exist, namely the structure-agnostic models that work with all structures, and the structure-specific models that deal with a specific structure type, e.g., perovskite[63–65], MXene[131], etc. The typical mean absolute errors (MAEs) of structure-specific models are 0.1 to 0.3 eV[28,28,129–131]. Despite the generalizability of the structure-agnostic models, they usually have higher errors. Up to now, there has not been reported a structure-specific model for garnet, whereas the state-of-the-art general model MEGNet[27] shows MAE of 0.43 eV on our garnet data set, which is higher than that of the general structure-specific models. Hence, it is necessary to develop a more accurate ML model to allow rapid assessment of garnet bandgaps.

In this work, we developed an accurate and interpretable ML model to predict the bandgap of garnets. We devised an ML-DFT hybrid workflow for screening Eu^{2+}-doped red-emission phosphor materials in the garnet family. The workflow combines the deep neural networks (DNNs) for phase stability[132], a newly developed model for bandgap as a proxy for emission, and the thermal quenching prediction algorithm[133]. Following the workflow, we successfully identified two promising candidates, Eu^{2+}-doped $CaEr_2Mg_2Si_3O_{12}$ and $CaTb_2Mg_2Si_3O_{12}$, which have high synthesizability, desired emission, and thermal stability.

3.2 Methods

3.2.1 DFT

Host structure All DFT calculations were performed using Vienna *ab initio* simulation package (VASP) within the projector augmented-wave approach[101,102]. The exchange-correlation interaction was described using the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation (GGA) functional[103]. The plane-wave energy cut-off was set at 520 eV, and the energies were converged to within 5×10^{-5} eV atom^{-1}. Symmetrically distinct orderings within the 80-atom primitive garnet unit cell for mixed compositions were generated using the enumlib library[104] *via* the python materials genomics (pymatgen) package[33]. For the calculation of the GGA bandgaps of the hosts, the k-point line density along the high symmetry line of the Brillouin zone is set at 20.

Doped structure To obtain the doped structure, Eu^{2+} replaces one of the +2 cation occupying C site or A site in the primitive cell. PBE calculations with a Hubbard U[134] parameter of 2.5 eV for Eu was used for these doped systems, same as the previous work done on oxide phosphors[135]. To compare energies and calculate doping formation energies, the structure relaxation and energy calculation were computed using the same settings as the host. The doping formation energy $E_f(Eu_M^{\times})$ was calculated using the formalism illustrated by Zhu et al.[136], where the Kröger-Vink notation[137] for defect is used and M is the +2 cation replaced by Eu^{2+}. The structures were further relaxed until the electronic energy and the atomic forces were converged to within 1.25×10^{-6} eV atom^{-1}, and 0.01 eV Å^{-1}.

The excited $4f^6 5d^1$ state of Eu^{2+} doped garnets was approximated using constrained DFT (CDFT) method, where the occupancy of the top most Eu 4f state (at the valence band) was transferred to the lowest 5d state (at the conduction band) and kept fixed during the calculations using the ground state structure. The energy difference between this state and the ground state is considered the excitation energy. The $4f^6 5d^1$ ground-level crystal structure was obtained through structural optimization under the same electron configuration. The energy difference of the obtained structure under the electron occupancy of the excited state ($4f^6 5d^1$) and that of the ground state ($4f^7$) is considered as the emission energy[135]. We tested the method on a known Eu^{2+}-doped phosphor, $Sr_3Y_2Ge_3O_{12}$:Eu^{2+}[138]. The calculated and experimentally reported excitation wavelengths are 450 nm and 468 nm, and that of emission wavelengths are 632 nm and 612 nm (Figure B.5).

3.2.2 Feature and model developments

Feature All the elemental attributes were obtained through pymatgen[33], except the number of valence electrons (NVE), which was obtained from Magpie[40].

Feature selection is the procedural approach to find the relevant subset of input variables. Simplification of the input variables brings mainly four benefits, i.e., enhancing the model interpretability, shortening the training time, avoiding the “curse of dimensionality” and reducing overfitting. In this work, we adopted an exhaustive feature selection method, where the cross-validation (CV) scores of the models trained with all possible combinations of features under the same hyper-parameters are recorded. The best feature set emerges at the turning point where adding any new feature leads to little change in the performance of the model.

ML model development eXtreme gradient boosting (XGBoost)[139] is a decision-tree-based ensemble ML algorithm that uses a gradient boosting framework. It provides a parallel tree boosting that solve many data science problems in a fast and accurate fashion, which also makes it our choice. The training of the gradient-boosted tree model was carried out using the XGBoost library[139]. 1823 data points were split in the ratio of 4:1 for training and test, respectively. During the training, 5-fold CV was performed.

Model interpretation SHapley Additive exPlanations (SHAP) is a technique that explains the output of a ML model by applying a game-theoretic approach to calculate the importance of individual input features to a given model prediction[140]. A positive (negative) SHAP value means that, at the given feature value, there are more instances with predicted bandgap higher (lower) than the average prediction. The analysis is often presented in a summary plot, and/or a partial dependent plot (PDP) of chosen features. In the former, the SHAP values of all the data and for all the features are presented. Usually, information such as the rank of the feature importance and the general trend of the impact of each feature on the model prediction can be obtained. The PDPs, on the other hand, show the marginal effect one feature has on the predicted outcome of an ML model[141].

3.3 Results

3.3.1 Data overview

Garnet structures have the general formula of $C_3A_2D_3O_{12}$, where C, A, and D refer to the three symmetrically distinct cation sites with Wyckoff symbols 24*c*(dodecahedron), 16*a*(octahedron) and 24*d*(tetrahedron), respectively, in the prototypical cubic $Ia\bar{3}d$ garnet crystal. By making variations of the species on the C, A, and D sites (Figure 3.1(b)), we have generated "unmix" garnets with the formula $C_3A_2D_3O_{12}$ and "mixed" garnets of the formulas $C_3A'A''D_3O_{12}$, $C'C''A_2D_3O_{12}$ and $C_3A_2D'D''_2O_{12}$. The structure generation strategy leads to a total of 20406 charge-neutral garnet compositions. We performed DFT band structure calculations using the PBE functional on 1823 of the generated garnets, which comprise 517 unmix ($C_3A_2D_3O_{12}$), 517 C-mixed ($C'C''A_2D_3O_{12}$), 484 A-mixed ($C_3A'A''D_3O_{12}$) and 305 D-mixed ($C_3A_2D'D''_2O_{12}$) garnets (Figure 3.1(a)). All the E_{bg}^{DFT}s are spread in the range of 1-5 eV, with the highest population lies in between 2-4 eV, as shown in Figure3.1(a). The distributions are unbiased among categories and have a reasonable population of data in the range of 3-4 eV. In particular, the distribution of C-mixed, which spans from 1.5 to 4.5 eV, is slightly "narrower" than that of the A- and D-mixed. The E_{bg}^{DFT} medians of both A-mixed and D-mixed are lower than that of the unmix. These observations could potentially suggest that introducing doping in A site and D site could be an effective strategy to change the E_{bg}, whereas doping in C site may have a milder effect.

3.3.2 Feature selection and model selection

We adopted the gradient boosting tree regression models for the bandgap prediction task and an exhaustive feature selection method to locate the most compact and effective set of features. The XGBoost algorithm is used for the model training, with the hyper-parameter settings as follows: $n\_estimators = 200$, $max\_depth = 6$, $learning\_rate = 0.1$, $gamma = 0$, and default values for the rest.

We performed the feature selection in two steps, i.e., the attribute selection and the feature selection. Here, the attribute refers to the elemental property, and the feature is the combination of the attributes of species in C, A, and D sites. The latter is the actual inputs of the model.

Attribute selection We started with a list of elemental properties that relate to the crystal electronic structure. It should include periodic table information of constituent elements (atomic number (Z), group number (*Group*), row number (*Row*), and Mendeleev number ($Mendeleev$)), the size of the atoms (atomic radius (AR)), and the electronic structure (electronegativity (χ), ionization energy (IE), $polarizability$, and the number of valence electrons (NVE)). There is redundant information among the chosen attributes to some degree. For example, the *Group* and NVE are the same for elements in periods 1-3, and the χ and IE have exactly the opposite trend theoretically. Therefore we calculated the Pearson correlation matrix for all the attributes (Figure3.2(a)), and for the pairs with correlation coefficients above 0.75, i.e., $Mendeleev$ and χ, Z and *row*, AR and χ, $Mendeleev$ and AR, χ and IE, AR and IE, and $Mendeleev$ and IE, we kept the ones with which the model performs better with single attribute features. For example, in the pair of $Mendeleev$ and χ, the 5-fold CV MAE of the feature $Mendeleev$ is 0.22 eV and that of the χ is 0.19 eV (Figure B.2), therefore we kept χ and discarded $Mendeleev$. After the elimination process, the final attributes are Z, $Group$, IE, and NVE.

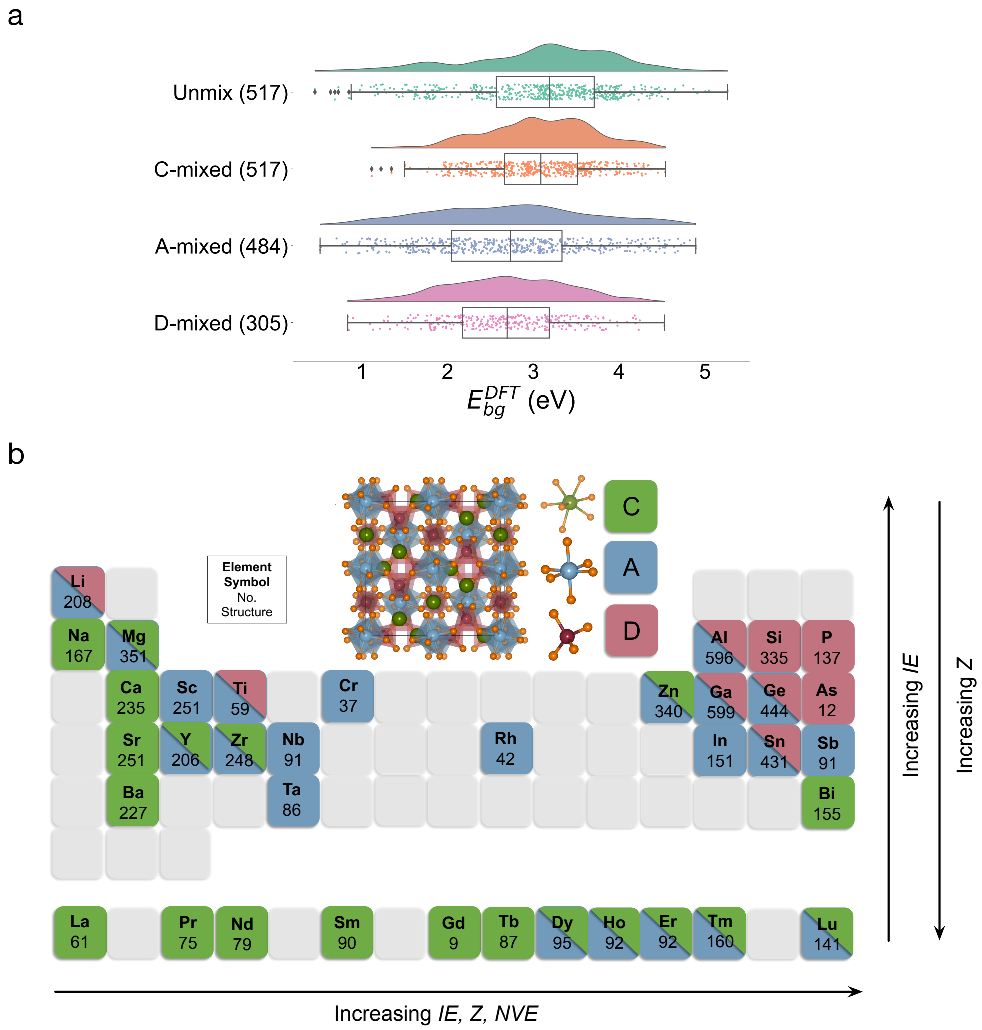


Figure 3.1: Overview of data. (a) Data distribution for different data categories. The numbers in the brackets indicate the number of structures calculated from each category. (b) The structure distribution over elements. The elements' positions are organized based on the periodic table. In each box, the top symbol is the element symbol, and the numbers are the total number of structures containing the element. The color of the box indicates the site preference of the element (green for the C site, blue for the A site, and pink for the D site), which is adapted from ref. 1.

Feature selection Given the selected 4 attributes, there are 12 features per structure: 4 for each of the C, A, and D sites. This gives a total of 4095 feature subsets. We exhaustively examined the performances of all the feature subsets using the same training data and model settings. CV scores for all 4095 combinations can be found in Figure B.3, and the best-achieved CV scores of each dimension of features are shown in 3.2(b). We observe the performance of model converges at around 0.1 eV when the dimension of feature reaches 6, and the best feature array in the dimension of 6 is Z_C, IE_C, Z_A, IE_A, Z_D and NVE_D, with which the model achieves the mean CV score of 0.13 eV.

Figure3.2(c) presents the parity plot of the selected model. The model's test MAE is the same as the mean CV score, 0.13 eV, suggesting the model's superior generalizability. The MAE of 0.13 eV is on par with the state-of-the-art structure-specific models, with an extremely compact set of features.

3.3.3 Model interpretation

Scrutiny on the model's behavior is critical in materials science since the model needs to make both statistical and physical/chemical sense. We employed the SHAP technique (see Method) to shed light on the relationships between features and the model predictions, of which the results are shown in Figure 3.3. To begin with, the features in the figure are shown in the descending order of importance from top to bottom, where we observe that Z_D is the most important feature, followed by IE and Z from both A and C sites, and the least importance goes to the NVE of the D site. The horizontal location in the left part of the Figure 3.3 shows whether the effect of that feature value is associated with a higher or lower SHAP value, and the color is an indication of the feature values. One should quickly notice that all the features are negatively correlated with the prediction since most of the instances with lower feature values (blue points) have positive SHAP values and vice versa. Similar correlations can also be captured from the training data itself. Figure B.4 shows that the feature values and the distributions of PBE bandgap

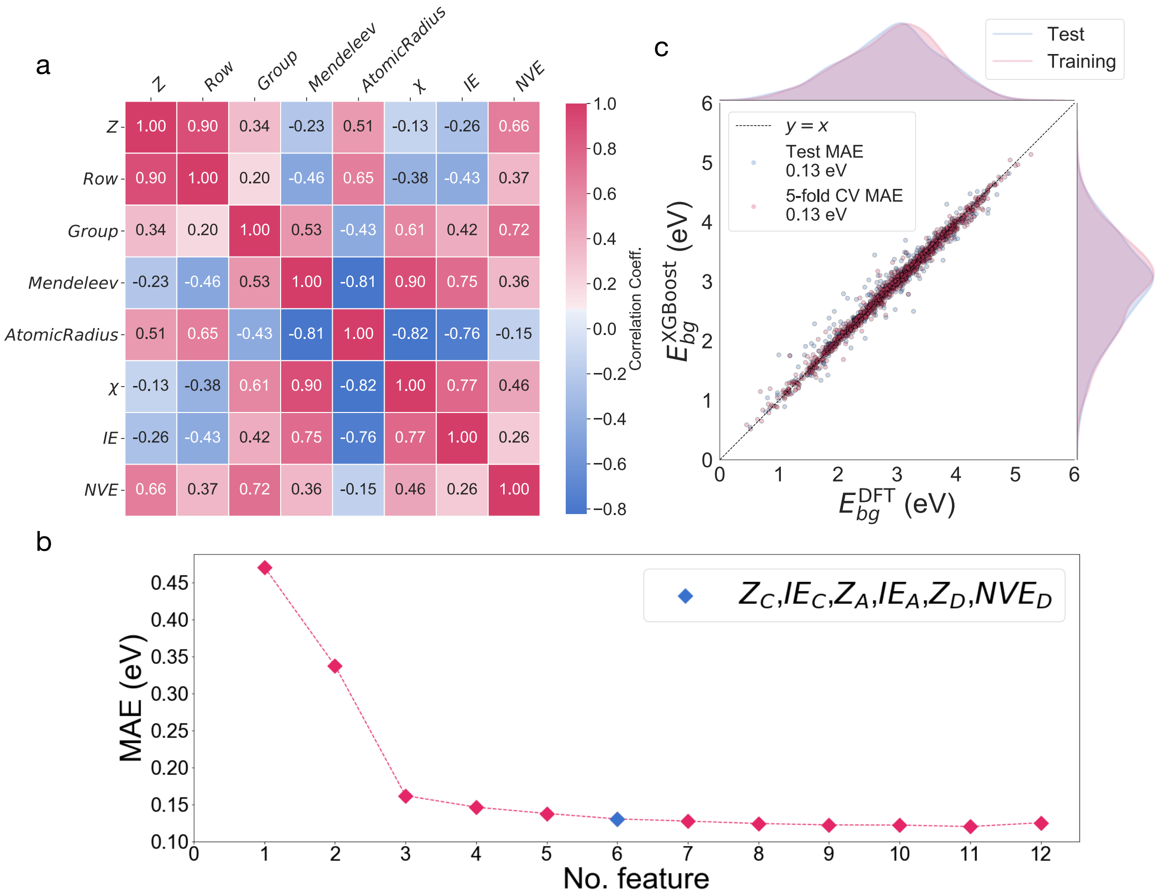


Figure 3.2: Feature selection. (a) The Pearson correlation matrix of the initial 10 atomic attributes. Pairs of atomic attributes with a correlation coefficient larger than 0.75 are considered highly correlated. (b) The lowest MAE achieved versus dimension of features ranging from 1 to 12. (c) The parity plot of the model using the feature set of (Z_C,IE_C,Z_A,IE_A,Z_D,NVE_D).

for the elements in various sites, where patterns of the negative correlation between the features of IE_C, IE_A, Z_D and NVE_D, and the mean of the PBE bandgap are observed. To rationalize the trend, we start from the most physical intuitive attribute, IE, and explain other attributes by relating to it. IE describes the atomic electronic structure directly, and the higher the IE, the less ionic the M-O (M: metal, O: oxygen) bonds become and the smaller the overlap between metal valence bands and the oxygen $2p$ bands, hence the smaller the energy gap. An increase in the Z can increase either the group number (increasing the IE) or the row number (decreasing the IE). However, changing the group number usually induces a more significant variation in the IE than changing the row number. Therefore, the net effect of increasing the Z should have the most similar trend as increasing the IE, i.e., the higher the Z, the smaller the prediction. The increase of the NVE can result from the increase of the group number in the same period, where the effective nuclear charge felt by each electron rises as the NVE increases, hence the IE increases. It explains a similar negative correlation of the NVE to the predicted bandgap as that of the IE. In terms of the ranking of feature importance, the plot shows that features from D and A sites outrank those from C site, which agrees with our observation from the breakdown of the E_{bg}^{DFT} distributions in Figure 3.1(a). One possible explanation is that the elements that prefer D sites are mostly from p blocks (Figure 3.1(b)), making them highly possible to form such polyatomic anions as ${PO_4}^{3-}$ in a tetrahedron coordination environment, and thus mainly affect the valence band positions.

We continue the analysis with the partial dependence plots (PDP) in the right part of Figure 3.3. Before we start, it is worth mentioning that the PDP's validity requires independence in the features, which has been assured by our rigorously-performed feature selection procedure. From the PDP plots of the Z features, increasing Z_C up to about 60 (the beginning of lanthanides) decreases the bandgap predictions, after which the trend reverses. It is similar with Z_A, despite more noises in the data. For Z_D, it reveals a more straightforward monotonic pattern, where increasing Z_D decreases the prediction. We can also observe three plateaus of the SHAP values at around 10-20, 25-37, and 37-50, which should map to p elements in the periods of 2 (Al, Si, P), 3

(Ga, Ge, As), and 4 (In, Sn), respectively. In terms of IEs, both for IE_C and IE_A, small IE values have limited contribution to the prediction and increase IE after around 6.7 eV starts to reduce the bandgap prediction. Finally, the PDP of NVE from the D site shows increasing of the valence electrons up to 8 leads to the decrease of the prediction, and after 8 the trend reverses. It can be explained by the shielding effect, where when the NVE is larger than 8, the electrons start to fill the more localized d orbitals and shield the nuclear charge, making the valence electron easier to remove and hence the larger bandgaps. Again we notice a strong resemblance between the model's behavior and the pattern of the training data. According to Figure B.4(c), the D elements of Sn, Ge, Ga, and As, which all have the Z above 30, and the NVE above 8, have the smallest mean of the PBE bandgap. This agrees with the negative SHAP values associated with the same range of Z and NVE as shown in the PDPs.

To summarize, the developed model captures the observed patterns in the training data, and the behaviors of the model agree well with physical and chemical intuitions.

3.3.4 Screening of Eu^{2+}-doped red-emission phosphor

Design of workflow The screening workflow considers the material cost, safety, phase stability, emission color, thermal stability, dopabilitiy, and dynamic stability. The cost and safety are often assured by the constraints on the constituent elements of the candidates. The phase stability of garnets can be rapidly assessed by our previously developed ML models[132]. For thermal stability, the algorithm developed by Amachraa et al. based on Voronoi area renders us the ability to approximate the percentage of intensity that can be maintained by the phosphor material when temperature elevates from 300 to 500 K in sub-minute[133]. The dopability, which is usually measured by the doping formation energy, and dynamic stability, as is often approached by phonon spectrum, can be completed by DFT for a narrowed list of candidates. The emission color screening is the only remaining challenge in the design of the workflow.

Based on the underlying physics of photoluminescence, the emission wavelength should

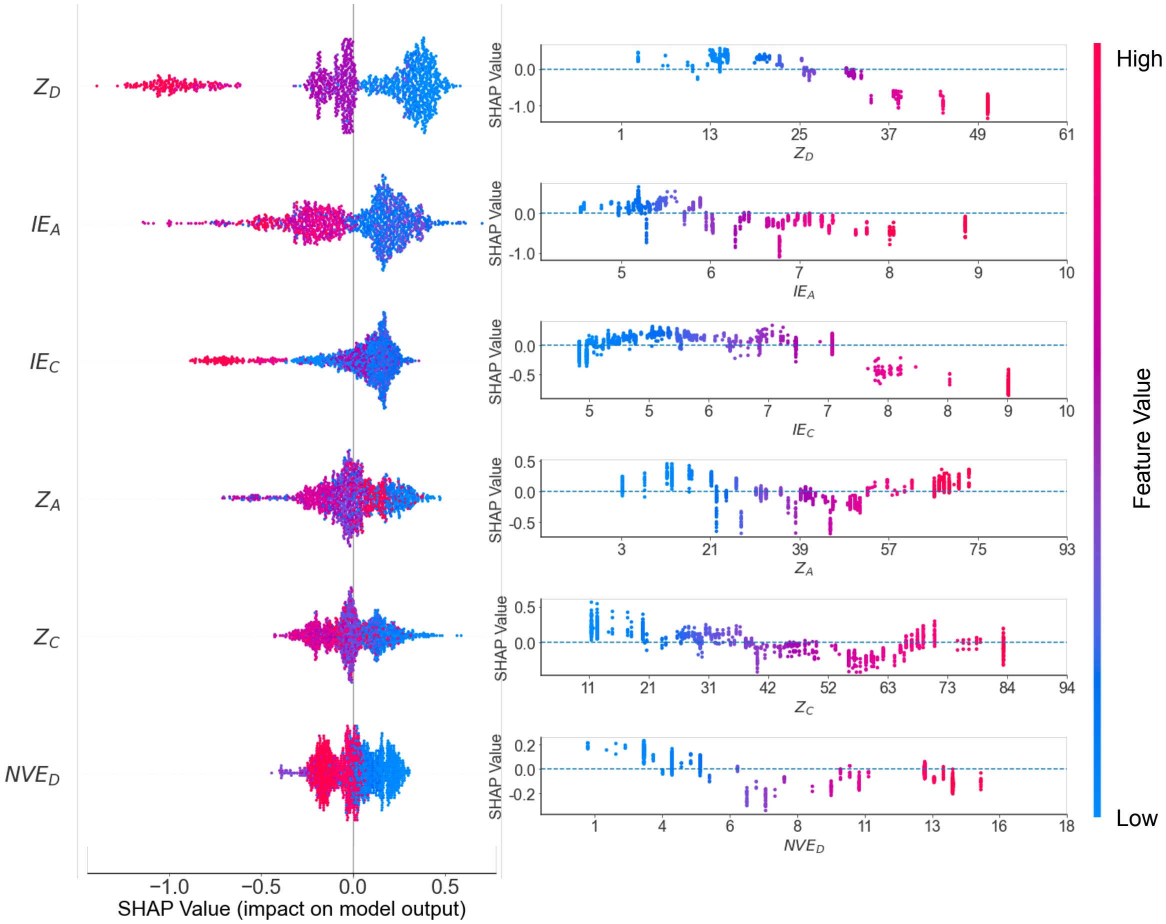


Figure 3.3: The SHAP analysis. The summary plot where all the SHAP values are shown (left) is joined by the partial dependence plots (PDP) of each feature (right). The blue dashed lines in the PDPs are the zero lines of SHAP values.

be negatively related to the bandgap. Indeed, according to Wang et al.[9] and Figure B.1, which summarizes the emission wavelengths and the calculated Perdew–Burke-Ernzerhof (PBE)[103] bandgaps of reported Eu^{2+}-doped phosphors, most of the Eu^{2+} red emission phosphors require the host (PBE) bandgap to be below 4 eV but above 3 eV. Particularly, the one red phosphors in the figure that adopts the garnet structure, i.e. $Sr_3Y_2Ge_3O_{12}$, has a PBE bandgap of 3.12 eV, well within the window of 3-4 eV. Therefore, the PBE bandgap can be an effective proxy for the emission energy or color of the phosphors. Based on the assumptions, we assembled the hybrid phosphor screening workflow as shown in Figure 3.4.

Screening results Starting from the 20406 charge-neutral candidates discussed in Section 3.3.1, we excluded the compositions containing elements Yb, Ho, Dy, Eu, Sc, Rh, Cd, As, and Pb from cost and toxicity considerations. Furthermore, we limited the stoichiometric ratio of rare-earth (RE) elements to be less than 2.5%, i.e., less than two RE elements per standard formula. This step filtered out most compositions and left 5554 candidates. 1357 out of the 5554 compositions are predicted to have E_{hull} = 0, which signals phase stability. Out of these 1357 stable garnets, the model developed by this work predicts 667 to have a PBE bandgap between 3-4 eV. The DFT verification on the E_{hull} and E_{bg} confirms 183 out of the 667 are valid. Furthermore, the Voronoi area analysis predicts that only 22 (shown in Table 3.1) can maintain more than 85% of the efficiency when the temperature is elevated from 300K to 500K.

A similarity noticed among the candidates is that the C site species are mostly Ca and Sr. According to Figure 3.3(b), when Z_C is 20 (Ca) or 38 (Sr), and IE_C is 6.1 (Ca) or 5.7 (Sr) eV, the SHAP values are close to but above 0, meaning the majority of the instances have the predicted bandgap close to but higher than the mean of the target (2.92 eV), well within our desired range.

We successfully verified the emission energies *via* the CDFT method for two out of these candidates. As shown in the band structures (BS) and densities of states (DOS) from Figure 3.5 (e) and (f), for the doped candidates, the PBE bandgaps are both $\sim$ 4 eV, which are expected based on the insights we extracted from the model, manifesting that when there are light elements

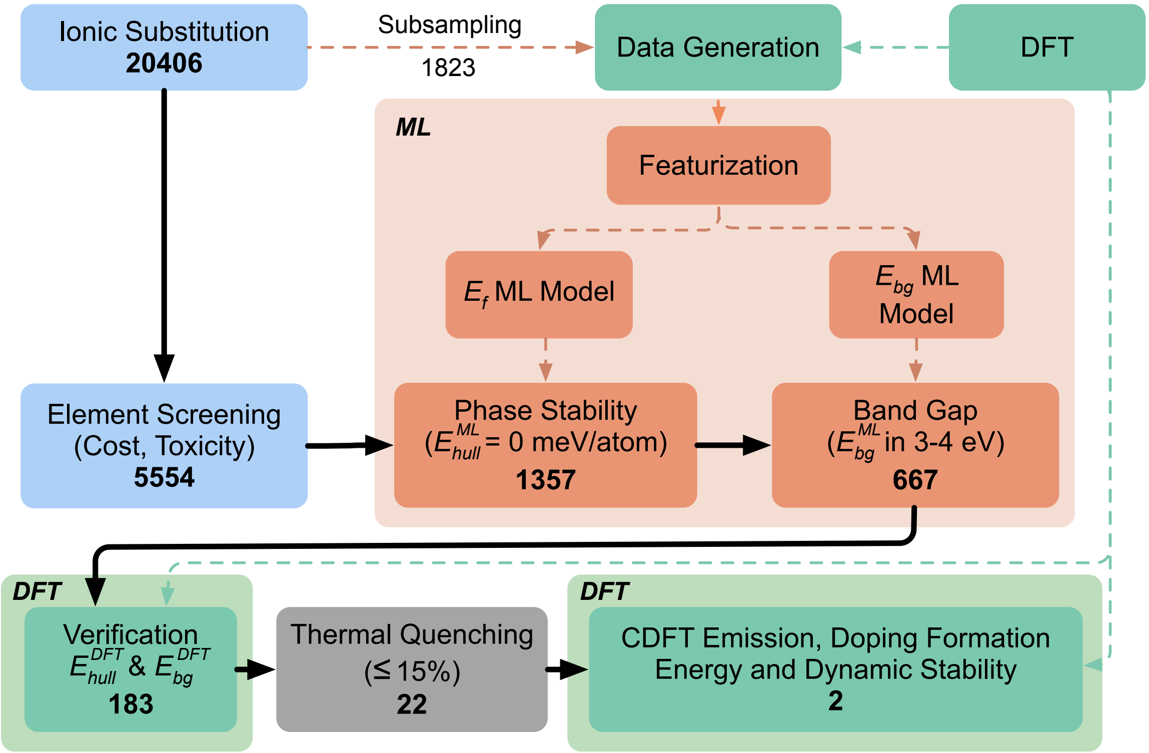

Ionic Substitution
20406
Subsampling
1823
Data Generation
DFT
ML
Featurization
E_f ML Model
E_bg ML Model
Element Screening (Cost, Toxicity)
5554
Phase Stability (E_hull^ML = 0 meV/atom)
1357
Band Gap (E_bg^ML in 3-4 eV)
667
DFT
Verification E_hull^DFT & E_bg^DFT
183
Thermal Quenching (≤ 15%)
22
DFT
CDFT Emission, Doping Formation Energy and Dynamic Stability
2

Figure 3.4: The workflow of high-throughput screening (HTS) of Eu^{2+}-doped red-emission garnet phosphors using density functional theory (DFT) and machine learning (ML). The cubic $Ia\bar{3}d$ prototypical garnet structure contains three cation sites, i.e., 24*c* (dodecahedron), 16*a* (octahedron), and 24*d* (tetrahedron). By varying the composition of each site, we created 20406 candidate garnet materials. For the development of ML models, the phase stability and bandgap of 1823 out of the 20406 structures were calculated by DFT and were used as training data. The HT screening starts with elemental screening based on the cost and toxicity of the constituent elements. Then the two ML models were used to identify candidates with stable phase (E_{hull}^{ML} = 0 eV atom^{-1}) and desired bandgap (E_{bg}^{ML} between 3 to 4 eV). DFT was performed to verify E_{hull}^{DFT} and E_{bg}^{DFT} for the candidates hitherto, and at the same time, the optimized structures of the candidates were obtained. The algorithm then approximated thermal quenching ratio (TQ) based on Voronoi area[133] and candidates with TQ less than 15 % were kept. Doping formation energy and the candidates' phonon dispersion spectrum were also calculated for the final candidates to shed light on dopability and dynamic stability. Following this workflow, we successfully identified two promising Eu^{2+}-doped garnet phosphors, i.e. $CaEr_2Mg_2Si_3O_{12}$ and $CaTb_2Mg_2Si_3O_{12}$.

on both A (Mg) and D (Si) sites, the PBE bandgaps are likely to be higher than the average prediction of 2.92 eV. Furthermore, we located the lowest bands in the conduction band minimum (CBM) with the most 5d character from the BS and DOS. Afterward, we calculated the excited state by adjusting the electron occupancy from the highest 4f band to the lowest 5d band. The emission energies for $CaEr_2Mg_2Si_3O_{12}$ and $CaTb_2Mg_2Si_3O_{12}$ are calculated to be 639 nm and 685 nm, respectively (Figure 3.5(c) and (d)). The previous discussion on the emission energy is based on the PBE bandgap of the host. Now that we have the doped structure at hand, we shall revisit the topic from the perspective of the activator's local environment. First of all, Eu^{2+}, in both candidates, replaces the smaller C site cation, i.e., Er^{3+} and Tb^{3+}, making the overall bond lengths shorter, which leads to a stronger crystal field splitting. Secondly, the difference in the two candidates' emission energy can be explained by the local environment's distortion. As defined by Wang et al.[142], the distortion index D for a polyhedron local environment can be calculated as follows:

$$D = \frac{1}{n}\sum_{i=1}^{n}\frac{|l_i - l_{av}|}{l_{av}}, \tag{3.1}$$

where l_i is the distance from the center atom (Eu) to the ith coordinating atom (O), l_{av} is the average bond length, and n is the coordination number (n = 8 in the dodecahedron environment). The D for the EuO_8 polyhedron in $CaEr_2Mg_2Si_3O_{12}$ and $CaTb_2Mg_2Si_3O_{12}$ are 0.011 and 0.014, respectively (Figure 3.5 (a) and (b)), indicating that the Tb compound is more distorted. The stronger distortion leads to the more significant splitting of the 5d bands, and therefore the lower CBM and smaller the emission energy.

Regarding the synthesizability, DFT calculations confirmed that E_{hull} of the both candidates are 0 eV atom^{-1}, and there are no imaginary frequencies in their phonon dispersion spectra (Figure 3.5(g) and (h)). To our best knowledge, there are no reports of the synthesis of the two candidates. However, similar garnet of the formula $CaY_2Mg_2Si_3O_{12}$ has been synthesized[143], and Meng et al. discussed the effect of Mg-Si replacing Al-Al in $(Gd,Lu)_3Al_5O_{12}$[144], which suggests a possible synthesis route of $Ca(Er,Tb)_2Mg_2Si_3O_{12}$ from $Ca(Er,Tb)_2Al_5O_{12}$. These evidences,

Table 3.1: Candidates with E_{hull} = 0, E_{bg} between 3 and 4 eV, and thermal quenching within 15% (300K to 500K). The formulas marked in bold are final candidates.

Formula	E_{bg}^{DFT} (eV)	E_{bg}^{ML} (eV)	**TQ** (%)
$Ca_3Lu_2SiGe_2O_{12}$	3.47	3.54	6
$Ca_3Tm_2SiGe_2O_{12}$	3.50	3.47	7
$Ca_3Er_2SiGe_2O_{12}$	3.42	3.43	8
$Ca_3Zr_2SiGa_2O_{12}$	3.73	3.62	9
$Ca_3Sn_2SiAl_2O_{12}$	3.44	3.34	9
$Sr_3Y_2Ti_3O_{12}$	3.96	3.95	10
$CaY_2Mg_2Si_3O_{12}$	3.66	3.58	10
$\mathbf{CaEr_2Mg_2Si_3O_{12}}$	4.00	3.38	11
$Sr_3Tm_2Ti_3O_{12}$	3.96	3.89	11
$Sr_3Er_2Ti_3O_{12}$	3.96	3.87	11
$Ca_3Al_2SiGe_2O_{12}$	3.76	3.67	12
$Ca_3Zr_2GeAl_2O_{12}$	3.72	3.8	13
$Ca_3LuInGe_3O_{12}$	3.02	3.38	13
$Ca_3TmInGe_3O_{12}$	3.02	3.37	14
$Sr_3Zr_2SiGa_2O_{12}$	3.49	3.35	14
$Ca_3Tm_2Si_3O_{12}$	3.95	3.97	14
$MgCa_2Al_2Ge_3O_{12}$	3.39	3.5	14
$SmCa_2Zr_2Al_3O_{12}$	4.35	3.78	15
$SrCa_2Lu_2Ge_3O_{12}$	3.28	3.28	15
$Sr_3Lu_2Ti_3O_{12}$	3.97	3.88	15
$Ca_3ErInGe_3O_{12}$	3.00	3.36	15
$\mathbf{CaTb_2Mg_2Si_3O_{12}}$	3.89	3.28	15

both theoretical and experimental, suggest that these are two highly synthesizable hosts. The doping formation energy($E_f(Eu_M^{\times})$) was also calculated for the candidates. The results are 0.26 eV per Eu^{2+} and 0.82 eV per Eu^{2+} for $CaEr_2Mg_2Si_3O_{12}$ and $CaTb_2Mg_2Si_3O_{12}$, respectively. They are both higher than that of the experimentally discovered Eu^{2+}-doped $Sr_3Y_2Ge_3O_{12}$, which is 0.14 eV per Eu^{2+}. However, compounds with doping formation energy higher than 1 eV per dopant have been reported in previous work[136], suggesting probable dopability of our candidates.

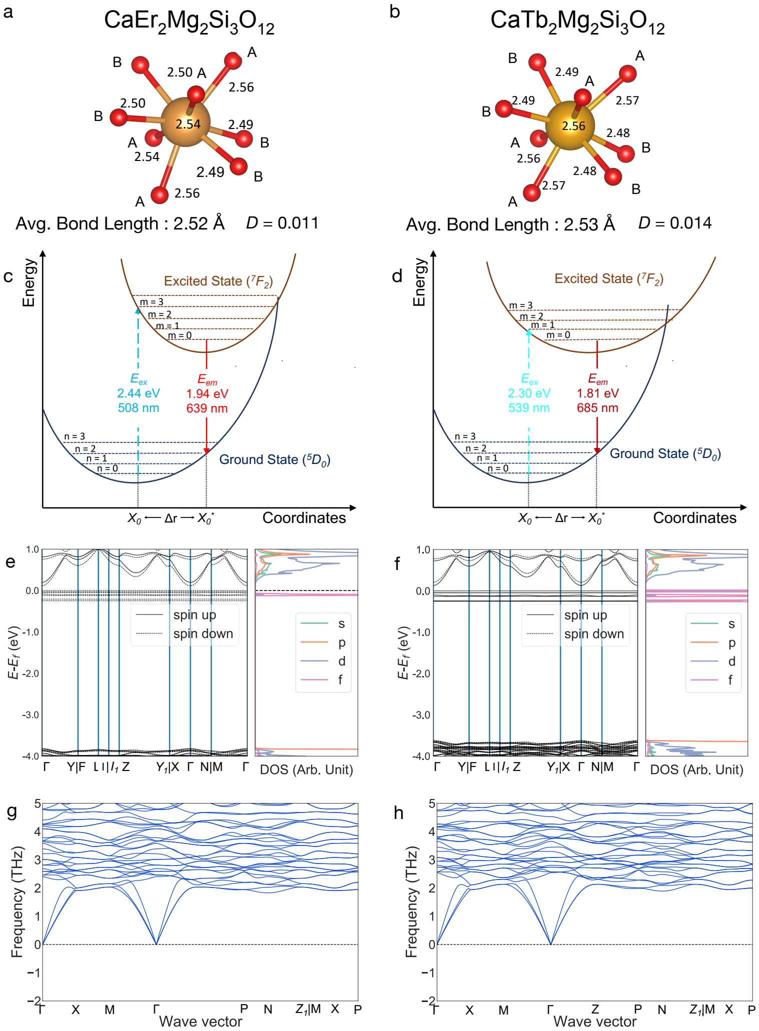


Figure 3.5: The DFT verification of candidates. (a) and (b) are the local environments of the Eu^{2+} in the candidates. The numbers around the bonds indicate the bond length in Å. Atoms marked with A and B represent two equivalent sets of sites in the dodecahedron coordination environment.(c) and (d) are the configurational coordinate diagrams for the Eu^{2+} in the candidates. Excitation is allowed from the vibrational level n = 0 of the ground state to the excited state and results in the excitation energy E_{ex}. The relaxation of the system from the lowest vibrational levels (m =0) of the excited state to the ground state results in the emission energy E_{em}. The displacement $\Delta r = X_0^* - X_0$ is the polyhedron average bond length difference between the excited and the ground states of Eu^{2+}. The CDFT calculated E_{ex} and E_{em} are shown in the diagrams. (e) and (f) are the bandstructures and density of the states of the Eu^{2+} phosphors. (g) and (h) are the phonon dispersion spectra.

3.4 Discussion

The interpretability of the ML models is important in two aspects. One is for the "debugging" of the model, and another is to provide new scientific insights. In our case, the model's interpretation adds to the value of the model by shedding light on the design of the garnets. According to the model, changing the D site species seems to be the most effective approach to engineer the bandgap. Decreasing Z_D, whether by replacing or mixing with lighter elements, increases the bandgap. For example, we find 3 cases out of the 22 candidates have $SiGe_2$ composition on the D site. They can be seen as doping a lighter element with smaller Z and NVE into a heavier element Ge, which increases the bandgap. Indeed, the PBE bandgaps of $Ca_3RE_2Ge_3O_{12}$ for RE = Lu, Tm, and Er are 3.32, 3.27, and 3.25 eV, respectively, all smaller than that of their Si-mixed counterparts by $\sim$ 0.2 eV. In general, species with smaller IE lowers the bandgap, and for species in the periodic table before lanthanides, the larger the Z, the smaller the bandgap. For example, Sr in C site could lead to smaller bandgap than Ca because Z_{Ca} is smaller than Z_{Sr} while the contributions of IE_C at IE_{Ca}(6.11 eV) and IE_{Sr}(5.69 eV) are similar. A similar rule applied to A site species as well.

Up to now, we have not proven that the inter- or extrapolations of the model for the finer grid of the compositions are valid. Therefore there could be room for fine-tuning of the candidates, especially on the Ca-Er/Tb ratios. For making the bandgap adjustment, mixing Al with Si can help increase the bandgap, while mixing with Ge for can lead to a decrease. We also noticed that for the two candidates, the excitation wavelengths, even though still in the cyan to the blue range, are higher than the ideal blue LED emission, which is 450 nm. The key to solving this mismatch is to tune the band curvature of the 5d band to enlarge the excitation energy while the emission energy remains in the range of red.

The necessity of hybridizing ML models into the high-throughput screening scheme has been illustrated in this work. Based on the statistics of the calculation of 1823 training data, the average CPU hours required to obtain the bandgap of the garnet primitive cell using a single node of Intel Xeon Phi ”Knight’s Landing” with 68 cores per node @ 1.4 GHz is 10 hours. Given the minimal compositions to be calculated by DFT, in this case, 5554, it would have taken about 78 months to finish by DFT solely. Now, with the assistance of ML models, such screening can be finished in weeks.

In this work, we have limited ourselves only to perform screening for novel phosphor materials. However, the developed models and even the whole workflow can work for other applications with minimum modifications. For example, screening for garnet photovoltaics is a suitable target, as the solar absorber materials also project a requirement for phase stability and bandgap on the candidates.

3.5 Conclusion

To conclude, we have developed an accurate, interpretable ML model that predicts the bandgap for garnet structures. Our model’s MAE is 0.13 eV, far below the common MAEs of 0.2 eV for structure-specific ML models. The feature selection was performed systematically and exhaustively to ensure an optimal and compact feature subset. These efforts lead to a highly interpretable model that makes physical and chemical sense and could effectively guide the design of novel materials. Furthermore, we integrated our two garnet models, targeting the phase stability and bandgap, respectively, to develop an ML-DFT hybrid high-throughput screening workflow to search for novel red-emission Eu^{2+}-doped garnet phosphors. Out of 20406 compositions, we identified two up-and-coming candidates, two candidates, $Ca(Er,Tb)_2Mg_2Si_3O_{12}$, which were verified theoretically to have emission in red, a substantial chance of synthesizability (both the host and the doped structure), and are predicted to have less than 15% thermal quenching from 300K to 500K. We believe it is a successful demonstration of accelerating the discovery of novel materials through statistical learning.

Chapter 3, in full, is under preparation for publication of the material ”High Throughput Screening of Eu^{2+}-Doped Red-Emission Garnet Phosphors Using Density Functional Theory and Machine Learning”, Weike Ye, Chi Chen, Mahdi Amachraa, Yunxing Zuo, Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 4

A universal machine learning model for elemental grain boundary energies

4.1 Introduction

Grain boundaries (GBs) play an important role in determining the strength, toughness, and corrosion resistance of materials[70,71]. A key property of a GB is its energy, which determines grain growth and the GB distribution. While the GB energy can be accurately calculated using electronic structure methods such as density functional theory (DFT) calculations, the requirement for relatively large supercells to model the inherently low symmetry GB structure limits such computationally intensive approaches to relatively small Σ GBs. Nevertheless, substantial databases of GB energies and other properties have been developed using high-throughput DFT. For example, the GB database (GBDB)[76] developed by the present authors contain the calculated GB energies and work of separation of more than 50 elemental metals for both tilt and twist GBs up to $\Sigma = 9$.

Alternatively, machine learning (ML) techniques have emerged as a means to develop models that can directly predict the GB energy from compositional and structural features[72–75]. However, existing models are limited in scope by chemistry or structure type, such as fcc Cu[72], Ni[73,74], or Al systems[75]. These limitations are a result of the choice of data source; these prior works have been developed using data sets computed using embedded atom method (EAM) potentials. While much less computationally intensive than DFT methods, EAM calculations are far less accurate, especially for non-fcc metals[76], and are available for only a limited subset of elements. Further, all these prior works rely on featurization approaches such as the Smooth Overlap of Atomic Positions (SOAP)[73,74] and the pair-correlation function (PCF)[75] that generates a large number of features (relative to the data set size) which do not provide direct interpretability.

In this letter, we outline a fundamentally different, physics-informed approach to developing a universal ML model for the GB energy of metals. We will demonstrate that the energy of small Σ GBs of metals can be predicted to within a mean absolute error (MAE) of 0.12 J m^{-2} using an eXtreme Gradient Boosting (XGBoost) model of four GB geometric features.

Extrapolation to high Σs GBs results in only a modest increase in MAE to 0.17 J m^{-2}.

4.2 Results

4.2.1 Normalization of E_{GB}

The starting point of this work is in re-evaluating the choice of target for our ML GB model. While prior works have attempted to directly predict the absolute GB energy, we do not believe this to be an optimal choice of target. The GB energy E_{GB} is the excess energy of the GB compared to the bulk per unit area, which can be obtained computationally as:

$$E_{GB} = \frac{E_{GB,supercell} - n \cdot E_{bulk}^{atom}}{2A} \tag{4.1}$$

where $E_{GB,supercell}$ is the energy of the supercell GB model, n is the number of atoms in the GB model, E_{bulk}^{atom} is the energy per atom of the bulk, A is the area of the GB and the factor of 2 accounts for the fact that there are two GBs per supercell model. E_{GB} is related to the energy necessary to break or stretch bonds at the GB from their bulk equilibrium configuration. This energy to stretch or break bonds scales with the cohesive energy of the metal E_{coh}[145] (see Figure C.1), which ranges from ~ 1.1 eV atom^{-1} for the alkali metals to ~ 8.9 eV atom^{-1} for tungsten. To remove this chemical scaling effect, we have elected to use the normalized GB energy $\hat{E_{GB}} = E_{GB}/E_{coh}$ as our choice of target.

4.2.2 Feature selection

Based on the coincident-site-lattice (CSL) theory[146,147], the GB can be specified at a macroscopic level by five degrees of freedom (DOF), namely two DOFs from the plane normal of the GB (or alternatively the Miller indices (hkl)), two DOFs from the rotation axis ($[uvw]$) and one DOF from the misorientation angle (θ). As integer Miller indices are non-optimal for a regression task, the (hkl) and $[uvw]$ were converted to the inter-planar distance of the GB plane (d_{GB}) and inter-planar distance of the normal plane to the rotation axis (d_{rot}), respectively (see Methods), and the cosine of the misorientation angle ($\cos(\theta)$) was used instead.

To these geometric GB features, we added three additional features related to bond stretching and breaking at the GB that are partially inspired by prior works in the literature. To describe the bond deformation, we used the average change in bond lengths between the GB supercell and its bulk conventional lattice, $\Delta(\bar{B}L) = \sum_{i=1}^{n}(BL_{GB}^{i} - BL_{0})/n$, where BL_{GB}^{i} is the bond length of the ith bond in the GB supercell, BL_0 is the bond length in the corresponding bulk conventional structure, and n is the number of bonds counted in the GB supercell. Here, the bonds are identified by performing a local environment analysis via a Voronoi tessellation-based algorithm implemented in the Python Materials Genomics (pymatgen) package[33]. A positive (negative) $\Delta(\bar{B}L)$ indicates overall bond stretching (compressing) at the GB. According to the Read-Shockley dislocation model[148], E_{GB} of GBs with small misorientation angles is proportional to the shear modulus G. Ratanaphan et al.[145] have also shown previously that the GB energies of bcc Mo and Fe are related to $G \cdot a_0$, where a_0 is the cubic lattice parameter. The multi-linear regression models developed by Zheng et al.[76] extended this conclusion to more bcc, face-centered cubic (fcc), and hexagonal closest packed (hcp) metals. Therefore, we include the Voigt-Reuss-Hill shear modulus G, and the bulk lattice parameter a_0 into the feature candidates. Figure 4.1 summarizes the initial set of six features considered in work.

A potential risk of domain-knowledge-driven feature selection is that some of the features may be correlated or redundant. For instance, G has a direct relationship with E_{coh}, which was used to normalize the GB energy. Therefore, we performed an exhaustive evaluation of all the 63 subsets of the initial 6 features (Figure4.1(a)). Figure 4.1(b) shows the performances of the optimal subset for feature subsets of each dimension, which shows that the model's performance converges at the number of features (n_f) of 4 when both the MAE(E_{GB})s of the training and the test data reach plateaus. We hence locate the optimal feature dimension at four, and the best feature subset (d_{GB}, cos(θ), a_0, $\Delta(\bar{B}L)$). Note that G, the only non-geometric feature, is excluded from the optimal subset, suggesting that the normalization scheme of the target is an effective strategy to shield most of the chemical scaling effect.

4.2.3 Model performance

The final ML model for E_{GB} was obtained by feeding the optimal feature set and normalized target into a tree-based pipeline optimization tool (TPOT)[149], as shown in Figure 4.2(a). To increase model flexibility, a polynomial transformation was performed on the four input features, resulting in a total of 14 compound features. Following this pipeline, we achieved the MAE(E_{GB}) of 0.06 and 0.12 J m^{-2} for the training and test data, respectively (Figure 4.2 (b)). The distribution of the normalized absolute errors shows 43 out of 53 elements have MAE(E_{GB})s less than 0.1 J m^{-2} (Figure 4.2(c)). Elements with the highest errors are such metals as Fe and Cr. The uncertainty in the magnetic ordering at the ground-state GB supercell of the two metals may lead to higher errors in the DFT calculations, hence higher errors for the models.

4.2.4 Model interpretation

One benefit of tree-based ensemble learning algorithms such as XGboost is the ease of retrieving the feature importance scores. However, in our case, the scores calculated from the XGBoost model should be taken skeptically due to the high correlations between the polynomial features (Figure C.3). To bypass the problem, we treated the pipeline as a whole, and calculated the permutation importance for the input four features instead. From Figure 4.3(b), we noticed that d_{GB} and cos(θ) are the two most important features. It agrees with the previous study[145] that the macroscopic geometry of the boundary plays an important role in determining the grain boundary energy. Furthermore, the feature d_{GB} being dominantly more important than cos(θ) echos with the conclusion drawn from Rohrer et al.[150] which states that variations in the grain

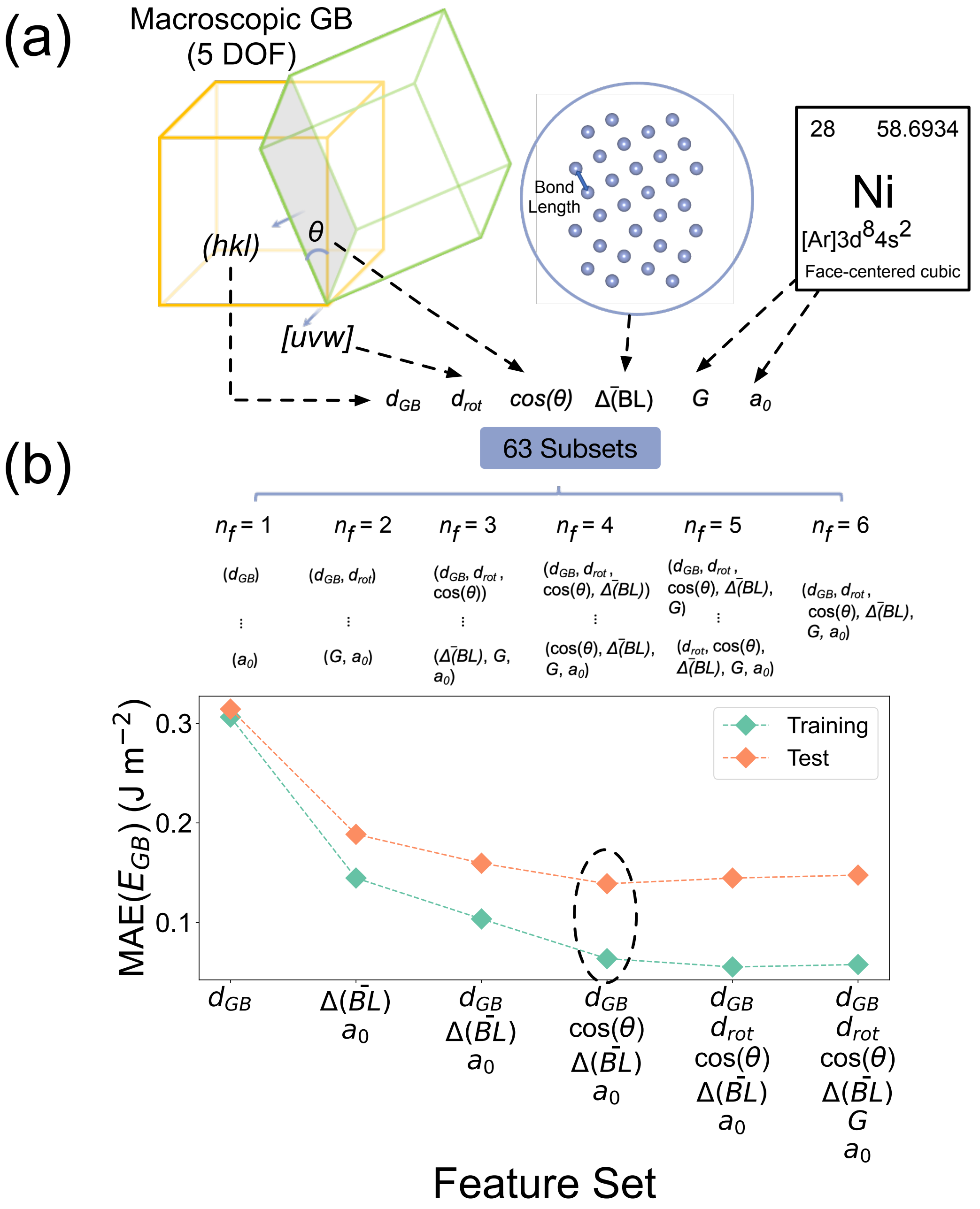


Figure 4.1: Feature engineering. (a) The knowledge-driven selection of initial feature candidates based on the macroscopic geometry, microscopic bonding environment in the GB supercell, and the corresponding elemental information. (b) The data-driven feature selection process. For the initial 6 features, there are in total of 63 feature subsets, which can be categorized by the number of features (n_f). The scatter plot shows the performances of the optimal subset in each category. The global optimal feature set is (d_{GB}, cos(θ), a_0, $\Delta(\bar{B}L)$), with which both the train and the test MAE(E_{GB}) reaches the plateau.

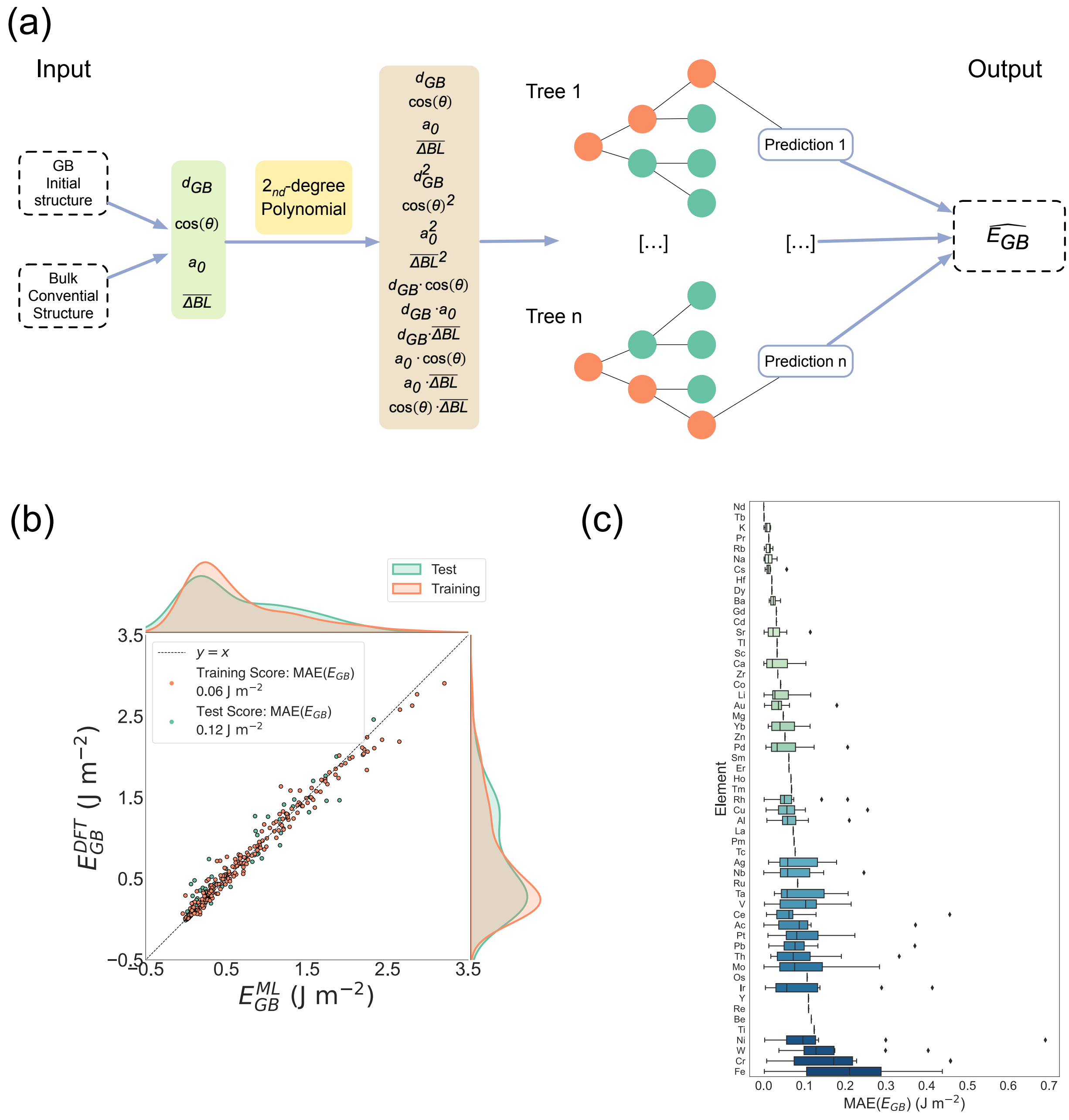


Figure 4.2: The pipeline and the performance. (a) The schematic illustration of the pipeline developed in this work. There are in total of 14 2_{nd}-degree polynomial terms associated with the optimized feature subset. The XGBoost model takes the 14 compounded features as input and output the predicted $\hat{E_{GB}}$. (b) The parity plot demonstrating the performance of the pipeline. The MAE(E_{GB}) for the training and the test data sets are 0.06 and 0.12 J m^{-2}, respectively. (c) The box plot of the normalized absolute error for each element. The elements are presented in the increasing order of the MAE(E_{GB}) from top to bottom.

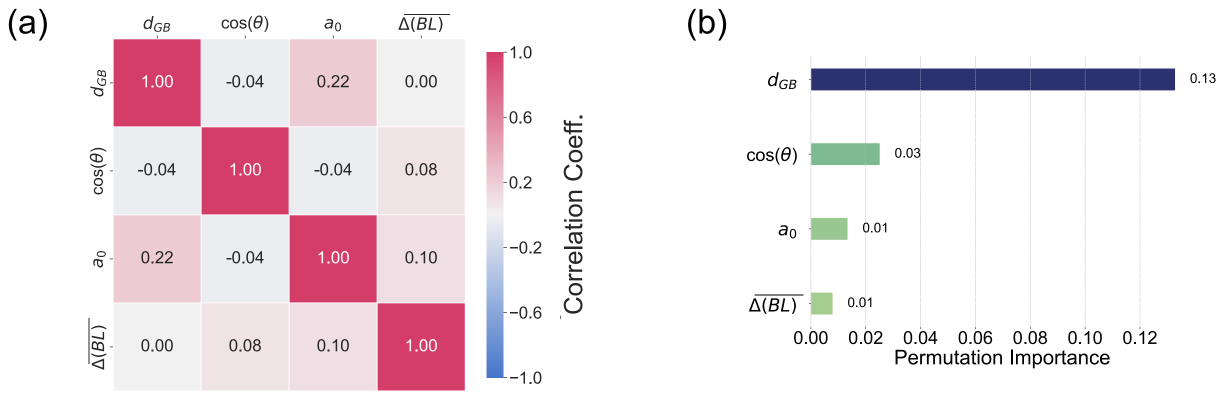


Figure 4.3: The feature importance analysis. (a) The correlation matrix for the optimized feature subsets. The four features can be considered independent. (b) The permutation feature importance calculated for the four input features of the pipeline.

boundary plane induce greater change in the energy than the variations in the misorientation. Such agreements between the model's behavior and the physical intuitions are strong evidence that the model has a solid grasp of the fundamental physics behind the grain boundary energy.

4.2.5 Model verification

It is often more important to explore the candidates outside of the current materials pool in actual materials science applications. It poses a challenge on the ML models to have great extrapolability towards the unknown structures. In our case, the model has only learned from data with a very limited range of low Σs (i.e., 3, 5, 7, 9) due to the limitation of the computation capacity. However, it is known that boundaries with larger fraction of coincident may have different properties compared to the ones with lower fraction due to the more severe deformation, making it necessary to test the extrapolability of the developed model on GBs with larger Σs. Therefore, we prepared an extrapolation test set, which contains 48 GBs of five elements (Ta, Pd, Cu, Pt, Li) with the Σ ranging from 17 to 66, far outside of the Σ range of the training data (Figure C.2). The model achieved a satisfactory MAE(E_{GB}) of 0.17 J m^{-2} on this data set, only a modest 0.05 J m^{-2} increase compared to the error of the test set, signaling a reliable extrapobility of the model.

Another evidence of the validity of the model is its qualitative reproducing of a well-acknowledged trend of GB energies, i.e., for fcc Ni, the symmetric twist boundaries that are joined by the widely-spaced (111), (100), and (110) planes have relatively low energies compared to that of GBs adopting other types[151,152]. Figure 4.4(b) shows the distribution of E_{GB}^{ML}s for a group of 76 GBs of fcc Ni, which contains 15, 19, and 6 symmetric twist GBs (STGBs) bounded by the planes of (111), (110), and (100), respectively, and 35 GBs of normal tilt or mixed GB types. The results show that the average energies of the three STGB categories are indeed lower than the energies of other GB configurations, especially (111) and (100) STGBs. Note that 69 out of the 79 GBs have the supercells containing more than 200 atoms, including 17 with more than 1000 atoms, making them computational prohibitive and thus impossible to determine the accuracy quantitatively. Nevertheless, it showcases the model's capability to qualitatively reproduce the well-documented energy trend of GBs outside of the range of training data and could potentially serve as a solution to the scaling difficulty of DFT.

4.3 Discussion

In this work, we found that normalizing the grain boundary energy by the elemental cohesive energy could reduce the chemical scaling effect. It suggests that the chemical influence on the grain boundary energy is dominated by E_{coh}, which agrees with the order-of-magnitude energetic analysis showing that E_{coh}'s contribution outweighs that of the G by almost a magnitude[153]. However, the extension of the normalization strategy beyond elemental systems is unlikely to succeed due to the more complex chemical interactions. We suspect that for alloys, the formation energy of the heterogeneous bonds plays a non-negligible role in the anisotropy of the E_{GB}[153].

It is worth mentioning that d_{rot} was also excluded from the final feature set. While it may

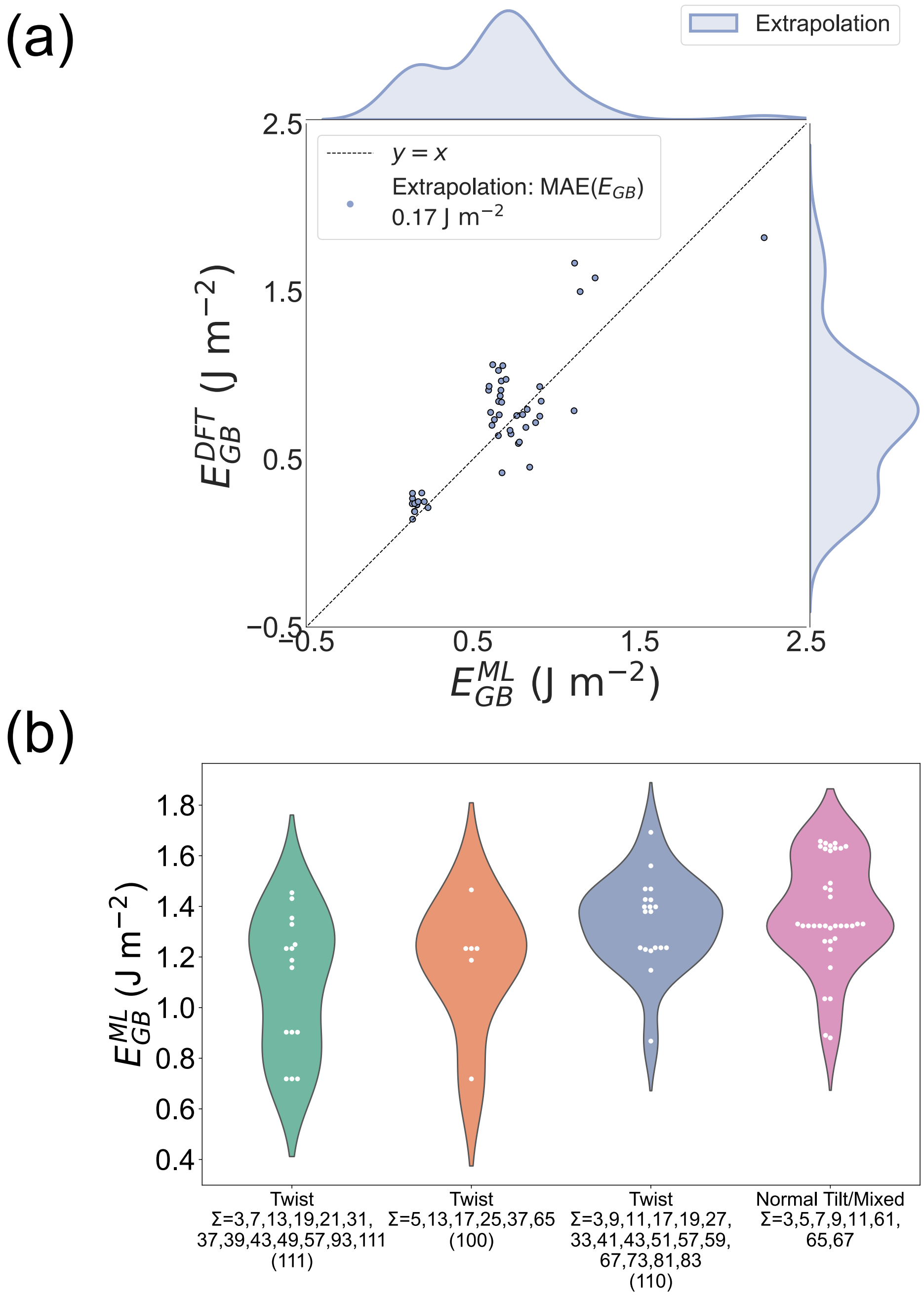


Figure 4.4: (a) The parity plot illustrating the model's performance on the extrapolation test data set. The MAE(E_{GB}) is 0.17 J m^{-2}, merely 0.05 J m^{-2} higher than the test MAE(E_{GB}).(b) The distributions of E_{GB}^{ML} of fcc Ni Σ3-111 (111) STGBs , Σ5-65 (100) STGBs, Σ5-65 (110) STGBs, and Σ3-67 normal tilt or mixed GBs. The four categories are arranged from left to right in the order of increasing mean energies. STGBs bounded by the (111) and (100) planes have noticeable lower mean energies compared to GBs of other configurations.

be a result of the fact that the grain boundary plane outweighs the misorientation in affecting E_{GB}[150], it is also likely to be a result of the limitation in our data. As illustrated in the Methods section, the MMI of the rotation axis is ≤ 1 for all the DFT-computed GBs in this work. More specifically, there are in total only four axes considered, i.e., [110], [100], [111], and [0001]. The low variance in d_{GB} makes it almost trivial for the model's performance. Future improvements can be made by creating GBs with a broader range of rotation axes.

4.4 Methods

4.4.1 Collection of data

The GB data used in this work were obtained from two sources. The first and the major part comes from the GBDB[76], which contains the energies of 316 GBs of 53 elements in fcc, bcc, hcp and double-hcp (dhcp) structures, after excluding Lu, Eu, and Hg due to the unavailability of the bulk elastic data. The Σs of the GBs range from 3 to 9. The upper limits of the maximum Miller index (MMI) for the rotation axis, and the grain boundary plane are 1 and 3, respectively. Interested readers are referred to ref 76 for the details on the GB structure generation and computational methods. The second part of GB data is from our calculations using the same computational methods as the previous work[76]. We calculated the energies for another 53 GBs of elements Ta, Pd, Cu, Pt and Li, which were generated by extending the limit of the Σ to 66, and the MMI of the grain boundary plane to be ≤ 8, while keeping the MMI of the rotation axis to be ≤ 1.

In the total of available 369 GBs, 321 GBs with $\Sigma \leq 9$ were used for the model development, which were divided into the training (258 GBs) and the test (63 GBs) set. The training data was selected by randomly sampling 80% of the GBs from elements with more than one GB entries, and including all the GBs from elements with single GB entry. The remaining 48 GBs with Σ ranging from 17 to 66 were specifically used to test the extrapolability of the model.

Details of the distributions of the chemistry and Σ for each data subsets can be found in Figure 3.1.

In order to test the model's ability to reproduce qualitatively the well-acknowledged trend of lower energies of the STGBs correspond to the three widely-spaced (111), (110), and (100) boundary planes for fcc Ni[151,152], we prepared a set of 76 Ni fcc GBs with wide variations in GB types. This data set contains 15 STGBs bounded by the (111) plane, 19 STGBs bounded by the (110) plane, 6 STGBs bounded by the (100) plane, 15 normal tilt GBs, and 20 mixed GBs. The Σ of the prepared data set ranges from 3 to 111, and the upper limit of the MMIs of the rotation axis and the grain boundary plane is 1 and 6, respectively. To model GBs with large Σs and joined by more closely-spaced planes, the supercells are usually too large in size to perform DFT calculations. For example, 69 out of the 79 GBs have supercells containing more than 200 atoms, including 17 with more than 1000 atoms. Therefore, this data set is only prepared for observing the qualitative trend of the E_{GB}^{ML}.

4.4.2 Inter-planar distance

The features d_{GB} and d_{rot} are inter-planar distances of the grain boundary plane ((hkl)) and the normal plane to the rotation axis ($[uvw]$). The formulas of d_{GB} and d_{rot} are as follows:

$$\begin{aligned}\frac{1}{d_{GB}^2} &= \frac{1}{d_{hkl}^2}\\ &= \frac{h^2+k^2+l^2}{a^2}(\text{cubic crystals})\\ &= \frac{4}{3}\frac{h^2+hk+k^2}{a^2}+\frac{l^2}{c^2}(\text{hexagonal crystals})\end{aligned} \tag{4.2}$$

$$
\begin{aligned}
\frac{1}{d_{rot}^2} &= \frac{1}{d_{uvw}^2} \\
&= \frac{u^2+v^2+w^2}{a^2}(\text{cubic crystals}) \\
&= \frac{4}{3}\frac{u^2+uv+v^2}{a^2}+\frac{w^2}{c^2}(\text{hexagonal crystals})
\end{aligned}
\tag{4.3}
$$

where a and c are the crystal lattice constants of the bulk conventional crystal. Note that $[uvw]$ used here should be normalized Miller indices of the rotation axis. For hexagonal systems, the Miller indices of the planes are first converted from the 4-index notation to the 3-index notation to calculate d_{GB} and d_{rot}.

4.4.3 Model development

The optimized machine learning pipeline was selected with the aid of a tree-based pipeline optimization tool (TPOT)[149]. Briefly, machine learning pipelines can be represented by binary expression trees with ML operators as primitives. TPOT automatically generates and optimizes the ML pipelines based on the accuracy and the complexity using genetic programming. In the current implementation of TPOT (`https://github.com/EpistasisLab/tpot`), the ML operators include a wide range of algorithms implemented in scikit-learn[52] and other advanced algorithms such as XGBoost[139]. In this work, we set the population size, the generations, and the offspring size at 50, 10, and 50, respectively, to allow for the evaluation of a total of 550 × (50 + 10 × 50) pipelines by TPOT.

The optimized model pipeline found by TPOT[149] is an XGBoost model preceded by a polynomial feature preprocessing step. The *learning rate*, *max depth*, *n estimator*, and *min child weight* are 0.1, 5, 100 and 7, respectively, for optimal learning ability. The *subsample ratio* is set at 0.7 to regulate over-fitting. Default values are used for all other hyper-parameters of the XGBoost model.

To inspect the model's behavior, we calculated the permutation feature importance. It is calculated by randomly shuffling the values of one feature, then calculating the decrease in the score of the model[154]. The calculation was performed by using the permutation importance method implemented in the open-source scikit-learn[52] package.

Chapter 4, in full, is under preparation for publication of the material ”A Universal Machine Learning Model for Elemental Grain Boundary Energies”, Weike Ye, Hui Zheng, Chi Chen, Shyue Ping Ong. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Summary and Outlook

In the first work, we have developed deep neural network models that are able to predict the DFT formation energies of garnets and perovskites to within 7-34 meV atom^{-1}. The substantially lowered MAE compared to existing works is achieved by using only the electronegativity and ionic radius of the species on each symmetrically distinct site as features, and the formation energy referenced to the binary oxides as the target. By introducing a binary encoding scheme, the models were successfully extended from the unmixed garnets and perovskites to mixed compositions, opening the applicability of the models to the vast unexplored chemistry spaces. Finally, we have shown that these models can be used to classify stable/unstable garnet and perovskite compositions with $\geq 80\%$ accuracy.

In the second work, we have developed an accurate, interpretable ML model that predicts the bandgap for garnet structures. Our model's MAE is 0.13 eV, far below the common MAEs of 0.2 eV for structure-specific ML models. The feature selection was performed systematically and exhaustively to ensure the most economic feature set. These efforts lead to a highly interpretable model that makes physical and chemical sense and could effectively guide the design of novel materials. Furthermore, we integrated our two garnet models to develop an ML-DFT hybrid high-throughput screening workflow, which led to the discovery of two candidates, $\mathrm{Ca(Er,Tb)_2Mg_2Si_3O_{12}}$, with desired properties. We believe it is a successful demonstration of applying ML to accelerate materials discovery.

In the third work, we successfully developed an accurate, universal machine learning pipeline for predicting the grain boundary energy across a wide variety of elemental systems. The model was trained on a data set based on the first-principles calculation that is more accurate and has broader chemistry scope than the data calculated using the EAM potential. We showcases that only four geometric features, d_{GB}, cos(θ), a_0 and $\Delta(\bar{B}L)$, are enough to predict the grain boundary energy after normalization by the bulk cohesive energy. The model developed achieves a test error of 0.12 J m^{-2}, and demonstrates great extrapolability to larger Σs with a modest increase of MAE to 0.17 J m^{-2}. The model also successfully identified the low energy GB types in the fcc

Ni system which agrees with the trend discovered experimentally. To conclude, the model's high accuracy and superior generalizability make it a potential surrogate of DFT calculations and a significant enhancement to DFT's accessibility.

To conclude, we have successfully developed highly interpretable ML models that predict the essential materials properties including the phase stability, the bandgap, and the grain boundary energy to the state-of-the-art accuracy. In all the models, we showcase the identification of optimal features with much more compact length compared to existing works that uses pure data-driven approach to perform feature selection. We also provide knowledge-driven engineering of the learning target that allows the more efficient learning and broadened applicability. Furthermore, we present real application of the developed models in accelerating materials discovery by devising ML-DFT hybrid HTS workflow and identifying novel phosphors with desired emission. Last but not least, in the process of the model development and verification, important insights are revealed in understanding the fundamental physics and the materials design. Meanwhile, we also notice that there are possible avenues for future work. To name a few, the compositional features that used in the first work are not extendable across structure types; the error with GGA bandgap limited its capability in narrowing the candidates pool; lastly, the simple normalization of the grain boundary energy in the third work can hardly extend to alloy systems.

Appendix A

Supporting information: Deep neural networks for accurate predictions of crystal stability

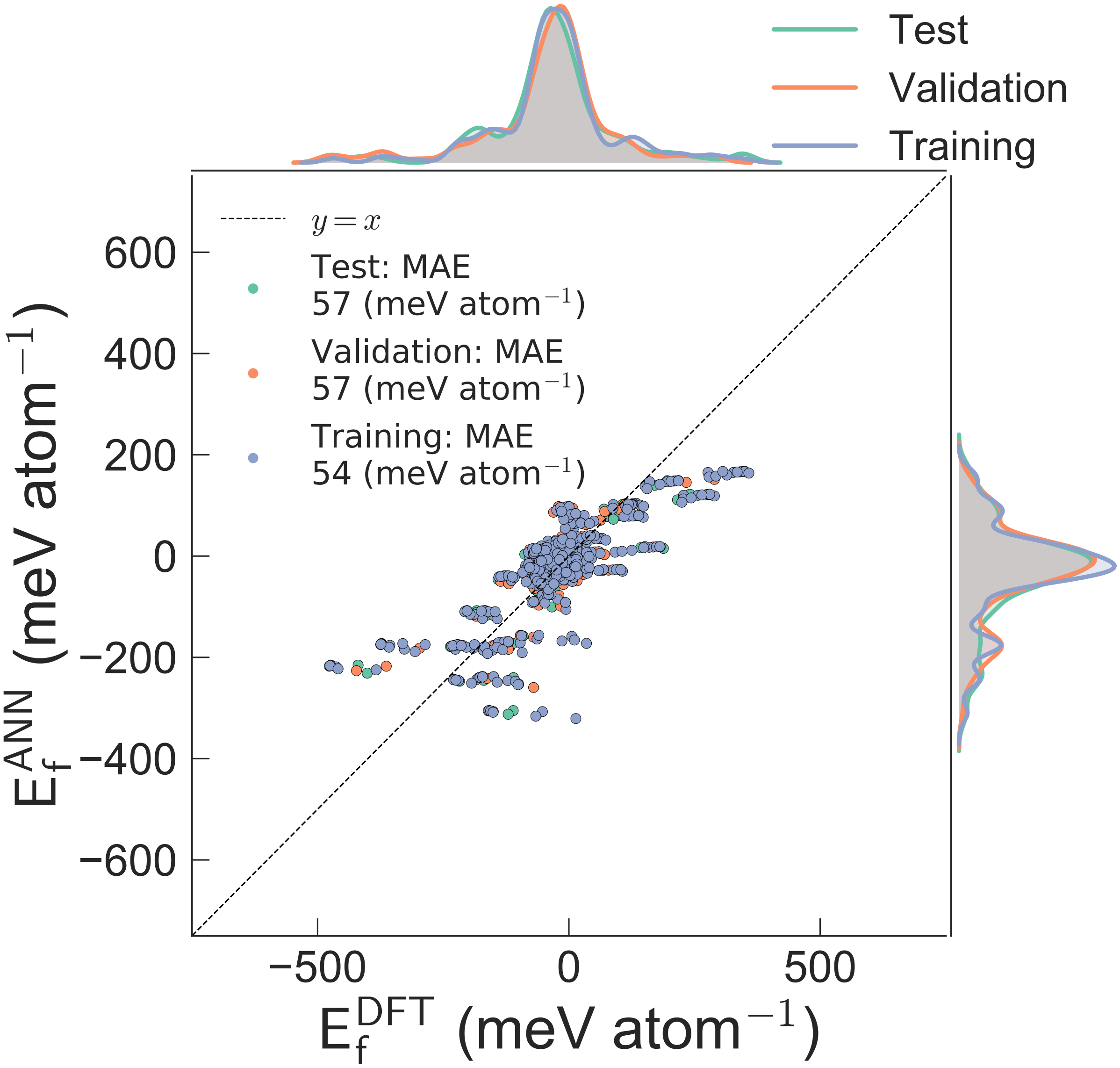


Figure A.1: Performance of multiple linear regression model on E_f^{DFT} of unmixed garnets. The high training, validation and test mean absolute errors (MAEs) of 54, 57 and 57 meV atom^{-1} indicate that a simple linear functional form is insufficient to model the relationship between E_f^{DFT} and the Pauling electronegativity and ionic radii descriptors. The R^2 for training, validation and test data are 0.63, 0.63 and 0.63, respectively. The black line (dashed) in the figure is the identity line serving as reference.

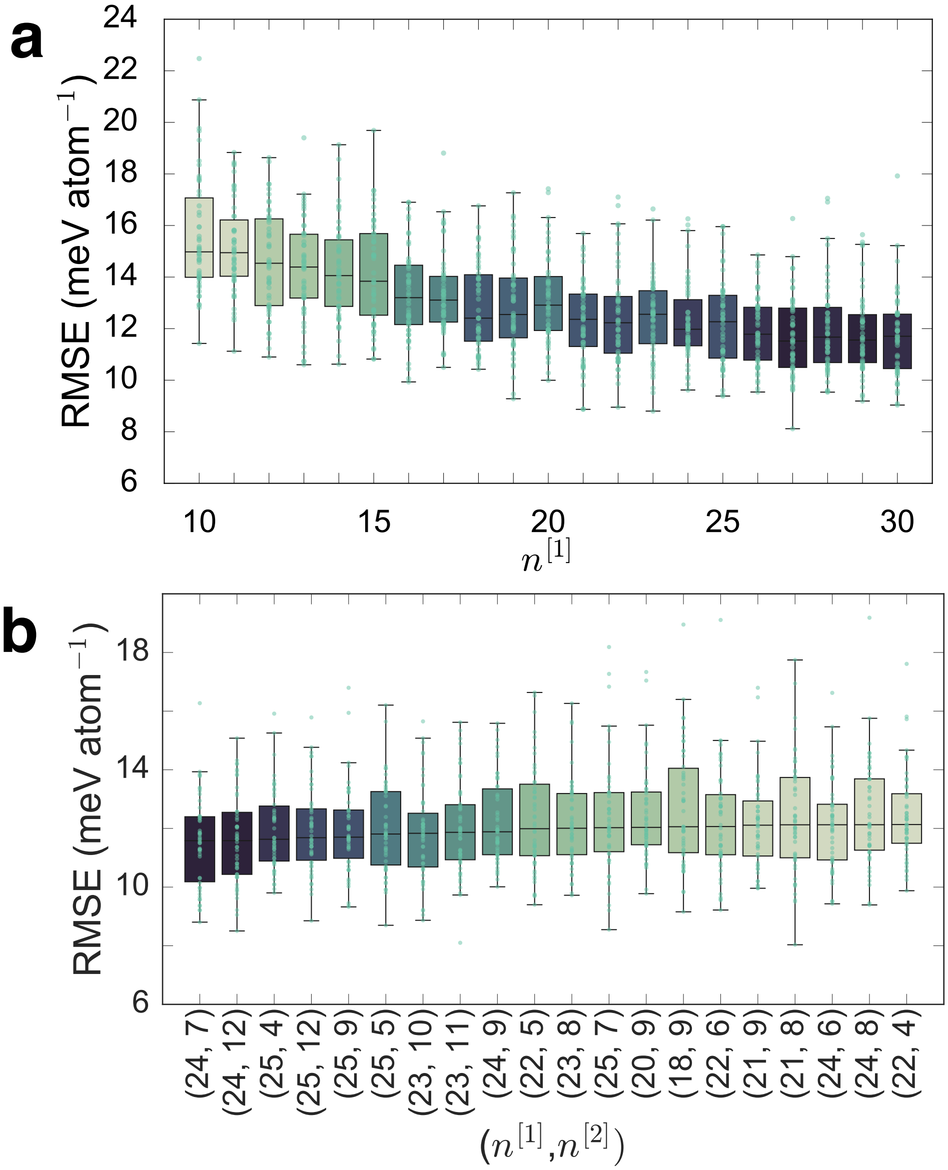


Figure A.2: Optimization of artificial neural network (ANN) architecture. a, Plot of the root mean square error (RMSE) loss metric versus number of neurons in a single-hidden-layer ANN model. The RMSE converges at $n^{[1]}$ ~20, and the smallest standard deviation is observed at $n^{[1]}$=24. b, Plot of the RMSE loss metric versus number of neurons in a two-hidden-layer deep neural network (DNN) model for unmixed garnets. Only the 20 best-performing models are shown for brevity. The RMSE loss metric achieved by the DNN model is similar to that of the single-hidden-layer ANN model. The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution.

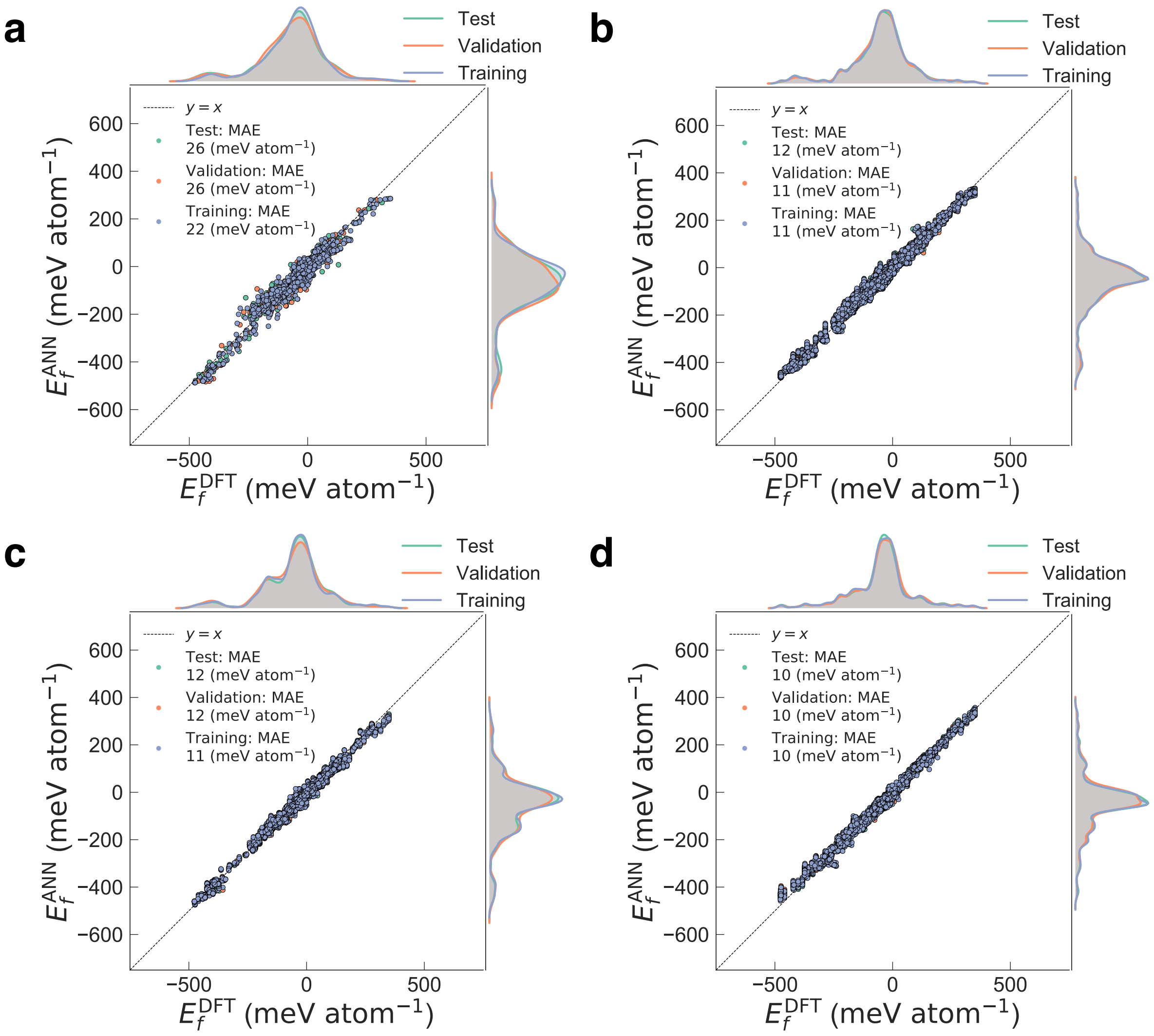


Figure A.3: Performance of optimized artificial neural network models for garnets. Plot of E_f^{NN} against E_f^{DFT} for a. “averaged” ANN model trained on all unmixed and mixed garnets, b. ordered DNN model trained on unmixed garnets with C-mixed garnets(standard deviations of E_f^{DFT} for training, validation and test set are: 130, 128 and 130 meV atom^{-1}), c. ordered DNN model trained on unmixed garnets with A-mixed garnets (standard deviations of E_f^{DFT} for training, validation and test set are: 132, 134 and 131 meV atom^{-1}), and d. ordered DNN model trained on unmixed garnets with D-mixed garnets (standard deviations of E_f^{DFT} for training, validation and test set are: 126, 126 and 127 meV aatom^{-1}). The black lines (dashed) in all subfigures are the identity lines serving as references.

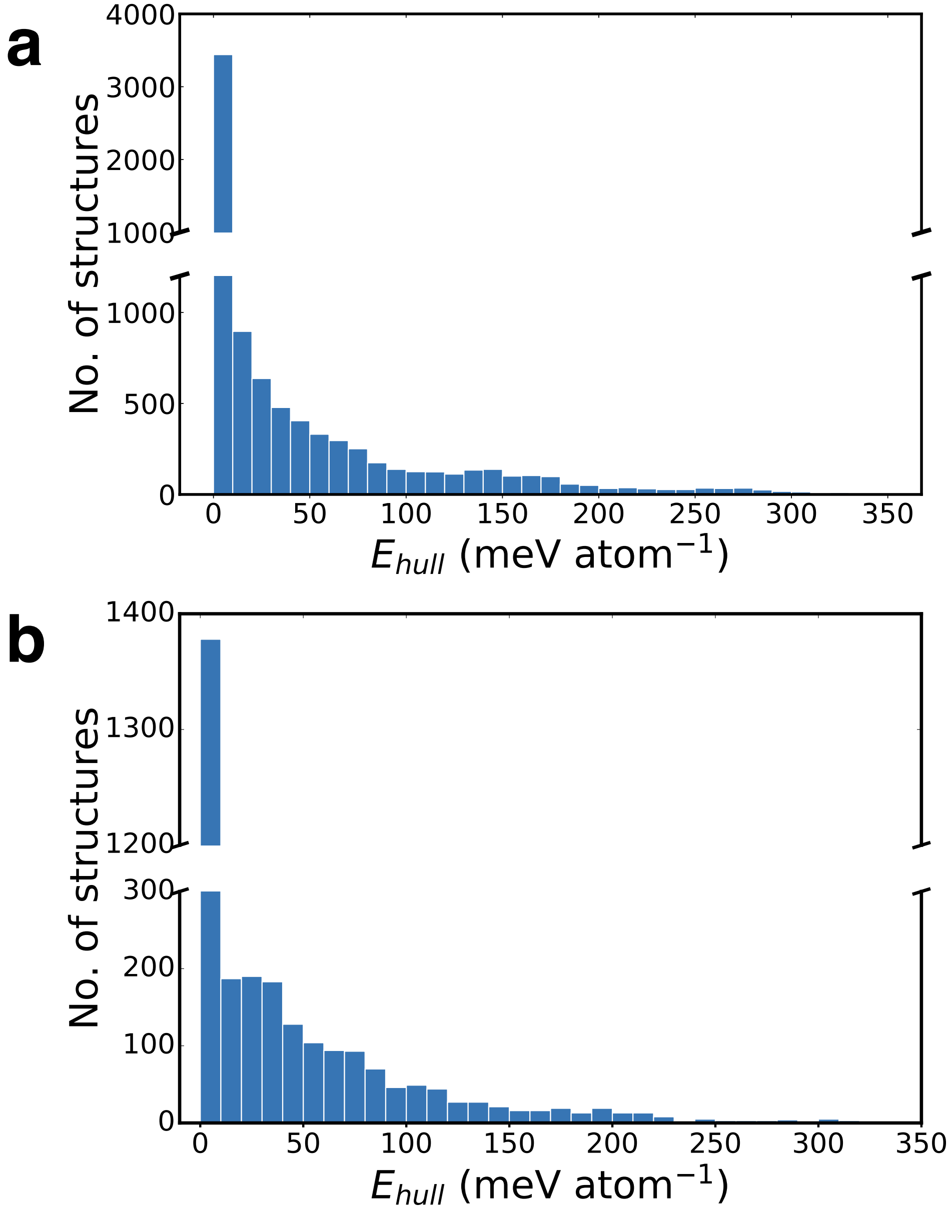


Figure A.4: Histograms of E_{hull} predicted using the optimized neural network models for garnets and perovskites. a. A total of 8,427 garnet compositions were generated based on 2:1 mixing on the C or D sites, or 1:1 mixing on the A site. Only the ordering with the lowest E_{hull} is presented at each composition. Of the 8,385 compositions, 2,307 compositions are predicted to have E_{hull}=0 meV atom−1. b. A total of 2,791 perovskite compositions were generated based on 1:1 mixing on the A or D sites. Only the ordering with lowest E_{hull} is presented at each composition. Of the 2,791 compositions, 1,147 compositions are predicted to have E_{hull} = 0 meV atom−1.

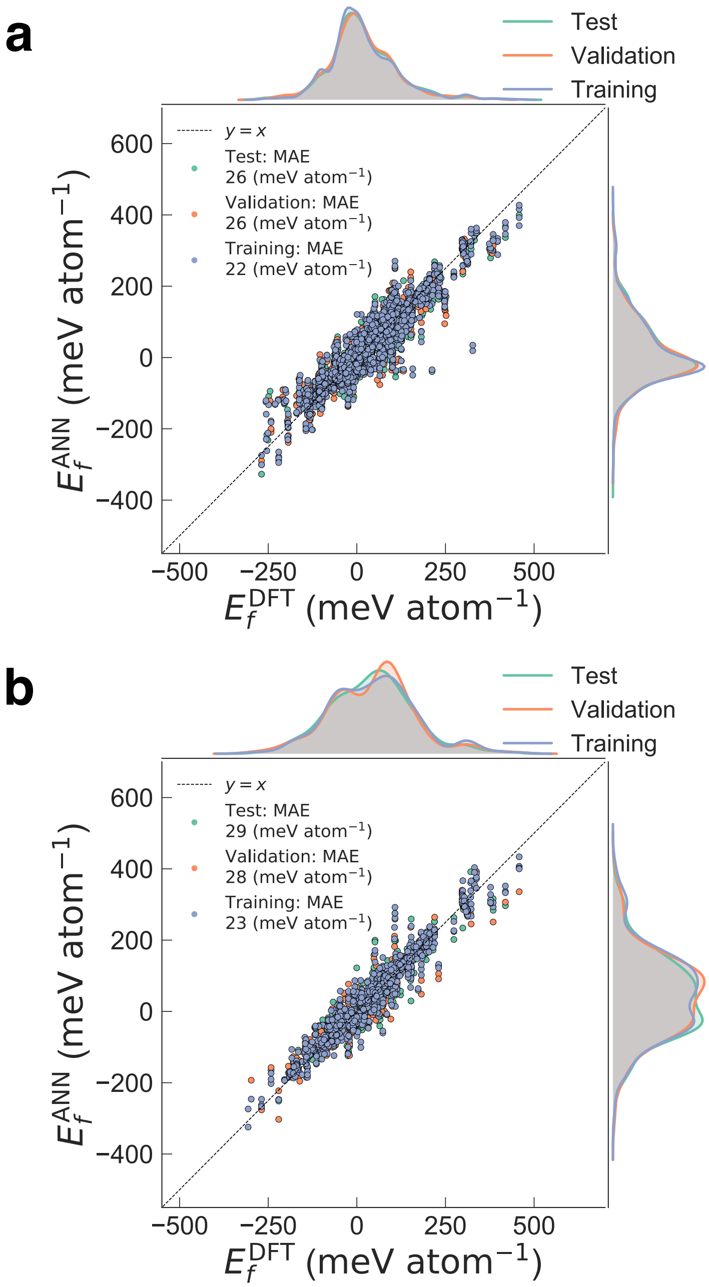


Figure A.5: Performance of optimized artificial neural network models for perovskites. Plot of E_f^{ANN} against E_f^{DFT} for a. ordered ANN model trained on unmixed with A-mixed perovskites (standard deviation of E_f^{DFT} for training, validation and test sets are: 95, 94 and 96 meV atom^{-1}), and b. ordered ANN model trained on unmixed with B-mixed perovskites (standard deviations of E_f^{DFT} for training, validation and test sets are: 121, 117 and 115 meV atom^{-1}). The black lines (dashed) in a. and b. are the identity lines serving as references.

Table A.1: Binary oxides used as reference states for E_f calculations.

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
Rh	3	Rh_2O_3	181829	mp-542734
Ga	3	Ga_2O_3	166198	mp-886
Sc	3	Sc_2O_3	647397	mp-216
Nd	3	Nd_2O_3	645664	mp-1045
Au	3	Au_2O_3	8014	mp-27253
B	3	B_2O_3	36066	mp-306
Mn	3	Mn_2O_3	76087	mp-542877
Hf	4	HfO_2	27313	mp-352
Zr	4	ZrO_2	68782	mp-2858
Ge	4	GeO_2	92551	mp-470
Ti	4	TiO_2	69331	mp-2657
Si	4	SiO_2	200726	mp-7000
Ru	4	RuO_2	56007	mp-825
Sn	4	SnO_2	39173	mp-856
Pt	4	PtO_2	647320	mp-1285
Mo	4	MoO_2	36263	mp-510536
Re	4	ReO_2	24060	mp-7228
Se	4	SeO_2	412234	mp-726
Te	4	TeO_2	26844	mp-2125
In	3	In_2O_3	181833	mp-22598
Tc	4	TcO_2	173153	mp-33137
Ir	4	IrO_2	640887	mp-2723

Continued on next page

Table A.1 – continued from previous page

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
Os	4	OsO_2	30400	mp-996
Nb	5	Nb_2O_5	25750	[1]
P	5	P_2O_5	40865	mp-562613
Sb	5	Sb_2O_5	1422	mp-1705
Ta	5	Ta_2O_5	[2]	mvc-4415
As	5	As_2O_5	10015	mp-555434
V	5	V_2O_5	40488	mp-25620
W	6	WO_3	50728	mp-19342
Fe	3	$\alpha-Fe_2O_3$	161283	mp-24972
Fe	2	FeO	633029	mp-18905
Ag	1	Ag_2O	173984	mp-353
Al	3	Al_2O_3	60419	mp-1143
Au	3	Au_2O_3	8014	mp-27253
As	5	As_2O_5	10015	mp-555434
Ba	2	BaO	616004	mp-1342
Bi	3	Bi_2O_3	15072	mp-23262
Ca	2	CaO	60704	mp-2605
Cd	2	CdO	181057	mp-1132
Ce	3	Ce_2O_3	96202	mp-542313
Ce	4	CeO_2	164225	mp-20194
Co	2	CoO	9865	mp-19079
Co	3	Co_2O_3		mvc-852
Cr	3	Cr_2O_3	201102	mp-19399

Continued on next page

Table A.1 – continued from previous page

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
Cr	4	CrO_2	166021	mp-19177
Cs	1	Cs_2O	27919	mp-7988
Cu	2	CuO	653723	mp-1692
Dy	3	Dy_2O_3	96208	mp-2345
Er	3	Er_2O_3	39521	mp-679
Eu	3	Eu_2O_3	40472	[1]
Fe	2	FeO	633029	mp-18905
Fe	3	Fe_2O_3	161283	mp-24972
Fe	4	FeO_2		mp-850222
Ga	3	Ga_2O_3	34243	mp-886
Gd	3	Gd_2O_3	152449	mp-504886
Ge	4	GeO_2	158592	mp-470
Hf	4	HfO_2	172165	mp-352
Hg	2	HgO	40316	mp-1224
Ho	3	Ho_2O_3	44516	mp-812
I	5	I_2O_5	182672	mp-23261
In	3	In_2O_3	640179	mp-22598
Ir	4	IrO_2	84577	mp-2723
K	1	K_2O	44674	mp-971
La	3	La_2O_3	96201	mp-2292
Li	1	Li_2O	54368	mp-1960

Continued on next page

[1] There is no corresponding entry in MP. The energy was obtained by applying DFT calculation on the structure using MP-compatible parameters.

Table A.1 – continued from previous page

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
Lu	3	Lu_2O_3	642477	mp-1427
Mg	2	MgO	41990	mp-1265
Mn	2	MnO	28898	mp-714882
Mn	3	Mn_2O_3	9091	mp-542877
Mn	4	MnO_2	20227	mp-19395
Mo	4	MoO_2	99714	mp-510536
Na	1	Na_2O	180570	mp-2352
Nb	4	NbO_2	35181	mp-557057
Nb	5	Nb_2O_5	25750	
Nd	3	Nd_2O_3	645664	mp-1045
Ni	2	NiO	61318	mp-19009
Os	4	OsO_2	30400	mp-996
P	5	P_2O_5	40865	mp-562613
Pb	2	PbO	99777	mp-672237
Pb	4	PbO_2	43460	mp-20725
Pd	4	PdO_2	647283	mp-1018886
Pd	2	PdO	29281	mp-1336
Pr	3	Pr_2O_3	96203	mp-16705
Pr	4	PrO_2	647300	mp-1302
Pt	2	PtO	164290	mp-7947
Pt	4	PtO_2	647320	mp-1285
Pu	4	PuO_2	55456	mp-1959
Rb	1	Rb_2O	180572	mp-1394

Continued on next page

Table A.1 – continued from previous page

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
Re	4	ReO_2	24060	mp-7228
Rh	3	Rh_2O_3	108941	mp-542734
Sc	3	Sc_2O_3	647397	mp-216
Se	4	SeO_2	59712	mp-726
Si	4	SiO_2	200726	mp-7000
Sm	3	Sm_2O_3	647461	mp-218
Sn	4	SnO_2	39173	mp-856
Sr	2	SrO	180194	mp-2472
Tc	4	TcO_2	173152	mp-33137
Ta	5	Ta_2O_5	[2]	mvc-4415
Tb	3	Tb_2O_3	40474	mp-1056
Tb	4	TbO_2	647500	mp-2458
Te	4	TeO_2	26844	mp-2125
Ti	3	Ti_2O_3	77696	mp-458
Ti	4	TiO_2	202240	mp-2657
Tl	1	Tl_2O	16220	mp-27484
Tl	3	Tl_2O_3	74090	mp-1658
Tm	3	Tm_2O_3	647581	mp-1767
V	3	V_2O_3	260212	mp-25787
V	4	VO_2	1504	mp-19094
V	5	V_2O_5	99808	mp-25620

Continued on next page

[2] This structure is not included in ICSD, but the DFT calculation from MP shows that it has a calculated formation energy of -23.489 eV per formula unit(fu), which is close to reported experimental value (-21.209 eV per fu)[155]

Table A.1 – continued from previous page

Element	Oxidation State	Binary Oxide	ICSD ID	Materials ID
W	4	WO_2	8217	mp-19372
Y	4	Y_2O_3	23811	mp-2652
Yb	3	Yb_2O_3	62872	mp-2814
Zn	2	ZnO	647681	mp-2133
Zr	4	ZrO_2	172161	mp-2858

Table A.2: Species on the C, A and D sites in garnet, adapted from ref. 1

Site	Ions
C	Ba^{2+}, Na^{+}, Sr^{2+}, Ca^{2+}, Tb^{3+}, La^{3+}, Pr^{3+}, Nd^{3+}, Sm^{3+}, Gd^{3+}, Eu^{3+}, Dy^{3+}, Y^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, Lu^{3+}, Hf^{4+}, Mg^{2+}, Zr^{4+}, Zn^{2+}, Cd^{2+}, Bi^{3+}
A	Dy^{3+}, Y^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, Lu^{3+}, Hf^{4+}, Mg^{2+}, Zr^{4+}, Sc^{3+}, Ta^{5+}, Ti^{4+}, Nb^{5+}, Al^{3+}, Zn^{2+}, Cr^{3+}, In^{3+}, Ga^{3+}, Sn^{4+}, Ge^{4+}, Sb^{5+}, Ru^{4+}, Rh^{3+}
D	Li^{+}, Ti^{4+}, Al^{3+}, Ga^{3+}, Si^{4+}, Sn^{4+}, Ge^{4+}, As^{5+}, P^{5+}

Table A.3: Species on the A and B sites in perovskites

Site	Ions
A	Ba^{2+}, Sr^{2+}, Ca^{2+}, La^{3+}, Tb^{3+}, Ce^{3+}, Ce^{4+}, Pr^{3+}, Nd^{3+}, Sm^{3+}, Gd^{3+}, Dy^{3+}, Y^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, Mg^{2+}, Sc^{3+}, Mn^{2+}, Al^{3+}, Tl^{3+}, Zn^{2+}, Cd^{2+}, Ni^{2+}, Sn^{4+}, Bi^{3+}, Pd^{2+}, Pt^{2+}, Rh^{3+}, Pb^{2+}
B	La^{3+}, Tb^{3+}, Ce^{3+}, Ce^{4+}, Pr^{3+}, Nd^{3+}, Sm^{3+}, Eu^{3+}, Gd^{3+}, Dy^{3+}, Y^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, Lu^{3+}, Hf^{4+}, Mg^{2+}, Zr^{4+}, Sc^{3+}, Ta^{5+}, Ti^{4+}, Mn^{2+}, Mn^{4+}, Al^{3+}, Tl^{3+}, V^{5+}, Cr^{3+}, In^{3+}, Ga^{3+}, Fe^{2+}, Fe^{3+}, Co^{2+}, Co^{3+}, Cu^{2+}, Re^{4+}, Si^{4+}, Tc^{4+}, Ni^{2+}, Sn^{4+}, Ge^{4+}, Bi^{3+}, Mo^{4+}, Ir^{4+}, Os^{4+}, Pd^{4+}, Ru^{4+}, Pt^{4+}, Rh^{3+}, Pb^{4+}, W^{4+}, Au^{3+}

Table A.4: Accuracy of DFT formation energies versus experiments.

Formula	E_f^{EXP} (meV atom^{-1})	E_f^{DFT} (meV atom^{-1})	**Source**
$Dy_3Al_5O_{12}$	-51(977 K)	-54	Ref. 156
$Ho_3Al_5O_{12}$	-53(977 K)	-51	Ref. 156
$Er_3Al_5O_{12}$	-50(977 K)	-49	Ref. 156
$Tm_3Al_5O_{12}$	-50(977 K)	-46	Ref. 156
$Lu_3Al_5O_{12}$	-38(977 K)	-37	Ref. 156
$Y_3Al_5O_{12}$	-60(977 K)	-51	Ref. 156
$Sm_3Ga_5O_{12}$	-76(977 K)	-67	Ref. 156
$Eu_3Ga_5O_{12}$	-72(977 K)	-28	Ref. 156
$Gd_3Ga_5O_{12}$	-76(977 K)	-53	Ref. 156
$Dy_3Ga_5O_{12}$	-62(977 K)	-53	Ref. 156
$Ho_3Ga_5O_{12}$	-66(977 K)	-48	Ref. 156
$Er_3Ga_5O_{12}$	-62(977 K)	-44	Ref. 156
$Tm_3Ga_5O_{12}$	-56(977 K)	-38	Ref. 156
$Lu_3Ga_5O_{12}$	-45(977 K)	-25	Ref. 156
$Y_3Ga_5O_{12}$	-69(977 K)	-52	Ref. 156
$Ca_3Al_2Si_3O_{12}$	-169	-132	Ref. 157

E_f^{EXP} is the enthalpy of formation of garnets from binary oxides, i.e., the enthalpy change of the reaction $\frac{3}{2}Ln_2O_3 + \frac{5}{2}M_2O_3 \longrightarrow Ln_3M_5O_{12}$ (Ln = Rare Earth, M=Al, Ga), and E_f^{DFT} is the DFT computed formation energy based on the same reaction. The mean absolute error (MAE) between E_f^{EXP} and E_f^{DFT} is $\sim$ 14 meV atom^{-1}.

Appendix B

Supporting information: High-throughput screening of Eu^{2+}-doped red-emission garnet phosphors using density functional theory and machine learning

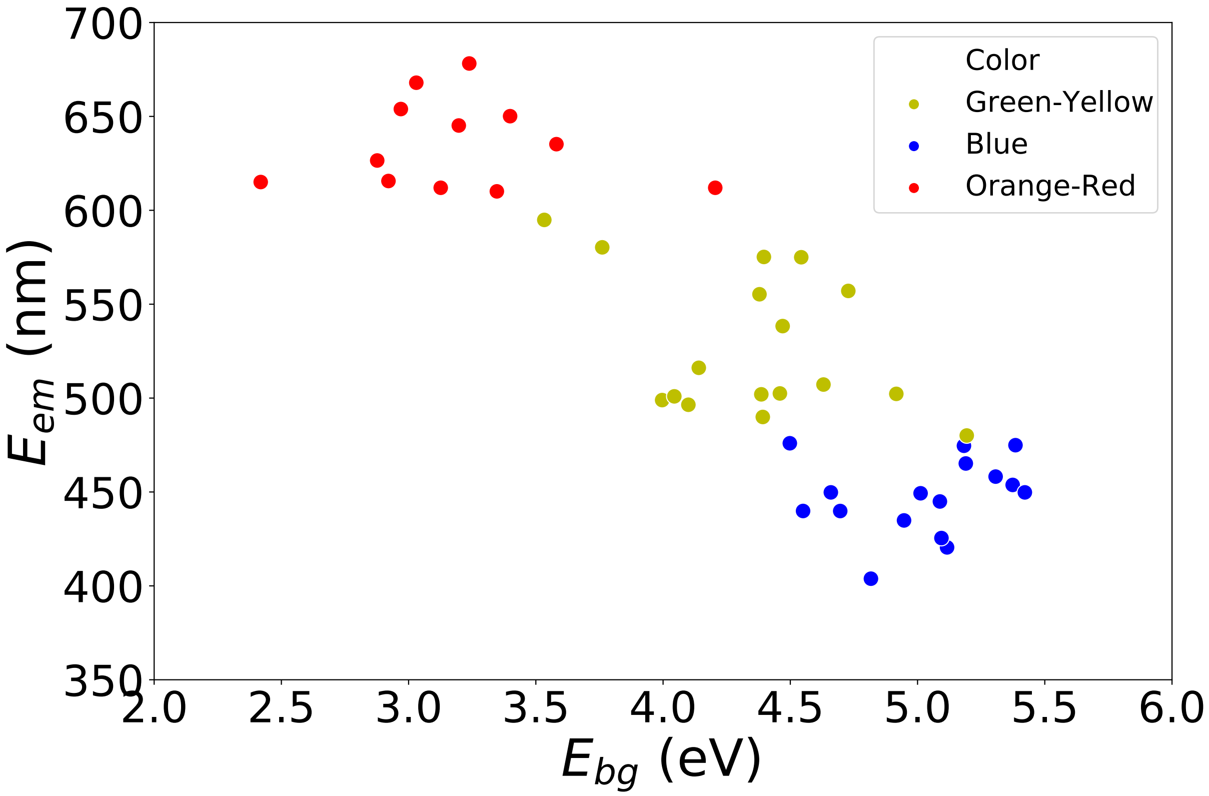


Figure B.1: The emission wavelength E_{em} vs. GGA bandgap E_{bg}^{DFT}. The emission energy (E_{em}) are obtained from[113,114,158–164]. The PBE bandgaps are obtained from Materials Project[11] and Table 1 in Ref. 9.

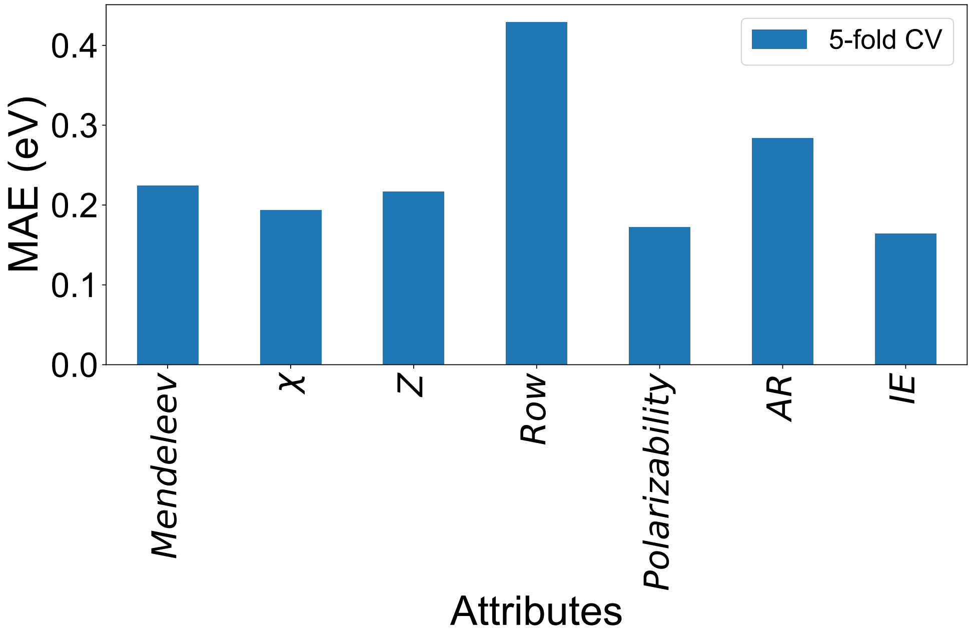


Figure B.2: The MAEs of XGBoost model using single attribute.

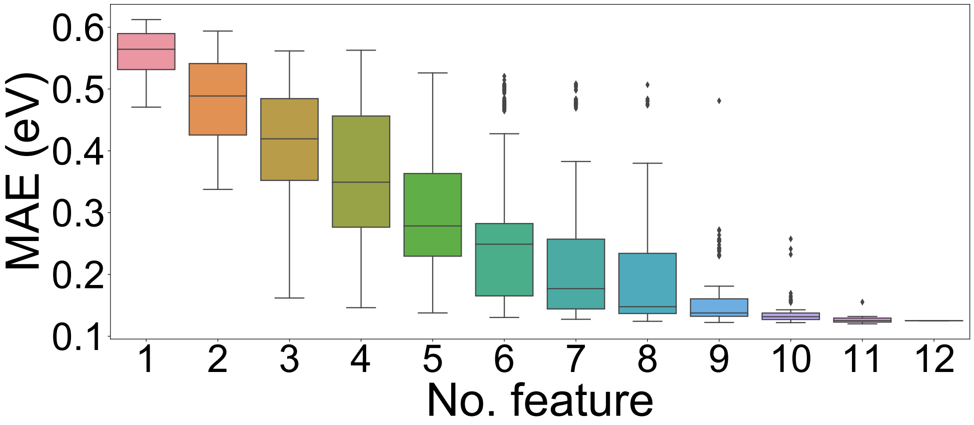


Figure B.3: The swamplot of exhaustive search result.

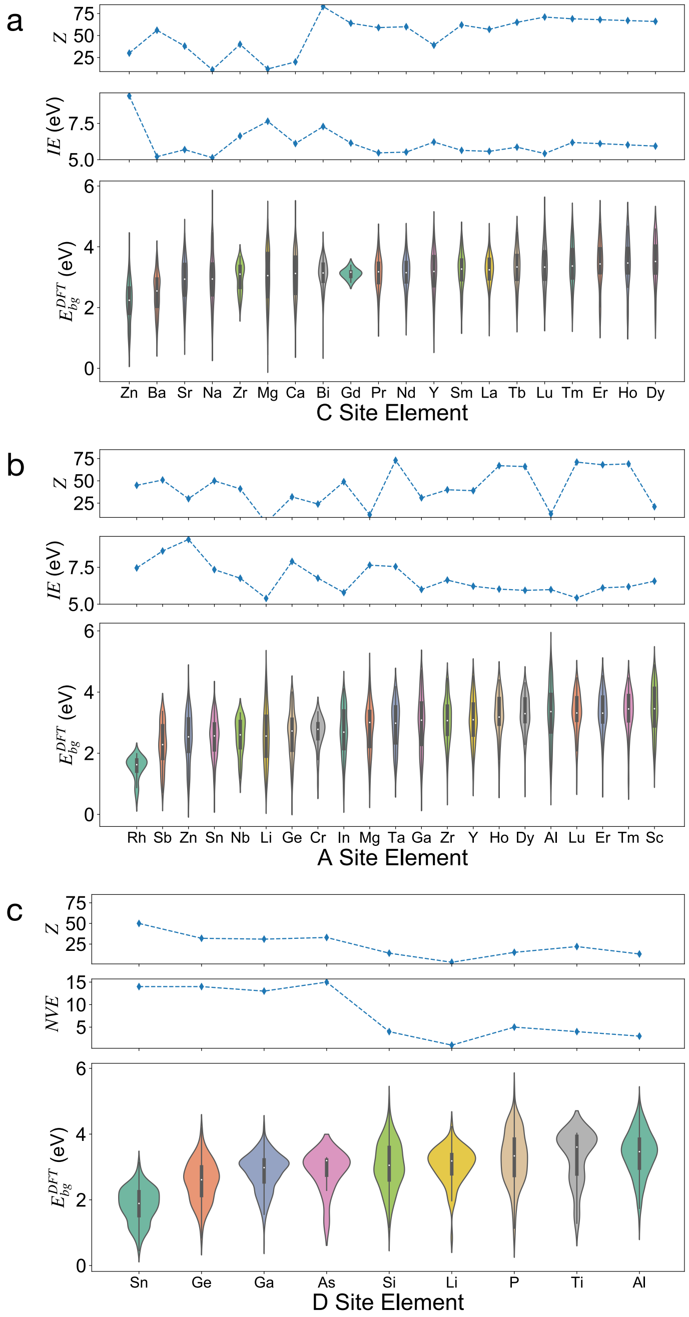


Figure B.4: The site features and the violin plot of the distribution of E_{bg}^{DFT} vs. element occupying (a) the C site elements, (b) the A site elements, and (c) the D site elements. The elements are sorted from left to right in the increasing order of the mean of the E_{bg}^{DFT}.

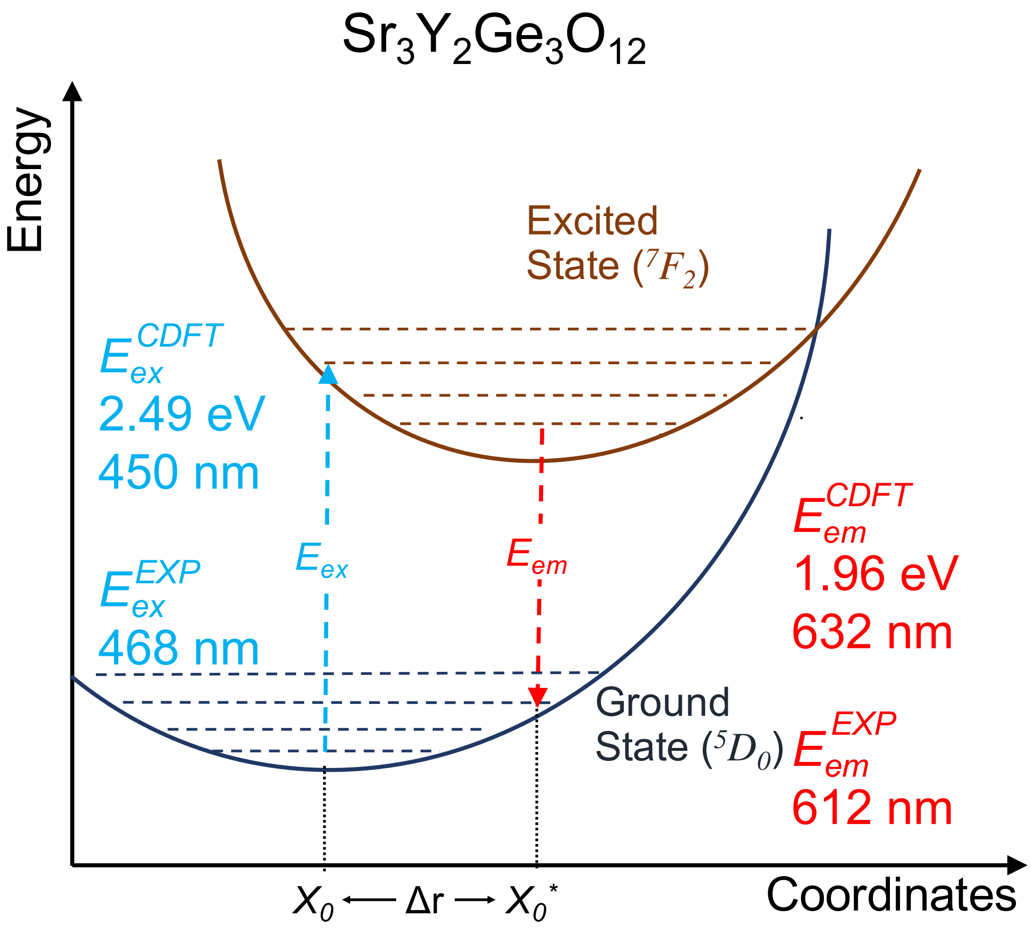


Figure B.5: The excitation E_{ex}^{CDFT} and emission E_{em}^{CDFT} wavelength calculated by CDFT are 450 nm and 632 nm, respectively. And the values reported from ref. 138 are 468 nm (E_{ex}^{EXP}) and 612 nm (E_{em}^{EXP}).

Appendix C

Supporting information: A universal machine learning model for elemental grain boundary energies

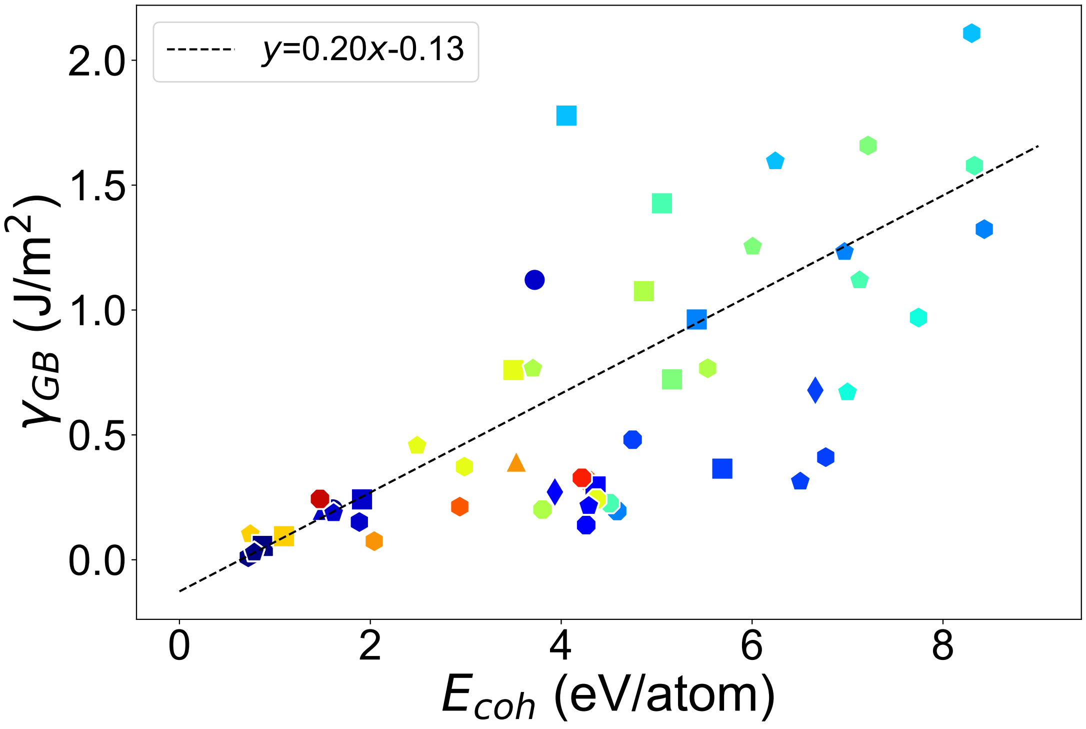


Figure C.1: The averaged elemental grain boundary energy plotted against the cohesive energy. The dotted line is a fitted linear function of *y*=0.20*x*-0.13, which helps to visualize the correlation between the E_{GB} and E_{coh}. The inset periodic table shows the marker and color scheme of the scatter plot.

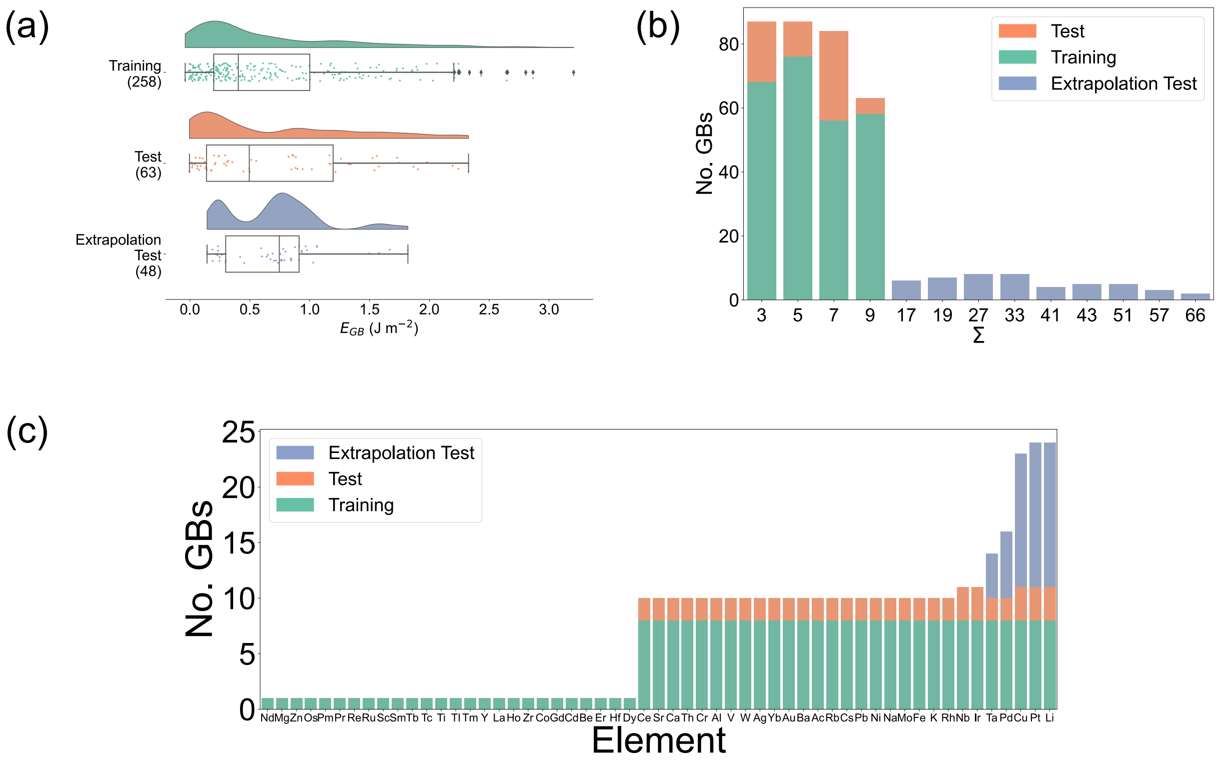
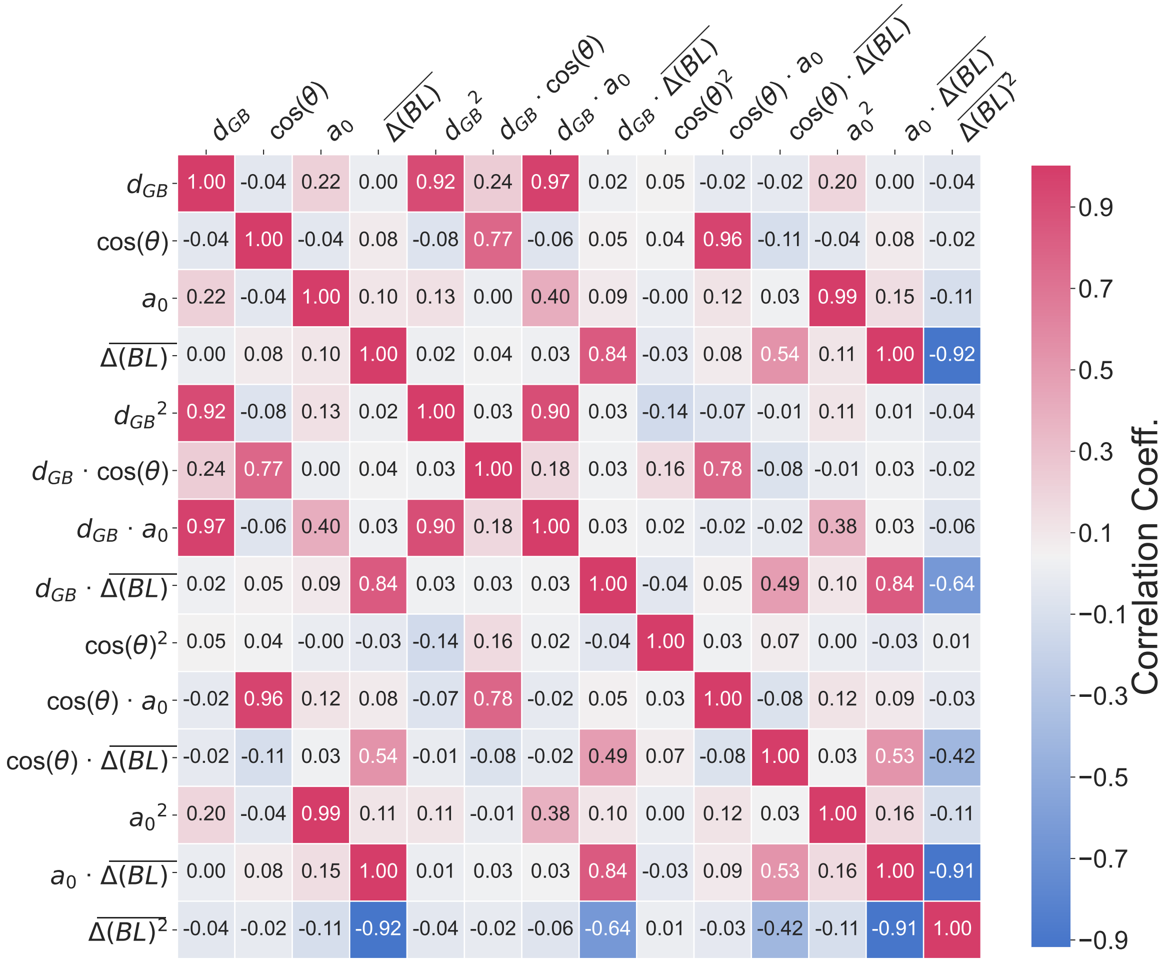


Figure C.2: (a) The distribution of the E_{GB}^{DFT} for different of data sets. The numbers in the bracket refer the number of data contained in the corresponding data set. (b) The Σ distribution. For the training and test set, we only used GBs with $\Sigma \leq 9$. In addition, we also prepared an external test data, of which the Σ ranges from 17 to 66, to test the extrapolability of the model on Σ. (c) The element distribution of the GBs.



	d_{GB}	$\cos(\theta)$	a_0	$\overline{\Delta(BL)}$	${d_{GB}}^2$	$d_{GB} \cdot \cos(\theta)$	$d_{GB} \cdot a_0$	$d_{GB} \cdot \overline{\Delta(BL)}$	$\cos(\theta)^2$	$\cos(\theta) \cdot a_0$	$\cos(\theta) \cdot \overline{\Delta(BL)}$	${a_0}^2$	$a_0 \cdot \overline{\Delta(BL)}$	$\overline{\Delta(BL)}^2$
d_{GB}	1.00	-0.04	0.22	0.00	0.92	0.24	0.97	0.02	0.05	-0.02	-0.02	0.20	0.00	-0.04
$\cos(\theta)$	-0.04	1.00	-0.04	0.08	-0.08	0.77	-0.06	0.05	0.04	0.96	-0.11	-0.04	0.08	-0.02
a_0	0.22	-0.04	1.00	0.10	0.13	0.00	0.40	0.09	-0.00	0.12	0.03	0.99	0.15	-0.11
$\overline{\Delta(BL)}$	0.00	0.08	0.10	1.00	0.02	0.04	0.03	0.84	-0.03	0.08	0.54	0.11	1.00	-0.92
${d_{GB}}^2$	0.92	-0.08	0.13	0.02	1.00	0.03	0.90	0.03	-0.14	-0.07	-0.01	0.11	0.01	-0.04
$d_{GB} \cdot \cos(\theta)$	0.24	0.77	0.00	0.04	0.03	1.00	0.18	0.03	0.16	0.78	-0.08	-0.01	0.03	-0.02
$d_{GB} \cdot a_0$	0.97	-0.06	0.40	0.03	0.90	0.18	1.00	0.03	0.02	-0.02	-0.02	0.38	0.03	-0.06
$d_{GB} \cdot \overline{\Delta(BL)}$	0.02	0.05	0.09	0.84	0.03	0.03	0.03	1.00	-0.04	0.05	0.49	0.10	0.84	-0.64
$\cos(\theta)^2$	0.05	0.04	-0.00	-0.03	-0.14	0.16	0.02	-0.04	1.00	0.03	0.07	0.00	-0.03	0.01
$\cos(\theta) \cdot a_0$	-0.02	0.96	0.12	0.08	-0.07	0.78	-0.02	0.05	0.03	1.00	-0.08	0.12	0.09	-0.03
$\cos(\theta) \cdot \overline{\Delta(BL)}$	-0.02	-0.11	0.03	0.54	-0.01	-0.08	-0.02	0.49	0.07	-0.08	1.00	0.03	0.53	-0.42
${a_0}^2$	0.20	-0.04	0.99	0.11	0.11	-0.01	0.38	0.10	0.00	0.12	0.03	1.00	0.16	-0.11
$a_0 \cdot \overline{\Delta(BL)}$	0.00	0.08	0.15	1.00	0.01	0.03	0.03	0.84	-0.03	0.09	0.53	0.16	1.00	-0.91
$\overline{\Delta(BL)}^2$	-0.04	-0.02	-0.11	-0.92	-0.04	-0.02	-0.06	-0.64	0.01	-0.03	-0.42	-0.11	-0.91	1.00

Figure C.3: The Pearson correlation matrix of the 2_{nd}-degree polynomial terms of the optimized feature subsets. There are 12 pairs of the features that have an absolute a correlation coefficient larger than 0.75, which are considered highly correlated.

Bibliography

[1] S Geller. Crystal chemistry of the garnets. *Zeitschrift für Kristallographie - Crystalline Materials*, 125(1-6):1–47, 1967. doi: 10.1524/zkri.1967.125.16.1.

[2] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, November 1964. doi: 10.1103/physrev.136.b864.

[3] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, November 1965. ISSN 0031-899X. doi: 10.1103/physrev.140.a1133.

[4] Geoffroy Hautier, Anubhav Jain, Hailong Chen, Charles Moore, Shyue Ping Ong, and Gerbrand Ceder. Novel mixed polyanions lithium-ion battery cathode materials predicted by high-throughput ab initio computations. *J. Mater. Chem.*, 21(43):17147–17153, October 2011. ISSN 1364-5501. doi: 10.1039/c1jm12216a.

[5] Geoffroy Hautier, Anubhav Jain, Shyue Ping Ong, Byoungwoo Kang, Charles Moore, Robert Doe, and Gerbrand Ceder. Phosphates as Lithium-Ion Battery Cathodes: An Evaluation Based on High-Throughput ab Initio Calculations. *Chem. Mater.*, 23(15): 3495–3508, August 2011. ISSN 0897-4756. doi: 10.1021/cm200949v.

[6] Shyue Ping Ong, Yifei Mo, William Davidson Richards, Lincoln Miara, Hyo Sug Lee, and Gerbrand Ceder. Phase stability, electrochemical stability and ionic conductivity of the Li10$\pm$1MP2X12 (M = Ge, Si, Sn, Al or P, and X = O, S or Se) family of superionic conductors. *Energy Environ. Sci.*, 6(1):148–156, December 2012. ISSN 1754-5706. doi: 10.1039/c2ee23355j.

[7] Jeff Greeley, Thomas F. Jaramillo, Jacob Bonde, Ib Chorkendorff, and Jens K. Nørskov. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature Mater*, 5(11):909–913, November 2006. ISSN 1476-4660. doi: 10.1038/nmat1752.

[8] Kesong Yang, Wahyu Setyawan, Shidong Wang, Marco Buongiorno Nardelli, and Stefano Curtarolo. A search model for topological insulators with high-throughput robustness descriptors. *Nature Mater*, 11(7):614–619, July 2012. ISSN 1476-4660. doi: 10.1038/nmat3332.

[9] Zhenbin Wang, Iek-Heng Chu, Fei Zhou, and Shyue Ping Ong. Electronic Structure Descriptor for the Discovery of Narrow-Band Red-Emitting Phosphors. *Chem. Mater.*, 28 (11):4024–4031, June 2016. ISSN 0897-4756. doi: 10.1021/acs.chemmater.6b01496.

[10] Zhenbin Wang, Jungmin Ha, Yoon Hwa Kim, Won Bin Im, Joanna McKittrick, and Shyue Ping Ong. Mining Unexplored Chemistries for Phosphors for High-Color-Quality White-Light-Emitting Diodes. *Joule*, 2(5):914–926, May 2018. ISSN 2542-4351. doi: 10.1016/j.joule.2018.01.015.

[11] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, July 2013. doi: 10.1063/1.4812323.

[12] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, November 2013. ISSN 1543-1851. doi: 10.1007/s11837-013-0755-4.

[13] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L. W. Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58:227–235, June 2012. ISSN 0927-0256. doi: 10.1016/j.commatsci.2012.02.002.

[14] Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. AiiDA: Automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111:218–230, 2016. doi: 10.1016/j.commatsci.2015.09.013.

[15] Jens S. Hummelshøj, Frank Abild-Pedersen, Felix Studt, Thomas Bligaard, and Jens K. Nørskov. CatApp: A web application for surface chemistry and heterogeneous catalysis. *Angewandte Chemie International Edition*, 51(1):272–274, 2012. doi: 10.1002/anie.20110 7947.

[16] Kamal Choudhary, Qin Zhang, Andrew CE Reid, Sugata Chowdhury, Nhan Van Nguyen, Zachary Trautt, Marcus W. Newrock, Faical Yannick Congo, and Francesca Tavazza. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Scientific data*, 5(1):1–12, 2018. doi: 10.1038/sdata.2018.82.

[17] NOMAD Repository. https://nomad-repository.eu/.

[18] Igor Ying Zhang, Xin Xu, Yousung Jung, and William A. Goddard. A fast doubly hybrid density functional method close to chemical accuracy using a local opposite spin ansatz. *Proceedings of the National Academy of Sciences*, 108(50):19896–19900, 2011. doi: 10/fccz52.

[19] Ying Zhang, Xin Xu, and William A. Goddard. Doubly hybrid density functional for accurate descriptions of nonbond interactions, thermochemistry, and thermochemical kinetics. *PNAS*, 106(13):4963–4968, March 2009. doi: 10.1073/pnas.0901093106.

[20] Hyunjun Ji, Yihan Shao, William A. Goddard, and Yousung Jung. Analytic Derivatives of Quartic-Scaling Doubly Hybrid XYGJ-OS Functional: Theory, Implementation, and Benchmark Comparison with M06-2X and MP2 Geometries for Nonbonded Complexes. *J. Chem. Theory Comput.*, 9(4):1971–1976, April 2013. ISSN 1549-9618. doi: 10/f4wxh3.

[21] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354–359, October 2017. ISSN 1476-4687. doi: 10.1038/nature24270.

[22] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860, February 2014. ISSN 0957-4174. doi: 10.1016/j.eswa.2013.08.015.

[23] Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep Reinforcement Learning framework for Autonomous Driving. *Electronic Imaging*, 2017 (19):70–76, January 2017. doi: 10.2352/issn.2470-1173.2017.19.avm-023.

[24] P Villars, N Onodera, and S Iwata. The Linus Pauling file (LPF) and its application to materials design. *Journal of Alloys and Compounds*, 279(1):1–7, September 1998. ISSN 0925-8388. doi: 10.1016/s0925-8388(98)00605-7.

[25] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown. The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.*, 23(2):66–69, May 1983. ISSN 0095-2338. doi: 10.1021/ci00038a003.

[26] Saulius Gražulis, Daniel Chateigner, Robert T. Downs, A. F. T. Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography Open Database–an open-access collection of crystal structures. *Journal of applied crystallography*, 42(4):726–729, 2009. doi: 10.1107/s0021889809016690.

[27] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.*, 31 (9):3564–3572, May 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b01294.

[28] Joohwi Lee, Atsuto Seko, Kazuki Shitara, Keita Nakayama, and Isao Tanaka. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B*, 93(11):115104, March 2016. doi: 10.1103/physrevb.93.115104.

[29] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.*, 120(14): 145301, April 2018. doi: 10.1103/physrevlett.120.145301.

[30] Muratahan Aykol, Soo Kim, and C. Wolverton. Van der Waals Interactions in Layered Lithium Cobalt Oxides. *J. Phys. Chem. C*, 119(33):19053–19058, August 2015. ISSN 1932-7447. doi: 10.1021/acs.jpcc.5b06240.

[31] Ariel Lozano, Bruno Escribano, Elena Akhmatskaya, and Javier Carrasco. Assessment of van der Waals inclusive density functional theory methods for layered electroactive materials. *Phys. Chem. Chem. Phys.*, 19(15):10133–10139, April 2017. ISSN 1463-9084. doi: 10.1039/c7cp00284j.

[32] Kamal Choudhary, Gowoon Cheon, Evan Reed, and Francesca Tavazza. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Phys. Rev. B*, 98(1):014107, July 2018. doi: 10.1103/physrevb.98.014107.

[33] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013. ISSN 0927-0256. doi: 10.1016/j.commatsci.2012.10.028.

[34] Anubhav Jain, Shyue Ping Ong, Wei Chen, Bharat Medasani, Xiaohui Qu, Michael Kocher, Miriam Brafman, Guido Petretto, Gian-Marco Rignanese, and Geoffroy Hautier. FireWorks: A dynamic workflow system designed for high-throughput applications. *Concurrency and Computation: Practice and Experience*, 27(17):5037–5059, 2015. doi: 10.1002/cpe.3505.

[35] Kiran Mathew, Joseph H. Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek heng Chu, Tess Smidt, Brandon Bocklund, Matthew Horton, John Dagdelen, Brandon Wood, Zi Kui Liu, Jeffrey Neaton, Shyue Ping Ong, Kristin Persson, and Anubhav Jain. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science*, 139:140–152, November 2017. ISSN 0927-0256. doi: 10.1016/j.commatsci.2017.07.030.

[36] Jesús Carrete, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. Nanograined Half-Heusler Semiconductors as Advanced Thermoelectrics: An Ab Initio High-Throughput Statistical Study. *Advanced Functional Materials*, 24(47):7427–7432, 2014. ISSN 1616-3028. doi: 10.1002/adfm.201401201.

[37] Jesús Carrete, Wu Li, Natalio Mingo, Shidong Wang, and Stefano Curtarolo. Finding Unprecedentedly Low-Thermal-Conductivity Half-Heusler Semiconductors via High-Throughput Materials Modeling. *Phys. Rev. X*, 4(1):011019, February 2014. doi: 10.1103/physrevx.4.011019.

[38] Henry Wu, Aren Lorenson, Ben Anderson, Liam Witteman, Haotian Wu, Bryce Meredig, and Dane Morgan. Robust FCC solute diffusion predictions from ab-initio machine learning methods. *Computational Materials Science*, 134:160–165, June 2017. ISSN 0927-0256. doi: 10.1016/j.commatsci.2017.03.052.

[39] Maarten De Jong, Wei Chen, Randy Notestine, Kristin Persson, Gerbrand Ceder, Anubhav Jain, Mark Asta, and Anthony Gamst. A statistical learning framework for materials science: Application to elastic moduli of k-nary inorganic polycrystalline compounds. *Scientific reports*, 6(1):1–11, 2016. doi: 10.1038/srep34256.

[40] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput Mater*, 2(1):1–7, August 2016. ISSN 2057-3960. doi: 10.1038/npjcompumats.2016.28.

[41] Danail Bonchev. *Chemical Graph Theory: Introduction and Fundamentals*, volume 1. CRC Press, 1991. ISBN 0-85626-454-7.

[42] Kristof T. Schütt, Huziel E. Sauceda, P.-J. Kindermans, Alexandre Tkatchenko, and K.-R. Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

[43] An automatic engine for predicting materials properties.: Hackingmaterials/automatminer. Hacking Materials Research Group, July 2019.

[44] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.*, 114(10):105503, March 2015. doi: 10.1103/physrevlett.114.105503.

[45] Chiho Kim, Ghanshyam Pilania, and Ramamurthy Ramprasad. From organized high-throughput data to phenomenological theory using machine learning: The example of dielectric breakdown. *Chemistry of Materials*, 28(5):1304–1311, 2016. doi: 10.1021/acs.chemmater.5b04109.

[46] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018. doi: 10.1103/physrevmaterials.2.083802.

[47] Robert B. Wexler, John Mark P. Martirez, and Andrew M. Rappe. Chemical pressure-driven enhancement of the hydrogen evolving activity of Ni2P from nonmetal surface doping interpreted via machine learning. *Journal of the American Chemical Society*, 140 (13):4678–4683, 2018. doi: 10.1021/jacs.8b00947.

[48] Jino Im, Seongwon Lee, Tae-Wook Ko, Hyun Woo Kim, YunKyong Hyon, and Hyunju Chang. Identifying Pb-free perovskites for solar cells by machine learning. *npj Computational Materials*, 5(1):1–8, 2019. doi: 10.1038/s41524-019-0177-0.

[49] Chi Chen, Zhi Deng, Richard Tran, Hanmei Tang, Iek-Heng Chu, and Shyue Ping Ong. Accurate force field for molybdenum by machine learning large materials data. *Physical Review Materials*, 1(4):043603, 2017. doi: 10.1103/physrevmaterials.1.043603.

[50] Mohammed Albanna, Kyle W. Binder, Sean V. Murphy, Jaehyun Kim, Shadi A. Qasem, Weixin Zhao, Josh Tan, Idris B. El-Amin, Dennis D. Dice, and Julie Marco. In situ bioprinting of autologous skin cells accelerates wound healing of extensive excisional full-thickness wounds. *Scientific reports*, 9(1):1–15, 2019. doi: 10.1038/s41598-018-38366-w.

[51] Shyue Ping Ong, Lei Wang, Byoungwoo Kang, and Gerbrand Ceder. Li-Fe-P-O2 Phase Diagram from First Principles Calculations. *Chem. Mater.*, 20(5):1798–1807, March 2008. ISSN 0897-4756. doi: 10.1021/cm702327g.

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[53] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[55] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. ElemNet : Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci Rep*, 8(1):17593, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-35934-y.

[56] Wei Li, Ryan Jacobs, and Dane Morgan. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Computational Materials Science*, 150: 454–463, July 2018. ISSN 0927-0256. doi: 10.1016/j.commatsci.2018.04.033.

[57] Jonathan Schmidt, Jingming Shi, Pedro Borlido, Liming Chen, Silvana Botti, and Miguel A. L. Marques. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chem. Mater.*, 29(12):5090–5103, June 2017. ISSN 0897-4756. doi: 10.1021/acs.chemmater.7b00156.

[58] Alexander S. Gzyl, Anton O. Oliynyk, and Arthur Mar. Half-Heusler Structures with Full-Heusler Counterparts: Machine-Learning Predictions and Experimental Validation. *Crystal Growth & Design*, 20(10):6469–6477, October 2020. ISSN 1528-7483. doi: 10.1021/acs.cgd.0c00646.

[59] Anton O. Oliynyk, Erin Antono, Taylor D. Sparks, Leila Ghadbeigi, Michael W. Gaultois, Bryce Meredig, and Arthur Mar. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.*, 28(20):7324–7331, October 2016. ISSN 0897-4756, 1520-5002. doi: 10.1021/acs.chemmater.6b02724.

[60] Kyoungdoc Kim, Logan Ward, Jiangang He, Amar Krishna, Ankit Agrawal, and C. Wolverton. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Phys. Rev. Materials*, 2(12):123801, December 2018. doi: 10.1103/physrevmaterials.2.123801.

[61] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, and Shyue Ping Ong. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat Comput Sci*, 1(1):46–53, January 2021. ISSN 2662-8457. doi: 10.1038/s43588-020-00002-x.

[62] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.*, 9(7):1668–1673, April 2018. doi: 10.1021/acs.jpclett.8b00124.

[63] Shuaihua Lu, Qionghua Zhou, Yixin Ouyang, Yilv Guo, Qiang Li, and Jinlan Wang. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nature Communications*, 9(1):3405, August 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05761-w.

[64] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman. Machine learning bandgaps of double perovskites. *Sci Rep*, 6(1):19375, January 2016. ISSN 2045-2322. doi: 10.1038/srep19375.

[65] Vladislav Gladkikh, Dong Yeon Kim, Amir Hajibabaei, Atanu Jana, Chang Woo Myung, and Kwang S. Kim. Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties. *J. Phys. Chem. C*, 124(16):8905–8918, April 2020. ISSN 1932-7447. doi: 10.1021/acs.jpcc.9b11768.

[66] Fabien Tran and Peter Blaha. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Physical review letters*, 102(22):226401, 2009. doi: 10.1103/physrevlett.102.226401.

[67] M. K. Y. Chan and Gerbrand Ceder. Efficient band gap prediction for solids. *Physical review letters*, 105(19):196403, 2010. doi: 10.1103/physrevlett.105.196403.

[68] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of chemical physics*, 118(18):8207–8215, 2003. doi: 10.1063/1.1564060.

[69] F. Fuchs, J. Furthmüller, F. Bechstedt, M. Shishkin, and G. Kresse. Quasiparticle band structure based on a generalized Kohn-Sham scheme. *Physical Review B*, 76(11):115109, 2007. doi: 10.1103/physrevb.76.115109.

[70] L. Tan, K. Sridharan, T. R. Allen, R. K. Nanstad, and D. A. McClintock. Microstructure tailoring for property improvements by grain boundary engineering. *Journal of Nuclear Materials*, 374(1):270–280, February 2008. ISSN 0022-3115. doi: 10.1016/j.jnucmat.20 07.08.015.

[71] M Shimada, H Kokawa, Z. J Wang, Y. S Sato, and I Karibe. Optimization of grain boundary character distribution for intergranular corrosion resistant 304 stainless steel by twin-induced grain boundary engineering. *Acta Materialia*, 50(9):2331–2341, May 2002. ISSN 1359-6454. doi: 10.1016/s1359-6454(02)00064-2.

[72] Shin Kiyohara, Hiromi Oda, Tomohiro Miyata, and Teruyasu Mizoguchi. Prediction of interface structures and energies via virtual screening. *Science Advances*, 2(11):e1600746, November 2016. ISSN 2375-2548. doi: 10.1126/sciadv.1600746.

[73] Conrad W. Rosenbrock, Eric R. Homer, Gábor Csányi, and Gus L. W. Hart. Discovering the building blocks of atomic systems using machine learning: Application to grain boundaries. *npj Computational Materials*, 3(1):1–7, August 2017. ISSN 2057-3960. doi: 10.1038/s41524-017-0027-x.

[74] Brandon D. Snow, Dustin D. Doty, and Oliver K. Johnson. A Simple Approach to Atomic Structure Characterization for Machine Learning of Grain Boundary Structure-Property Models. *Front. Mater.*, 6, 2019. ISSN 2296-8016. doi: 10.3389/fmats.2019.00120.

[75] Joshua A. Gomberg, Andrew J. Medford, and Surya R. Kalidindi. Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning. *Acta Materialia*, 133:100–108, July 2017. ISSN 1359-6454. doi: 10.1016/j.actamat.2017.05.00 9.

[76] Hui Zheng, Xiang-Guo Li, Richard Tran, Chi Chen, Matthew Horton, Donald Winston, Kristin Aslaug Persson, and Shyue Ping Ong. Grain boundary properties of elemental metals. *Acta Materialia*, 186:40–49, March 2020. ISSN 13596454. doi: 10.1016/j.actama t.2019.12.030.

[77] Lei Wang, Thomas Maxisch, and Gerbrand Ceder. Oxidation energies of transition metal oxides within the GGA+U framework. *Phys. Rev. B*, 73(19):195107, May 2006. doi: 10.1103/physrevb.73.195107.

[78] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad. Accelerating materials property predictions using machine learning. *Sci Rep*, 3(1):2810, September 2013. ISSN 2045-2322. doi: 10.1038/srep02810.

[79] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat Commun*, 8(1):15679, June 2017. ISSN 2041-1723. doi: 10.1038/ncomms15679.

[80] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B*, 89(9):094104, March 2014. doi: 10.1103/physrevb.89.094104.

[81] Felix A. Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Machine Learning Energies of 2 Million Elpasolite ABC2D6 Crystals. *Phys. Rev. Lett.*, 117(13):135502, September 2016. doi: 10.1103/physrevlett.117.135502.

[82] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I. Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B*, 96(2):024104, July 2017. doi: 10.1103/physrevb.96.024104.

[83] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B*, 95(14):144110, April 2017. doi: 10.1103/physrevb.95.144110.

[84] Wenhao Sun, Stephen T. Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D. Richards, Anthony C. Gamst, Kristin A. Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2(11): e1600225, November 2016. ISSN 2375-2548. doi: 10.1126/sciadv.1600225.

[85] Geoffroy Hautier, Shyue Ping Ong, Anubhav Jain, Charles J. Moore, and Gerbrand Ceder. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B*, 85(15):155208, April 2012. doi: 10.1103/physrevb.85.155208.

[86] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, May 2015. ISSN 1476-4687. doi: 10.1038/nature14539.

[87] Linus Pauling. The Principles Determining The Structure of Complex Ionic Crystals. *J. Am. Chem. Soc.*, 51(4):1010–1026, April 1929. ISSN 0002-7863. doi: 10.1021/ja01379a006.

[88] V. M. Goldschmidt. Die Gesetze der Krystallochemie. *Naturwissenschaften*, 14(21): 477–485, May 1926. ISSN 1432-1904. doi: 10.1007/bf01507527.

[89] Shuji Nakamura. Present performance of InGaN-based blue/green/yellow LEDs. In *Light-Emitting Diodes: Research, Manufacturing, and Applications*, volume 3002, pages 26–35. International Society for Optics and Photonics, April 1997. doi: 10.1117/12.271048.

[90] Michael P. O'Callaghan, Danny R. Lynham, Edmund J. Cussen, and George Z. Chen. Structure and Ionic-Transport Properties of Lithium-Containing Garnets Li3Ln3Te2O12 (Ln = Y, Pr, Nd, Sm-Lu). *Chem. Mater.*, 18(19):4681–4689, September 2006. ISSN 0897-4756. doi: 10.1021/cm060992t.

[91] Hongjian Peng, Qing Wu, and Lihong Xiao. Low temperature synthesis of Li5La3Nb2O12 with cubic garnet-type structure by sol–gel process. *J Sol-Gel Sci Technol*, 66(1):175–179, April 2013. ISSN 1573-4846. doi: 10.1007/s10971-013-2984-y.

[92] K.-I. Kobayashi, T. Kimura, H. Sawada, K. Terakura, and Y. Tokura. Room-temperature magnetoresistance in an oxide material with an ordered double-perovskite structure. *Nature*, 395(6703):677–680, October 1998. ISSN 1476-4687. doi: 10.1038/27167.

[93] R. J. Cava, B. Batlogg, R. B. van Dover, D. W. Murphy, S. Sunshine, T. Siegrist, J. P. Remeika, E. A. Rietman, S. Zahurak, and G. P. Espinosa. Bulk superconductivity at 91 K in single-phase oxygen-deficient perovskite Ba 2 YCu 3 O 9 - δ. *Phys. Rev. Lett.*, 58(16): 1676–1679, April 1987. ISSN 0031-9007. doi: 10.1103/physrevlett.58.1676.

[94] Ronald E. Cohen. Origin of ferroelectricity in perovskite oxides. *Nature*, 358(6382): 136–138, July 1992. ISSN 1476-4687. doi: 10.1038/358136a0.

[95] Ilya Grinberg, D. Vincent West, Maria Torres, Gaoyang Gou, David M. Stein, Liyan Wu, Guannan Chen, Eric M. Gallo, Andrew R. Akbashev, Peter K. Davies, Jonathan E. Spanier, and Andrew M. Rappe. Perovskite oxides for visible-light-absorbing ferroelectric and photovoltaic materials. *Nature*, 503(7477):509–512, November 2013. ISSN 1476-4687. doi: 10.1038/nature12622.

[96] Martin A. Green, Anita Ho-Baillie, and Henry J. Snaith. The emergence of perovskite solar cells. *Nature Photon*, 8(7):506–514, July 2014. ISSN 1749-4893. doi: 10.1038/npho ton.2014.134.

[97] Linus Pauling. The Nature of The Chemical Bond. IV. The Energy of Single Bonds and The Relative Electrogenativity of Atoms. *J. Am. Chem. Soc.*, 54(9):3570–3582, September 1932. ISSN 0002-7863. doi: 10.1021/ja01348a011.

[98] R. D. Shannon. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallographica Section A*, 32(5):751–767, 1976. ISSN 1600-5724. doi: 10.1107/s0567739476001551.

[99] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, January 1991. ISSN 0893-6080. doi: 10.1016/0893-6080(91)9 0009-t.

[100] François Chollet. Keras. 2015.

[101] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54(16):11169–11186, October 1996. doi: 10.1103/physrevb.54.11169.

[102] P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50(24):17953–17979, December 1994. doi: 10.1103/physrevb.50.17953.

[103] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, October 1996. doi: 10.1103/physrevlett.77.3865.

[104] Gus L. W. Hart, Lance J. Nelson, and Rodney W. Forcade. Generating derivative structures at a fixed concentration. *Computational Materials Science*, 59:101–107, June 2012. ISSN 0927-0256. doi: 10.1016/j.commatsci.2012.02.015.

[105] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

[106] E. Fred Schubert and Jong Kyu Kim. Solid-State Light Sources Getting Smart. *Science*, 308(5726):1274–1278, May 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1108712.

[107] Colin J. Humphreys. Solid-State Lighting. *MRS Bulletin*, 33(4):459–470, April 2008. ISSN 1938-1425, 0883-7694. doi: 10.1557/mrs2008.91.

[108] H. S. Jang, Y.-H. Won, and D. Y. Jeon. Improvement of electroluminescent property of blue LED coated with highly luminescent yellow-emitting phosphors. *Appl. Phys. B*, 95 (4):715–720, June 2009. ISSN 1432-0649. doi: 10.1007/s00340-009-3484-1.

[109] Ye Jin, Mu-Huai Fang, Marek Grinberg, Sebastian Mahlik, Tadeusz Lesniewski, M.G. Brik, Guan-Yu Luo, Jauyn Grace Lin, and Ru-Shi Liu. Narrow Red Emission Band Fluoride Phosphor KNaSiF6:Mn4+ for Warm White Light-Emitting Diodes. *ACS Appl. Mater. Interfaces*, 8(18):11194–11203, May 2016. ISSN 1944-8244. doi: 10.1021/acsami.6b01905.

[110] Toru Takahashi and Sadao Adachi. Mn4+ Activated Red Photoluminescence in K2SiF6 Phosphor. *J. Electrochem. Soc.*, 155(12):E183, 2008. ISSN 00134651. doi: 10.1149/1.2993159.

[111] Yue Tian, Baojiu Chen, Ruinian Hua, Jiashi Sun, Lihong Cheng, Haiyang Zhong, Xiangping Li, Jinsu Zhang, Yanfeng Zheng, Tingting Yu, Libo Huang, and Hongquan Yu. Optical transition, electron-phonon coupling and fluorescent quenching of La2(MoO4)3:Eu3+ phosphor. *Journal of Applied Physics*, 109(5):053511, March 2011. ISSN 0021-8979. doi: 10.1063/1.3551584.

[112] Ravi P. Rao. Growth and characterization of Y2O3:Eu3+ phosphor films by sol-gel process. *Solid State Communications*, 99(6):439–443, August 1996. ISSN 0038-1098. doi: 10.1016/0038-1098(96)00249-9.

[113] Kyota Uheda, Naoto Hirosaki, Yoshinobu Yamamoto, Atsushi Naito, Takuya Nakajima, and Hajime Yamamoto. Luminescence Properties of a Red Phosphor, CaAlSiN3 : Eu2 + , for White Light-Emitting Diodes. *Electrochem. Solid-State Lett.*, 9(4):H22, February 2006. ISSN 1944-8775. doi: 10.1149/1.2173192.

[114] Y. Q. Li, J. E. J. van Steen, J. W. H. van Krevel, G. Botty, A. C. A. Delsing, F. J. DiSalvo, G. de With, and H. T. Hintzen. Luminescence properties of red-emitting M2Si5N8:Eu2+ (M=Ca, Sr, Ba) LED conversion phosphors. *Journal of Alloys and Compounds*, 417(1): 273–279, June 2006. ISSN 0925-8388. doi: 10.1016/j.jallcom.2005.09.041.

[115] Rong-Jun Xie, Naoto Hirosaki, Takayuki Suehiro, Fang-Fang Xu, and Mamoru Mitomo. A Simple, Efficient Synthetic Route to Sr2Si5N8:Eu2+-Based Red Phosphors for White Light-Emitting Diodes. *Chem. Mater.*, 18(23):5578–5583, November 2006. ISSN 0897-4756. doi: 10.1021/cm061010n.

[116] Xiaoyong Huang. Red phosphor converts white LEDs. *Nature Photonics*, 8(10):748–749, October 2014. ISSN 1749-4893. doi: 10.1038/nphoton.2014.221.

[117] Daisuke Sekiguchi and Sadao Adachi. Synthesis and photoluminescence spectroscopy of BaGeF6:Mn4+ red phosphor. *Optical Materials*, 42:417–422, April 2015. ISSN 0925-3467. doi: 10.1016/j.optmat.2015.01.039.

[118] Hiromu Watanabe and Naoto Kijima. Crystal structure and luminescence properties of SrxCa1-xAlSiN3:Eu2+ mixed nitride phosphors. *Journal of Alloys and Compounds*, 475 (1):434–439, May 2009. ISSN 0925-8388. doi: 10.1016/j.jallcom.2008.07.054.

[119] Hui-Li Li, Rong-Jun Xie, Naoto Hirosaki, and Yoshiyuki Yajima. Synthesis and Photoluminescence Properties of Sr2Si5N8 : Eu2 + Red Phosphor by a Gas-Reduction and Nitridation Method. *J. Electrochem. Soc.*, 155(12):J378, October 2008. ISSN 1945-7111. doi: 10.1149/1.2999278.

[120] Mu-Huai Fang, Shu-Yi Meng, Natalia Majewska, Tadeusz Leśniewski, Sebastian Mahlik, Marek Grinberg, Hwo-Shuenn Sheu, and Ru-Shi Liu. Chemical Control of SrLi(Al1−xGax)3N4:Eu2+ Red Phosphors at Extreme Conditions for Application in Light-Emitting Diodes. *Chem. Mater.*, 31(12):4614–4618, June 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b01783.

[121] Chien-Chih Chiang, Ming-Shyong Tsai, and Min-Hsiung Hon. Luminescent Properties of Cerium-Activated Garnet Series Phosphor: Structure and Temperature Effects. *J. Electrochem. Soc.*, 155(6):B517, April 2008. ISSN 1945-7111. doi: 10.1149/1.2898093.

[122] Jia Liang, Balaji Devakumar, Liangling Sun, Shaoying Wang, Qi Sun, and Xiaoyong Huang. Full-visible-spectrum lighting enabled by an excellent cyan-emitting garnet phosphor. *Journal of Materials Chemistry C*, 8(14):4934–4943, 2020. doi: 10.1039/d0tc00006j.

[123] Can He, Haipeng Ji, Zhaohui Huang, Tiesheng Wang, Xiaoguang Zhang, Yangai Liu, Minghao Fang, Xiaowen Wu, Jiaqi Zhang, and Xin Min. Red-Shifted Emission in Y3MgSiAl3O12:Ce3+ Garnet Phosphor for Blue Light-Pumped White Light-Emitting Diodes. *J. Phys. Chem. C*, 122(27):15659–15665, July 2018. ISSN 1932-7447. doi: 10.1021/acs.jpcc.8b03940.

[124] Zhiguo Xia and Andries Meijerink. Ce3+-Doped garnet phosphors: Composition modification, luminescence properties and applications. *Chem. Soc. Rev.*, 46(1):275–299, January 2017. ISSN 1460-4744. doi: 10.1039/c6cs00551a.

[125] Shuxing Li, Yonghui Xia, Mahdi Amachraa, Nguyen Tuan Hung, Zhenbin Wang, Shyue Ping Ong, and Rong-Jun Xie. Data-Driven Discovery of Full-Visible-Spectrum Phosphor. *Chem. Mater.*, 31(16):6286–6294, August 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b02505.

[126] Hong-Jian Feng, Kan Wu, and Zun-Yi Deng. Predicting Inorganic Photovoltaic Materials with Efficiencies $>26\%$ via Structure-Relevant Machine Learning and Density Functional Calculations. *Cell Reports Physical Science*, page 100179, September 2020. ISSN 2666-3864. doi: 10.1016/j.xcrp.2020.100179.

[127] Mohammad Zafari, Deepak Kumar, Muhammad Umer, and Kwang S. Kim. Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts. *Journal of Materials Chemistry A*, 8(10):5209–5216, 2020. doi: 10.1039/c9ta12 608b.

[128] G. Pilania, J. E. Gubernatis, and T. Lookman. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, March 2017. ISSN 0927-0256. doi: 10.1016/j.commatsci.2016.12.004.

[129] L. Weston and C. Stampfl. Machine learning the band gap properties of kesterite I2-II-IV-V4 quaternary compounds for photovoltaics applications. *Phys. Rev. Materials*, 2(8): 085407, August 2018. doi: c.

[130] Yang Huang, Changyou Yu, Weiguang Chen, Yuhuai Liu, Chong Li, Chunyao Niu, Fei Wang, and Yu Jia. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *Journal of Materials Chemistry C*, 7(11):3238–3245, 2019. doi: 10.1039/c8tc05554h.

[131] Arunkumar Chitteth Rajan, Avanish Mishra, Swanti Satsangi, Rishabh Vaish, Hiroshi Mizuseki, Kwang-Ryeol Lee, and Abhishek K. Singh. Machine-Learning-Assisted Accurate Band Gap Predictions of Functionalized MXene. *Chem. Mater.*, 30(12):4031–4038, June 2018. ISSN 0897-4756. doi: 10.1021/acs.chemmater.8b00686.

[132] Weike Ye, Chi Chen, Zhenbin Wang, Iek-Heng Chu, and Shyue Ping Ong. Deep neural networks for accurate predictions of crystal stability. *Nat Commun*, 9(1):1–6, September 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06322-x.

[133] Mahdi Amachraa, Zhenbin Wang, Chi Chen, Shruti Hariyani, Hanmei Tang, Jakoah Brgoch, and Shyue Ping Ong. Predicting Thermal Quenching in Inorganic Phosphors. *Chem. Mater.*, 32(14):6256–6265, July 2020. ISSN 0897-4756. doi: 10.1021/acs.chemmater.0c02231.

[134] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B*, 57(3):1505–1509, January 1998. doi: 10.1103/physrevb.57.1505.

[135] A. Chaudhry, R. Boutchko, S. Chourou, G. Zhang, N. Grønbech-Jensen, and A. Canning. First-principles study of luminescence in Eu2+-doped inorganic scintillators. *Phys. Rev. B*, 89(15):155105, April 2014. doi: 10.1103/physrevb.89.155105.

[136] Zhuoying Zhu, Iek-Heng Chu, Zhi Deng, and Shyue Ping Ong. Role of Na+ Interstitials and Dopants in Enhancing the Na+ Conductivity of the Cubic Na3PS4 Superionic Conductor. *Chem. Mater.*, 27(24):8318–8325, December 2015. ISSN 0897-4756. doi: 10.1021/acs.chemmater.5b03656.

[137] F. A. Kröger and H. J. Vink. Relations between the Concentrations of Imperfections in Crystalline Solids. In Frederick Seitz and David Turnbull, editors, *Solid State Physics*, volume 3, pages 307–435. Academic Press, January 1956. doi: 10.1016/S0081-1947(08)60135-6.

[138] Sk Khaja Hussain and Jae Su Yu. Broad red-emission of Sr3Y2Ge3O12 :Eu2+ garnet phosphors under blue excitation for warm WLED applications. *RSC Advances*, 7(22):13281–13288, 2017. doi: 10.1039/c6ra28069b.

[139] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. doi: 10.1145/2939672.2939785.

[140] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[141] Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.

[142] Zhenbin Wang, Weike Ye, Iek-Heng Chu, and Shyue Ping Ong. Elucidating Structure–Composition–Property Relationships of the β-SiAlON:Eu2+ Phosphor. *Chem. Mater.*, 28(23):8622–8630, December 2016. ISSN 0897-4756. doi: 10.1021/acs.chemmater.6b03555.

[143] Edward S. Grew, Jeffrey H. Marsh, Martin G. Yates, Biljana Lazic, Thomas Armbruster, Andrew Locock, Samuel W. Bell, M. Darby Dyar, Heinz-Jürgen Bernhardt, and Olaf Medenbach. Menzerite-(Y), A New Species, (Y, RE)(Ca,Fe2+)2(Mg,Fe2+)(Fe3+,Al)(Si3)O12, from A Felsic Granulite, Parry Sound, Ontario, and A new Garnet Endmember, Y2CaMg2Si3O12. *The Canadian Mineralogist*, 48(5):1171–1193, October 2010. ISSN 0008-4476. doi: 10.3749/canmin.48.5.1171.

[144] Qinghong Meng, Xuejiao Wang, Qi Zhu, and Ji-Guang Li. The effects of Mg2+/Si4+ co-substitution for Al3+ on sintering and photoluminescence of (Gd,Lu)3Al5O12:Ce garnet ceramics. *Journal of the European Ceramic Society*, 40(8):3262–3269, July 2020. ISSN 0955-2219. doi: 10.1016/j.jeurceramsoc.2020.03.019.

[145] Sutatch Ratanaphan, David L. Olmsted, Vasily V. Bulatov, Elizabeth A. Holm, Anthony D. Rollett, and Gregory S. Rohrer. Grain boundary energies in body-centered cubic metals. *Acta Materialia*, 88:346–354, April 2015. ISSN 13596454. doi: 10.1016/j.actamat.2015.01.069.

[146] H. Grimmer, W. Bollmann, and D. H. Warrington. Coincidence-site lattices and complete pattern-shift in cubic crystals. *Acta Cryst A*, 30(2):197–207, March 1974. ISSN 0567-7394. doi: 10.1107/s056773947400043x.

[147] D. Wolf and J. F. Lutsko. On the geometrical relationship between tilt and twist grain boundaries. *Zeitschrift für Kristallographie - Crystalline Materials*, 189(1-4):239–262, November 1989. ISSN 2196-7105. doi: 10.1524/zkri.1989.189.14.239.

[148] W. T. Read and W. Shockley. Dislocation Models of Crystal Grain Boundaries. *Phys. Rev.*, 78(3):275–289, May 1950. doi: 10.1103/physrev.78.275.

[149] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, pages 485–492, New York, NY, USA, July 2016. Association for Computing Machinery. ISBN 978-1-4503-4206-3. doi: 10.1145/2908812.2908918.

[150] Gregory S. Rohrer. Grain boundary energy anisotropy: A review. *J Mater Sci*, 46(18):5881–5895, September 2011. ISSN 0022-2461, 1573-4803. doi: 10.1007/s10853-011-5677-3.

[151] Vasily V. Bulatov, Bryan W. Reed, and Mukul Kumar. Grain boundary energy function for fcc metals. *Acta Materialia*, 65:161–175, February 2014. ISSN 13596454. doi: 10.1016/j.actamat.2013.10.057.

[152] Jia Li, Shen J. Dillon, and Gregory S. Rohrer. Relative grain boundary area and energy distributions in nickel. *Acta Materialia*, 57(14):4304–4311, August 2009. ISSN 1359-6454. doi: 10.1016/j.actamat.2009.06.004.

[153] Michael A. Gibson and Christopher A. Schuh. A survey of ab-initio calculations shows that segregation-induced grain boundary embrittlement is predicted by bond-breaking arguments. *Scripta Materialia*, 113:55–58, March 2016. ISSN 13596462. doi: 10.1016/j.scriptamat.2015.09.041.

[154] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/a:1010933404324.

[155] Oswald Kubaschewski, Charles B Alcock, and PJ Spencer. Materials Thermochemistry. Revised. *Pergamon Press Ltd, Headington Hill Hall, Oxford OX 3 0 BW, UK, 1993. 363*, 1993.

[156] Yasushi Kanke and Alexandra Navrotsky. A Calorimetric Study of the Lanthanide Aluminum Oxides and the Lanthanide Gallium Oxides: Stability of the Perovskites and the Garnets. *Journal of Solid State Chemistry*, 141(2):424–436, December 1998. ISSN 0022-4596. doi: 10.1006/jssc.1998.7969.

[157] Ladislav Cemič. *Thermodynamics in Mineral Sciences*. Springer, 2005.

[158] C. J. Duan, X. J. Wang, W. M. Otten, A. C. A. Delsing, J. T. Zhao, and H. T. Hintzen. Preparation, Electronic Structure, and Photoluminescence Properties of Eu2+- and Ce3+/Li+-Activated Alkaline Earth Silicon Nitride MSiN2 (M = Sr, Ba). *Chem. Mater.*, 20(4): 1597–1605, February 2008. ISSN 0897-4756. doi: 10.1021/cm701875e.

[159] J. McKittrick, M. E. Hannah, A. Piquette, J. K. Han, J. I. Choi, M. Anc, M. Galvez, H. Lugauer, J. B. Talbot, and K. C. Mishra. Phosphor Selection Considerations for Near-UV LED Solid State Lighting. *ECS J. Solid State Sci. Technol.*, 2(2):R3119, December 2012. ISSN 2162-8777. doi: 10.1149/2.017302jss.

[160] Sebastian Schmiechen, Hajnalka Schneider, Peter Wagatha, Cora Hecht, Peter J. Schmidt, and Wolfgang Schnick. Toward New Phosphors for Application in Illumination-Grade White pc-LEDs: The Nitridomagnesosilicates Ca[Mg3SiN4]:Ce3+, Sr[Mg3SiN4]:Eu2+, and Eu[Mg3SiN4]. *Chem. Mater.*, 26(8):2712–2719, April 2014. ISSN 0897-4756. doi: 10.1021/cm500610v.

[161] Philipp Pust, Angela S. Wochnik, Elen Baumann, Peter J. Schmidt, Detlef Wiechert, Christina Scheu, and Wolfgang Schnick. Ca[LiAl3N4]:Eu2+—A Narrow-Band Red-Emitting Nitridolithoaluminate. *Chem. Mater.*, 26(11):3544–3549, June 2014. ISSN 0897-4756. doi: 10.1021/cm501162n.

[162] Philipp Pust, Volker Weiler, Cora Hecht, Andreas Tücks, Angela S. Wochnik, Ann-Kathrin Henß, Detlef Wiechert, Christina Scheu, Peter J. Schmidt, and Wolfgang Schnick. Narrow-band red-emitting Sr[LiAl 3 N 4]:Eu 2+ as a next-generation LED-phosphor material. *Nature Materials*, 13(9):891–896, September 2014. ISSN 1476-4660. doi: 10.1038/nmat4012.

[163] Ping Zhang, Lingxia Li, Mingxia Xu, and Lan Liu. The new red luminescent Sr3Al2O6:Eu2+ phosphor powders synthesized via sol–gel route by microwave-assisted. *Journal of Alloys and Compounds*, 456(1):216–219, May 2008. ISSN 0925-8388. doi: 10.1016/j.jallcom.2007.02.004.

[164] Shan-shan Yao, Li-hong Xue, and You-wei Yan. Concentration quenching of Eu2+ in Ba2Mg(BO3)2: Eu2+ phosphor. *Current Applied Physics*, 11(3):639–642, May 2011. ISSN 1567-1739. doi: 10.1016/j.cap.2010.10.018.