2019-12-01

Materials Prediction Using High-Throughput and Machine Learning Techniques

Chandramouli Nyshadham
Brigham Young University

Materials Prediction Using High-Throughput and Machine Learning Techniques

Chandramouli Nyshadham

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Gus L. Hart, Chair
Scott D. Bergeson
Eric R. Homer
Tony R. Martinez
David W. Neilsen

Department of Physics and Astronomy

Brigham Young University

ABSTRACT

Materials Prediction Using High-Throughput and Machine Learning Techniques

Chandramouli Nyshadham
Department of Physics and Astronomy, BYU
Doctor of Philosophy

Predicting new materials through virtually screening a large number of hypothetical materials using supercomputers has enabled materials discovery at an accelerated pace. However, the innumerable number of possible hypothetical materials necessitates the development of faster computational methods for speedier screening of materials reducing the time of discovery. In this thesis, I aim to understand and apply two computational methods for materials prediction. The first method deals with a computational high-throughput study of s uperalloys. Superalloys are materials which exhibit high-temperature strength. A combinatorial high-throughput search across 2224 ternary alloy systems revealed 102 potential superalloys of which 37 are brand new, all of which we patented. The second computational method deals with a machine-learning (ML) approach and aims at understanding the consistency among five different state-of-the-art machine-learning models in predicting the formation enthalpy of 10 different binary alloys. The study revealed that although the five different ML models approach the problem uniquely, their predictions are consistent with each other and that they are all capable of predicting multiple materials simultaneously.

My contribution to both the projects included conceiving the idea, performing calculations, interpreting the results, and writing significant portions of the two journal articles published related to each project. A follow-up work of both computational approaches, their impact, and future outlook of materials prediction are also presented.

Keywords: materials prediction, superalloys, high-throughput, machine learning, computational materials science, density functional theory, formation enthalpy

ACKNOWLEDGMENTS

The end product of my Ph.D. is not just this thesis but myself. During the last five years, I not only learned physics of materials but also experienced great life in the company of many wonderful people here at BYU and Provo, with whom I was fortunate enough to interact and work together. I'm thankful to every one of them and would like to acknowledge a few here. Coming from an engineering background with a passion for physics, it is hard to find an adviser in the physics department willing to believe in me and take me as his doctoral student. I was fortunate enough to find Dr. Hart, who believed in me (more than myself) and supported me throughout my Ph.D. He is a great adviser, and I learned a lot from him in both research and life. I was fortunate to attend and experience multiple physics conferences every year all over the country. I'm also thankful to him for giving me the opportunity and experience I had in collaborating with great people from Cambridge and Berlin on projects. I'm also grateful to his wife Cynthia and their children for their kind hospitality during Christmas times. I loved their company.

I thank all the graduate students (Conrad, Wiley, Jeremey, and Tyler), postdocs (Andrew, and Carlos), and undergrads (Jake, Derek, Spencer, Brayden, Kennedy, Hayden, and Nate) in the group with whom I worked and had great insightful discussions. I'm grateful to all the graduate committee members and graduate coordinator, Dr. Eric Hirschmann for their immense support. I'm thankful to my parents and near and dear family members for their support and encouragement. They made my life easier, happier, and joyful. I'm also thankful to the physics and astronomy department staff (Shelena and Nan), and professors for their constant and great support. I'm also thankful to my violin teacher (Molly Cowley), and my many friends who made my life in Provo a joyful experience. I'm thankful and acknowledge the funding from the office of naval research, ONR (MURI N00014-13-1-0635) and the physics and astronomy department funding. The funding helped me spend more time on learning and research. Finally, last but not least, I'm thankful for www.audible.com. I listened to 53 books on audible as of now. This new habit made my life more meaningful and helped me understand, experience, and live life better.

I dedicate this thesis work to my parents, Rajya Lakshmi Nyshadham, Siva Sankar Nyshadham, and my adviser Dr. Gus L. W. Hart. This thesis work is a result of their support and belief in me and my passion.

Contents

Contents v

List of Figures vii

List of Tables xii

1 A computational approach for materials prediction 1
- 1.1 Introduction . . . 1
- 1.2 High-throughput approach for materials prediction . . . 1
- 1.3 Machine learning for materials prediction . . . 2
- 1.4 Impact and follow-up work . . . 3
  - 1.4.1 High-throughput approach . . . 3
  - 1.4.2 Machine learning approach . . . 4
- 1.5 Summary . . . 4

2 A high-throughput search for ternary superalloys 5
- 2.1 Superalloys . . . 5
- 2.2 Structure-property relationship . . . 5
- 2.3 Project background . . . 6
- 2.4 Project . . . 6
  - 2.4.1 My contribution . . . 7
- 2.5 Followup work: effect of quaternary additions to Al-Co-W . . . 19

3 Machine-learned models for materials prediction 21
- 3.1 Machine-learning: new paradigm for materials prediction . . . 21
- 3.2 Accelerated high-throughput using ML . . . 21
- 3.3 Project . . . 21
  - 3.3.1 My contribution . . . 22
- 3.4 Followup work: 45 alloys MBTR models, FIBS, and CUR . . . 29
  - 3.4.1 Many-body tensor representation with kernel ridge regression (MBTR+KRR) on 45 binary alloys . . . 29
  - 3.4.2 On improving the accuracy of a model . . . 29
  - 3.4.3 Active learning based ML models . . . 30

4 Conclusion and future work 36

5 Appendix 38
- 5.1 Paper copyright licenses . . . 38
- 5.2 Supplementary information for the journal articles . . . 40
  - 5.2.1 A computational high-throughput search for new ternary superalloys . . . 40
  - 5.2.2 Machine-learned multi-system surrogate models for materials prediction . . . 40
- 5.3 Methodology for generating `DFT-45B` dataset . . . 48

Bibliography **49**

List of Figures

1.1 (on the left) Field emission scanning electron micrographs of Co-Al-W-Mo superalloy (taken from [21]). We can notice the precipitates of a crystal structure phase called $L1_2$ denoted as γ' within a solid solution of an fcc crystal structure host matrix (cobalt) marked as γ. a) and b) show the crytal structures of fcc and $L1_2$ (a face-centered cubic (fcc) structure with different atoms on faces and edges) respectively. 2

1.2 ML models built from first-principles data is used to compute properties of a large number of materials at a faster pace. The materials can then filtered in a high-throughput fashion for desired properties, thus saving time and increasing the speed of discovery. 3

2.1 Depiction of major slip system in a) bcc, b) fcc, and c) hcp crystal structures. Only one major slip plane is shown for each of the crystal structure. {110} (red) for bcc, {111} (green) for fcc and (0001) (blue) for hcp. Dislocation motion can easily occur along the slip planes. 6

2.2 (All the figures are inspired by similar figures in reference [30]) a) *Single dislocation in a crystal:* The irregularity in the crystal marked in the figure shows a dislocation. The dislocation in the figure is an edge dislocation where an extra half-plane of atoms is introduced midway through the crystal. b) *Multiple dislocations in a crystal:* Movement of one dislocation is inhibited by the other. More dislocations in a materials increases its strength. c) *Solid-solution strengthening:* We can see a different species of atom (red) replacing the host atom (blue) within the crystal structure. The solute (red) atom helps inhibits the dislocation motion strengthening the material. d)*Precipitate strengthening:* The red and blue atom is the precipitate within the host matrix of blue atoms. The distortion in the lattice created by the precipitate helps strengthen the material. Superalloys materials contain $L1_2$ precipitates within a fcc structure. 7

2.3 For each base element in $X_3[A_{0.5}, B_{0.5}]$, there are 40 elements chosen for 'A' or 'B' which includes 38 elements (highlighted in blue) chosen from the periodic table and the remaining two of three base elements 'X' (highlighted in red). . . . 10

2.4 The 32-atom special quasirandom structure (SQS-32) used to model a solid solution with an $L1_2$ structure (smaller cube in the figure). The blue, red, and green atoms correspond to X, A, and B in $X_3[A_{0.5}, B_{0.5}]$, respectively. 10

2.5 The two-phase equilibrium screening criterion discussed in Sec. 2.3. If a tie line between the host matrix and the $L1_2$ precipitate phase (light blue dotted line) is intersected by the tie line for another phase (eg., green line between $X_{3.6}A_1$ and X_4B_1) then the precipitate phase will not be in two-phase equilibrium with the host matrix for any concentration between X_4A_1 and X_4B_1. 11

2.6 Formation enthalpy vs. the decomposition energy for all 2224 ternary systems. Each triangle represents one Ni_3/Co_3/$Fe_3[A_{0.5}, B_{0.5}]$ structure, where A and B are any two different elements in the periodic table from Fig. 1. Co-based and Fe-based systems are displaced on the x-axis by 200 meV and 400 meV, respectively, for clarity. Ni-based, Co-based, and Fe-based systems are marked in blue, red, and green triangles, respectively. Systems enclosed within dotted lines are the ones identified to be better than the $Co_3[Al_{0.5}, W_{0.5}]$ with respect to these properties. 11

2.7 All the elements are arranged as per the chemical scale (χ) introduced by Pettifor in increasing order. Each circle, square, and diamond represents a ternary combination $X_3[A_{0.5}, B_{0.5}]$ with X = Ni, Co, or Fe, and A, B are the elements indicated along the x and y-axes, respectively. A square indicates that the SQS-32 crystal structure has a positive formation enthalpy. A diamond indicates that there exists no stable binary or ternary compounds in the respective ternary system. A colored circle indicates that the SQS-32 structure has a negative formation enthalpy. The color contrast from yellow to black indicates decreasing formation enthalpy of the crystal structure in ternary system. 12

2.8 All the elements are arranged as per the chemical scale (χ) introduced by Pettifor in increasing order. Each circle, square, and diamond represents a ternary combination $X_3[A_{0.5}, B_{0.5}]$ with X = Ni, Co, or Fe, and A, B are the elements indicated along the x and y-axes, respectively. A square indicates that the SQS-32 crystal structure has a positive formation enthalpy. A diamond indicates that there exists no stable binary or ternary compounds in the respective ternary system. The color contrast from yellow to black indicates increasing decomposition energy of the crystal structure in ternary system. 13

2.9 A comparison between the density range for the theoretical calculations performed inthis work and modern superalloys. Densities are computed for 102 ternary systems screened from the 2224 systems computed in this work., The red line shows the range of density for commercially-available superalloys at present. . . 14

2.10 The magnitude of the bulk modulus for Ni-A-x (A = Al, Hf, Nb, Sb, Sc, Si, Ta, Ti, V, and Zr) systems with the x-axis arranged according to the χ scale in Pettifor maps. In general, the systems display a maximum in the bulk modulus at or before Ni. Only systems with simultaneously lower E_d and ΔH_f than $Co_3[Al_{0.5}, W_{0.5}]$ are plotted. 14

2.11 The magnitude of the bulk modulus for Co-A-x (A = Hf, Mo, Nb, Si, Ta, Ti, V, and W) systems with the x-axis arranged according to the χ scale in Pettifor maps. In general, the magnitude of the bulk modulus increases with χ up to Re. Only systems with simultaneously lower E_d and ΔH_f than $Co_3[Al_{0.5}, W_{0.5}]$ are plotted. 14

2.12 Distance to the $T = 0$ K convex hull algorithm. **a)** The correct distance (shown in green) for d_1 is the minimum distance of structure S_1 to all hyperplanes defining the convex hull. In case of structure S_2, the minimum distance is not d_2 (green line), an artifact of the hyperplane description for hull facets. **b)** Projecting the points to the zero energy line guarantees that all points will lie within the hull, thus enabling the use of minimization algorithm to calculate the correct distance. The distance to the hull d is given as the difference of the projected distance d_2 from the distance to the zero energy line d_1. 16
2.13 Change in the formation enthalpy by adding a quaternary element to Al-Co-W alloy. The parent structures are 12 and 16-atom unit cell and are close to ($\sim$ 2 meV above) convex hull. We can see C, Si, Ta, Ti, and V lowering the formation enthalpy. At $\sim$6%, Si and C lower the enthalpy of Al-Co-W. At $\sim$8%, and $\sim$12% we see Ta and V lowering the enthalpy too. Also, Si, C, and Ti lower the formation enthalpy further with increasing concentration. 20

3.1 The accelerated high-throughput approach. Candidate structures and properties are generated by surrogate machine-learning models based on reference electronic structure calculations in a materials repository. Selected structures are validated by electronic structure calculations, preventing false positive errors. 24
3.2 Consistency in prediction errors of formation enthalpy of five machine learning surrogate models on the `DFT-10B` dataset. (a) Root mean squared error (RMSE) of predicted enthalpies of formation of each surrogate model on each binary alloy subset in meV/atom (colored bars). RMSE for MTP results is computed using pure atom total energies obtained from DFT. The consistency of errors across models indicates the validity of machine learning surrogate models to predict formation enthalpy of materials—prediction errors are similar, independent of the details of model parametrization. (b) Root mean squared error (RMSE) of predicted enthalpies of formation of each surrogate model on each binary alloy subset as a percentage of energy range. Note that relative errors are below 2.5% for all systems. 25
3.3 *Performance of MBTR+KRR model for multiple alloy systems.* Shown are deviation of mean absolute error (MAE, vertical axis) of an MBTR+KRR surrogate model trained on k (horizontal axis) alloy systems simultaneously from the average MAE of k models trained on each alloy subsystem separately.Whiskers, boxes, horizontal line and numbers inside the plot show the range of values, quartiles, median and sample size, respectively. Difference in error between individual and combined models is always less than 1 meV/atom. 25
3.4 *Influence of biased training and validation sets.* Shown are the root mean squared errors (meV/atom) as a function of training and validation set composition obtained using MBTR+KRR model. See main text for discussion. 26

3.5 Root mean squared error (RMSE) of predicted enthalpies of formation of MBTR + KRR surrogate model on each of 45 binary alloys in meV/atom. The relative errors (δ_{RMSE}), express formation enthalpy as a percentage of the range of energies of individual alloy dataset. 32
3.6 Mapping a set of data points (blue and red) from input space to kernel space using the kernel function ϕ. Mapping onto the higher dimensional space allows us to classify the blue and red data points, classified using non-linear function (brown line) in input space with a linear hyper plane (brown colored plane) in the kernel space. 32
3.7 Graphical representation of a) singular value decomposition (SVD) and b) CUR decomposition of a matrix. In SVD, the matrix A is decomposed into three matrices namely U, Σ, and V^T. Here U and V^T are dense and big matrices with Σ being small and sparse. In CUR decomposition, the matrix A is decomposed into three matrices namely C, U, and R, where the matrices C and R are sparse and are expressed in terms of a small number of actual columns and actual rows of the data matrix and are thus more interpretable than SVD. 33
3.8 Root mean squared error (RMSE) of predicted enthalpies of formation of MBTR + KRR surrogate model on NbNi system in meV/atom using LTCUR, LSCUR and FIBS algorithms. 34
3.9 CPU time taken for LTCUR, LSCUR and FIBS algorithms for ranking and picking N training structures from a given dataset. 35

5.1 *Alchemical similarity explains prediction errors.*Shown are the logarithmized root mean squared error (RMSE; compare Fig. 2 as a function of an analytic expression in the difference in row r and column c of the periodic table as well as atomic number z of the two chemical element species of a binary alloy. $R^2 = 0.81$. 43
5.2 *Improvement of MBTR+DNN model on all alloys.*Shown are the root mean squared error (RMSE) when trained on each alloy separately (blue bars) and on all alloys simultaneously (grey bars). 44
5.3 *Influence of unit cell size on errors*. Shown are the absolute errors (meV/atom) as a function of the number of atoms in the unit cell for a validation set of 595 randomly chosen structures using the MBTR+KRR model. The number in brackets and the dashed line indicate the root mean squared error (RMSE, meV/atom) and the median absolute error (meV/atom) on the same set. If small structures (one or two atoms in the unit cell) are not contained in the training data (that is, are shown in the plot) they tend to have larger errors, increasing overall RMSE as well. If all small structures are contained in the training data, the overall RMSE is low (AlMg, CoNi). Retraining models with small structures included in the training set improved RMSE in all cases, by an amount depending on how many structures were added. 45

5.4 Visualizing all 15950 structures (DFT-10B) using a t-SNE plot. Each structure in the higher-dimensional space (MBTR) is graphically represented on a 2D plane using t-SNE method. We can observe that 1 or 2 atom unit cells are not representative of larger unit cells in the dataset and are away from other higher atom unit cells. This is a possible reason for high prediction errors when 1 or 2 atom cells are not included in the training set. 46

List of Tables

2.1 Systems where the SQS structure computed in this work has a corresponding $L1_2$ phase reported in experiment. The experimental compounds are all close to the stoichiometry of the SQS structure, $X_{24}[A_4, B_4]$ 14
2.2 Candidates for precipitate-forming systems that have no previously reported phase diagrams in standard databases. These have a smaller decomposition energy and a lower formation enthalpy than the Co_3(Al, W) superalloy. All are in stable two-phase equilibrium with the host matrix and have a relative lattice mismatch with the host matrix of less than or equal to 5%. Promising candidates (see Section 3.6) are boxed. $***$ indicates that the quantity is not computed in this work . . . 15

3.1 State-of-the-art surrogate machine-learning models investigated in this work . . . 24
3.2 Performance of general models 25
3.3 Size distribution in the DFT-10B dataset 26

5.1 Shown are the minimum and maximum values of k-point density across all structures for each of the alloys for computing the DFT total energies. 43

CHAPTER 1

A computational approach for materials prediction

1.1 INTRODUCTION

The philosophical inquiry pursued during the five years of this Ph.D. work has been about understanding and using computational methods for materials prediction. From a computational materials scientist perspective, the quest for new and better materials with enhanced properties is a perennial goal for humanity, which requires searching through the endless number of possible materials [1] and is one of the main challenges at present. The combinatorial nature of materials space makes it infeasible to explore properties of more than a small fraction of materials within a reasonable time [2].

The goal is to reduce the time taken to invent new materials. We need better methodologies to invent the next generation material for future technology at a faster pace. The main obstacle to inventing a new material is the amount of *work* needed to be done to compute every property of all possible materials. The aim of materials science during the last three decades has been to speed up the *work* by inventing new, faster methods in both the experimental and theoretical domains. In this silicon age, the surge of computational power used across a wide variety of domains is used to solve materials prediction problem. This dissertation work is a computational effort to speed up the *work* of materials invention by using two computational methods for materials prediction. The first computational method is called the *high-throughput* approach [3, 4], and the second is a machine-learning approach [5–7].

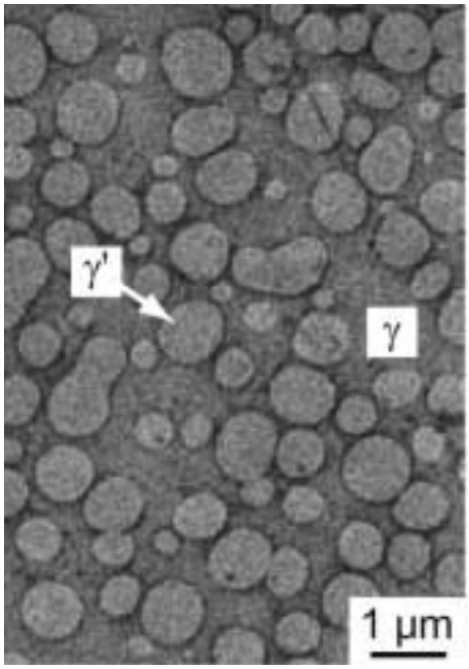
1.2 HIGH-THROUGHPUT APPROACH FOR MATERIALS PREDICTION

Computers and modern electronic structure codes such as density functional theory[1] (DFT) [8, 9] help us to virtually screen a large number of materials at an accelerated pace (high-throughput) compared to performing lab experiments on each material. The high-throughput approach [3, 4] is a computational method that takes advantage of robust DFT codes and the increasing speed of computers. In a high-throughput approach, the combinatorial space encompassing a large number of materials is screened intelligently to discover new materials with desired properties [1, 10].

The first project [11] in this dissertation work involves using a computational high-throughput approach for predicting new materials called superalloys [12]. Superalloys are materials with tremendous high-temperature strength, and their strength is a result of precipitates[2] of a secondary phase, called $L1_2$ (γ'), within a face-centered cubic (fcc) host matrix (γ) (see fig. 1.1). An investigation of 2224 ternary metallic systems using a high-throughput screening method for such precipitates ($L1_2$) within the fcc host matrices of cobalt, nickel, and iron revealed 37 new possible superalloy materials [11, 13]. Of these

[1] DFT is a computational modeling method based on quantum mechanics and used to compute the electronic structure of many-body systems.

[2] In the context here, a precipitate is a solid within a solid solution (alloy). Superalloys are materials that contain precipitates of a crystal structure phase called $L1_2$ (a face-centered cubic (fcc) structure with different atoms on faces and edges) within a solid solution of an fcc crystal structure.



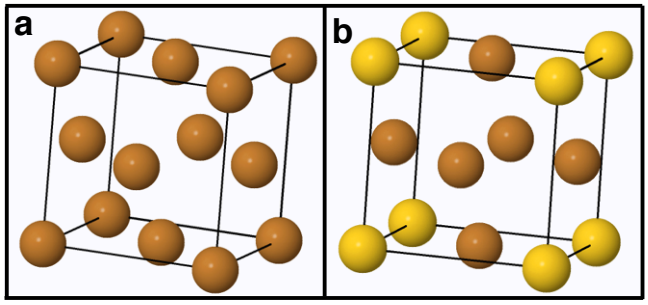


Figure 1.1 (on the left) Field emission scanning electron micrographs of Co-Al-W-Mo superalloy (taken from [21]). We can notice the precipitates of a crystal structure phase called $L1_2$ denoted as γ' within a solid solution of an fcc crystal structure host matrix (cobalt) marked as γ. a) and b) show the crytal structures of fcc and $L1_2$ (a face-centered cubic (fcc) structure with different atoms on faces and edges) respectively.

37, based on cost, availability, and toxicity, the six most promising candidates materials are reported, two of which have already been experimentally verified [14].

The limitation of the high-throughput method is that it is only as fast as the speed of available DFT codes and current supercomputers. The run time of a typical robust DFT code has polynomial time complexity [15]. At this pace, it is infeasible to predict properties of all possible materials in a reasonable time. Material scientists during the last few decades resorted to finding faster and accurate computational methods in comparison to DFT. The machine-learning (ML) approach is found to be a viable solution to this problem [16].

1.3 MACHINE LEARNING FOR MATERIALS PREDICTION

Machine-learning, a sub-field of artificial intelligence, uses data to build efficient models [17]. ML models relying on DFT data are sufficiently accurate and faster than high-throughput methods and enable a further speedup. ML models scale linearly $\mathcal{O}(N)$ with system size and can be as accurate as DFT. During the last decade, there have been many material repositories generated using DFT [4, 18–20]. Computational material scientists use existing data for data mining and material informatics. DFT data using machine learning methods help to build faster computational models for screening a large number of hypothetical, unexplored materials.

The use of machine learning methods combined with high-throughput techniques, i.e., accelerated-high-throughput (AHT), is the new paradigm for materials invention in computational materials science. In the accelerated high-throughput approach, ML models built from DFT or first-principles data are used to compute the properties of a large number of materials at a faster pace ($\mathcal{O}(N)$). The materials can then be screened in a high-throughput fashion for desired properties, and

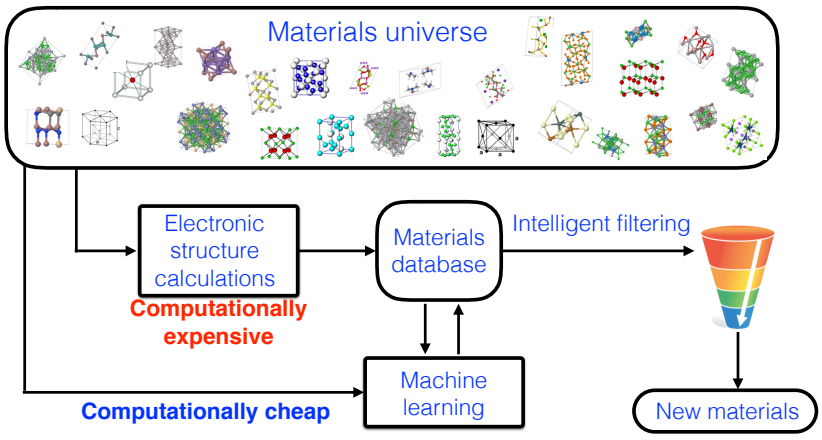


Figure 1.2 ML models built from first-principles data is used to compute properties of a large number of materials at a faster pace. The materials can then filtered in a high-throughput fashion for desired properties, thus saving time and increasing the speed of discovery.

the selected materials can be verified using accurate DFT calculations, thus increasing the speed of discovery (see fig. 1.2).

In the last few years, several machine learning models have been proposed in the literature [5–7, 22–27]. These machine-learned surrogate models to first-principles calculations consist of two parts. The first part is creating a unique representation of crystal structure [6, 24]. The second part is the learning algorithm used i.e, *regression*, which is traditionally one of the algorithms available in computer science literature.

Every machine-learned model proposed in the literature approaches the problem with a different representation. It is important to know if all ML models are consistent with each other in predicting the properties of materials. The second project [28] in this dissertation work aims to answers this question explicitly wherein five different state-of-the-art machine-learned surrogate models to DFT are used for predicting the formation enthalpy[3] of 10 different materials. All five ML models agree qualitatively of the prediction errors and, the models are shown to be capable of predicting formation enthalpies of multiple materials simultaneously.

1.4 IMPACT AND FOLLOW-UP WORK

1.4.1 High-throughput approach

Of the six promising candidate superalloys proposed in the high-throughput work, two of the alloys namely, Co-Ta-V and Co-Nb-V, were experimentally made by David Dunand's group at Northwestern University [14]. The $L1_2$ (derivative of face-centered cubic (fcc) unit cell) phase precipitate was experimentally observed in both alloys but was metastable, decomposing into other phases after two hours. Further study is necessary to understand how to stabilize the precipitates. A possible solution is to understand the effect of adding a fourth or fifth element to the ternary alloy.

One of the main problems in super-alloys research is that conventional superalloys comprise of more than five elements. It is computationally expensive to probe alloys

[3]The Formation enthalpy is the difference between the total energy of a compound and the sum of total energies of the corresponding stable, pure concentration elements constituting the compound.

beyond ternary combinations due to the sheer number of combinatoric possibilities. Nevertheless, such an effect of ternary and quaternary additions needs to be studied. In this regard, DFT calculations studying the effect of 10 different elements additions to Al-Co-W alloy at ∼6 at.%, ∼8 at.%, and ∼12 at.% is studied and discussed in the follow-up work in Chapter 2.

1.4.2 Machine learning approach

Analyzing the prediction errors of five state-of-the-art machine learning models revealed that all ML models agree qualitatively on the prediction errors of 10 different materials. An interesting result from this second project is that materials which are hard to learn (prediction errors are high) are hard for all ML models and materials which are easy to learn are easy for all models. It is important to analyze why some materials are hard to learn than others.

In this regard, as follow-up work, a study on 45 more binary alloy systems was performed. The results showed that harder systems tend to have a wider range of formation enthalpies than easy-to-learn systems. A possible solution is to either add more number of parameters to the model or add more training data which can help reduce the error for these hard systems. Adding more parameters when data is not big can lead to overfitting. This begs the question, how does one pick the least amount of additional training data which can help to improve errors and build efficient models? In this regard, three methods for intelligent selection of training data are explored in an *active learning* framework. This work is discussed in chapter 3.

1.5 SUMMARY

The following chapters in this dissertation focus on the understanding and usage of two computational methods, namely, the high-throughput approach and machine-learning approach for materials prediction. The contents of chapter 2 on HT approach, and chapter 3 on the ML approach are centered around peer-reviewed articles, typeset in the style of the journal in which they were published. Chapters, 2 and 3 also include follow-up work on these two published projects, answering further questions Chapter 4 proposes future work.

CHAPTER 2

A high-throughput search for ternary superalloys

2.1 SUPERALLOYS

The strength of materials is one of the critical properties determining the ability of a material to withstand load without permanent deformation or catastrophic failure. We carried out a computational high-throughput study to identify new candidate materials with high-temperature strength called superalloys. Superalloys are used in a wide variety of applications such as jet engines, solar power plants, chemical industries, and others. An alloy that is effective at temperatures of 500° Celsius and above can be classified as a high-temperature alloy [12]. Conventional superalloys consist of a mixture of at least one base metal (nickel, cobalt, or iron) and few other metals. Superalloys which can withstand even higher temperatures than currently known materials would enable us to increase the efficiency of machines by raising the operating temperature. For example, if we operate jet engines even one degree hotter than current temperature limits, the engine efficiency increases and can lead to saving billions of dollars on fuel.

2.2 STRUCTURE-PROPERTY RELATIONSHIP

Applications of superalloys require the materials to be tough and exhibit great strengths at high temperatures [12]. A balance of strength with ductility gives toughness to a material [29]. From a computational material scientist's perspective, the properties of a material are related to the crystal structure of the material. Closely packed crystal structure makes it easy for planes of atoms to slide by each other and allow more plastic deformation than non-closely packed structures, thereby making them more ductile materials. For this reason, a face-centered crystal (fcc) structure exhibits more ductility than a body-centered structure (bcc). Another factor that contributes to ductility is the symmetry of structures. Cubic lattice structures, because of their symmetry provides closely packed planes in several directions compared to a hexagonal close-packed (hcp) structure, which is less symmetric (see fig. 2.1).

Most conventional high-temperature superalloys are nickel-based alloys. They usually contain an $L1_2$ (derivative of face-centered cubic (fcc) unit cell) phase precipitates within an fcc host matrix (cobalt, nickel, or iron) (see fig. 1.1). These precipitates inhibit the dislocation motion in the material resulting in higher-strength. Dislocation is a beautiful phenomenon in materials science. When small in number they cause the material to be weak, but when their number increases, they increase the strength of the materials as dislocations inhibit the propagation of other dislocations (see fig. 2.2b). The relation between the structure and strength of a material is one of the crucial factors enabling the computational high-throughput study of superalloys in this dissertation work. The quest for new superalloys from a computational perspective carried out in this work is a search for stable $L1_2$ precipitates (ternary, quaternary, or more) that can form within the base elements—cobalt, nickel or iron (see fig. 2.2d).

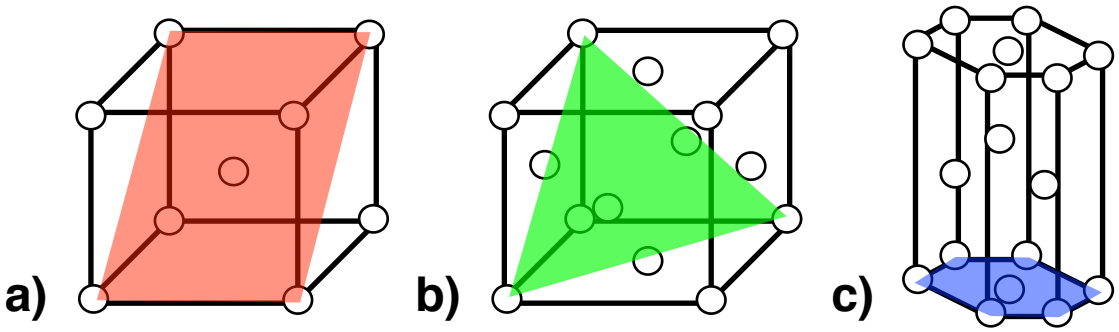


Figure 2.1 Depiction of major slip system in a) bcc, b) fcc, and c) hcp crystal structures. Only one major slip plane is shown for each of the crystal structure. {110} (red) for bcc, {111} (green) for fcc and (0001) (blue) for hcp. Dislocation motion can easily occur along the slip planes.

2.3 PROJECT BACKGROUND

In 2006, Sato *et al*., [21], reported a new cobalt-based superalloy (Co-Al-W) which has better high-temperature strength than that of conventional nickel-based superalloys. It was reported that the Co-Al-W superalloy exhibits coherent $L1_2$ Co_3(Al,W) (γ') precipitates in an face-centered cubic Co (γ) matrix. However, experimental observations suggested that the precipitate in Co_3(Al, W) system is metastable at 1173 K [31]. A theoretical investigation of the experimentally observed precipitate was carried out by Saal and Wolverton [32]. Saal and Wolverton modeled the precipitate (γ') at high-temperatures using a crystal structure called a *special quasi-random structure* (SQS)[1] [34]. Using the SQS approach, including the finite temperature contributions and point defect energetics, Saal and Wolverton found that the experimentally observed precipitate composition is consistent with a 32-atom, $L1_2$ like random SQS structure (see fig. 2 in journal article) with stoichiometry $Co_3[Al_{0.5}, W_{0.5}]$, and that the structure is metastable (66 meV/atom above convex hull).

2.4 PROJECT

Using this 32-atom, $L1_2$-like random structure (γ') reported by Saal and Wolverton, we performed an extensive combinatorial (high-throughput) search over 2224 ternary systems using the `AFLOW` [18, 35–37] framework for such stable precipitates(γ'). Our descriptors for screening potential precipitate in superalloys included formation enthalpy, decomposition energy, coherency with the host matrix, and lattice mismatch with the host matrix. Screening the 2224 systems with respect to the descriptors in comparison to the descriptor values of $Co_3[Al_{0.5}, W_{0.5}]$, we found 102 systems better than Al-Co-W system. Of these 102 systems, 37 are brand new and have no published phase diagrams. Further, based on cost and toxicity, we prioritized six most promising candidates. All the 102 materials are patented [13]. We also computed the bulk

[1]For a finite number of atoms (N), a SQS (special periodic *quasirandom crystal structure*) structure mimics the correlation functions of an infinite substitutional random alloy far more closely than does the standard approach of occupying each of the lattice sites randomly [33].

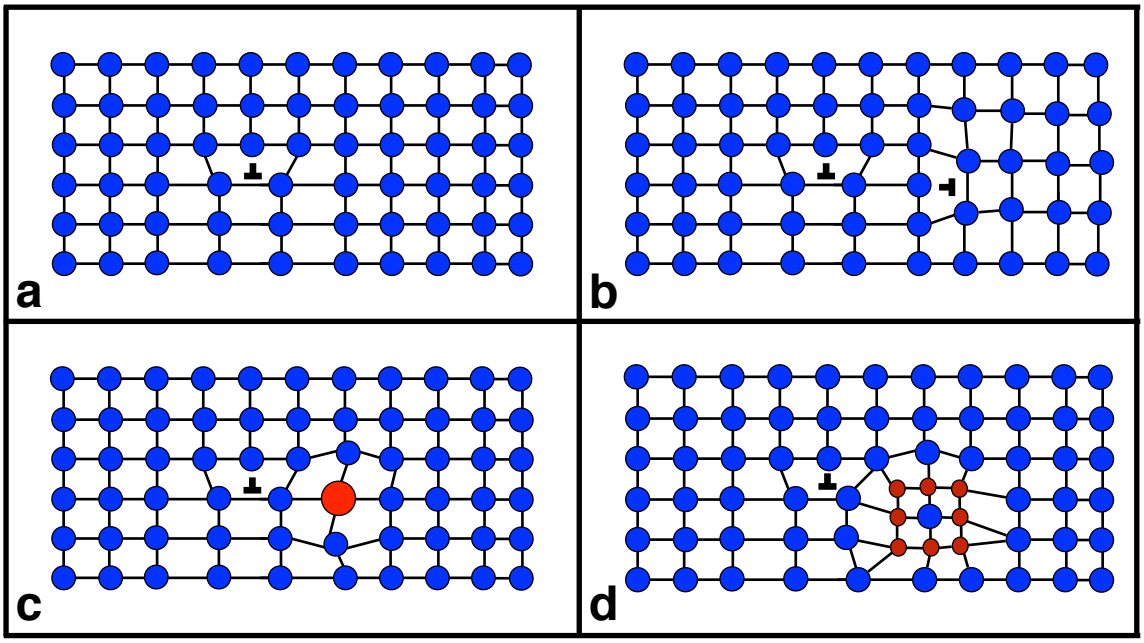


Figure 2.2 (All the figures are inspired by similar figures in reference [30]) a) *Single dislocation in a crystal:* The irregularity in the crystal marked in the figure shows a dislocation. The dislocation in the figure is an edge dislocation where an extra half-plane of atoms is introduced midway through the crystal. b) *Multiple dislocations in a crystal:* Movement of one dislocation is inhibited by the other. More dislocations in a materials increases its strength. c) *Solid-solution strengthening:* We can see a different species of atom (red) replacing the host atom (blue) within the crystal structure. The solute (red) atom helps inhibits the dislocation motion strengthening the material. d)*Precipitate strengthening:* The red and blue atom is the precipitate within the host matrix of blue atoms. The distortion in the lattice created by the precipitate helps strengthen the material. Superalloys materials contain $L1_2$ precipitates within a fcc structure.

modulus and density for these materials. All the details of the work are given in the paper attached in the following pages.

2.4.1 My contribution

In regards to this project, I conceived the idea, developed a code for computing ternary convex hull along with Jake Hansen (an undergraduate student that I mentored). I wrote scripts for collecting the data, generating all figures (except figure 1, and A10), generating the convex hull, and computing the density, and bulk modulus of all materials. I also analyzed the results, did an extensive literature search (ASM phase diagrams), wrote a significant portion of the paper, and followed up with the responses to reviewers. Corey Oses helped refine the text, figures, supplementary material, and response to reviewers. Ichiro Takeuchi provided his expertise in interpreting the results. Stefano Curtarolo performed the AFLOW calculations, provided his expertise

in interpreting results, and helped with the paper. Gus Hart guided the whole project, contributed many ideas, helped write the paper, and response to reviewers.

Acta Materialia 122 (2017) 438–447

Contents lists available at ScienceDirect

Acta Materialia

journal homepage: www.elsevier.com/locate/actamat

Full length article

A computational high-throughput search for new ternary superalloys

Chandramouli Nyshadham [a], Corey Oses [b], Jacob E. Hansen [a], Ichiro Takeuchi [c], Stefano Curtarolo [b, d], Gus L.W. Hart [a, *]

[a] Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602, USA
[b] Center for Materials Genomics, Duke University, Durham, NC 27708, USA
[c] Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA
[d] Materials Science, Electrical Engineering, Physics and Chemistry, Duke University, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form
12 August 2016
Accepted 12 September 2016
Available online 1 November 2016

Keywords:
First-principles calculations
Superalloys
High-throughput
Phase stability

ABSTRACT

In 2006, a novel cobalt-based superalloy was discovered [1] with mechanical properties better than some conventional nickel-based superalloys. As with conventional superalloys, its high performance arises from the precipitate-hardening effect of a coherent $L1_2$ phase, which is in two-phase equilibrium with the fcc matrix. Inspired by this unexpected discovery of an $L1_2$ ternary phase, we performed a first-principles search through 2224 ternary metallic systems for analogous precipitate-hardening phases of the form $X_3[A_{0.5},B_{0.5}]$, where X = Ni, Co, or Fe, and $[A,B]$ = Li, Be, Mg, Al, Si, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn Ga, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, or Tl. We found 102 systems that have a smaller decomposition energy and a lower formation enthalpy than the Co_3(Al, W) superalloy. They have a stable two-phase equilibrium with the host matrix within the concentration range $0 < x < 1$ ($X_3[A_x,B_{1-x}]$) and have a relative lattice mismatch with the host matrix of less than or equal to 5%. These new candidates, narrowed from 2224 systems, suggest possible experimental exploration for identifying new superalloys. Of these 102 systems, 37 are new; they have no reported phase diagrams in standard databases. Based on cost, experimental difficulty, and toxicity, we limit these 37 to a shorter list of six promising candidates of immediate interest. Our calculations are consistent with current experimental literature where data exists.

1. Introduction

Materials scientists have developed large experimental databases of known materials over the last century [2–5]. Similar computational databases are being compiled by exploiting the power of supercomputers and advanced electronic structure methods [6–11]. The challenge now is to leverage the data to discover new materials by building computational models [12] and employing machine learning methods [13–16]. Data mining and materials informatics approaches can also be used to identify structure/property relationships, which may suggest atomic combinations, stoichiometries, and structures not included in the database [12].

An emerging area in materials science is the computational prediction of new materials using high-throughput approaches [6,12–14,17–20]. Hundreds of thousands of hypothetical candidates can be explored much faster than by experimental means. In this work, a simple combinatorial search for ternary superalloys is performed in a high-throughput fashion. The extraordinary mechanical properties of superalloys at high temperatures make them useful for many important applications in the aerospace and power generation industries. One of the basic traits of superalloys is that they generally occur in a face-centered-cubic structure [21]. The most common base elements for superalloys are nickel, cobalt, and iron, but most are nickel-based. In 2006, a new cobalt-based superalloy, Co_3(Al, W), was confirmed to have better mechanical properties than many nickel-based superalloys [1].

This cobalt-based superalloy has the commonly occurring $L1_2$ phase which creates coherent precipitates in the fcc matrix. A theoretical investigation of Co_3(Al, W) was subsequently carried out by Saal and Wolverton [22]. To model the properties of the $L1_2$ solid solution phase observed at high temperature, Saal and Wolverton used an $L1_2$-based special quasirandom structure (SQS) [23].

* Corresponding author.
E-mail address: gus.hart@gmail.com (G.L.W. Hart).

http://dx.doi.org/10.1016/j.actamat.2016.09.017
1359-6454/

In order to identify the stoichiometry of the superalloy, they performed first-principles calculations for solid solutions $Co_3[Al_x, W_{1-x}]$ with varying concentrations of Al and W. Their study includes finite temperature effects and point defect energetics. They showed that an $L1_2$-like random structure with stoichiometry $Co_3[Al_{0.5}, W_{0.5}]$ is consistent with experiment. Interestingly, their solid-solution-like $Co_3[Al_{0.5}, W_{0.5}]$ structure is metastable and predicted to have a decomposition energy of 66 meV/atom (distance from the $T = 0$ K convex hull). They show that high-temperature effects make this phase thermodynamically competitive with other competing structures at elevated temperatures. The fact that a metastable structure ($Co_3[Al_{0.5}, W_{0.5}]$) with a decomposition energy as high as 66 meV/atom at T = 0 K, is competitive with many commercially available superalloys at higher temperatures motivates our search for similar ternary systems containing an $L1_2$-like solid solution phase.

Ideally, a computational search over potential superalloys would model actual engineering observables (e.g., hardness) and consider the influence of small concentrations of impurities, finite temperature effects, influence of vacancies, effects of polycrystallinity, etc. Unfortunately, such calculations are extremely challenging even for a single material and impractical for thousands of candidate systems as in this work.

In known superalloy systems, $L1_2$-based phases have large negative formation enthalpies, a small decomposition energy, and a relatively small lattice mismatch between the host matrix and the precipitate phase. Our search is for new ternary systems with these same metrics. We further screen candidate alloy systems for $L1_2$ precipitates either in two-phase equilibrium with the host matrix or likely to precipitate as metastable phases. Based on the relative lattice mismatch between the host element and the precipitate phases any compound with a relative lattice mismatch of >5% is excluded.

Using the solid-solution-like structure identified by Saal and Wolverton [22], we performed an extensive combinatorial search over 2224 ternary systems using the AFLOW framework [7,8]. We found 102 systems that are more stable (closer to the $T = 0$ K convex hull) and have a lower formation enthalpy than the $Co_3[Al_{0.5}, W_{0.5}]$ superalloy. All 102 systems are in two-phase equilibrium with the host matrix and have a relative lattice mismatch of less than or equal to 5%. Of these systems, 37 are new—they have no reported phase diagrams [4,5,24]. These new candidates, narrowed from thousands of possibilities, suggest experimental exploration for identifying new superalloys. Furthermore, by eliminating systems that are experimentally difficult to make or contain expensive or toxic elements, we identify six particularly promising systems.

2. Methodology

2.1. First-principles structure calculations

We performed first principles calculations using the software package AFLOW [7]. To model an $L1_2$-based solid solution, we used a 32-atom special quasirandom structure (SQS-32) [22,23,25] of the form $X_3[A_{0.5},B_{0.5}]$, where X is one of the base elements, nickel (Ni), cobalt (Co) or iron (Fe) (refer Fig. 1). These combinations lead to 780 different ternary structures for each base element totaling to 2340 SQS structures in 2224 different ternary systems.

All the calculations follow the AFLOW [26] standard, are hosted in the AFLOW repository [8], and can be easily accessed by using the

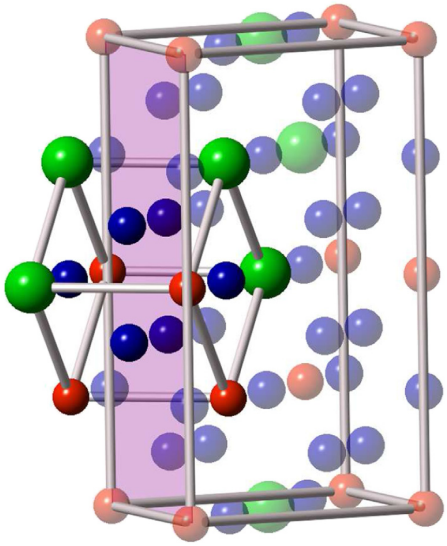


Fig. 2. The 32-atom special quasirandom structure (SQS-32) [25] used to model a solid solution with an $L1_2$ structure (smaller cube in the figure). The blue, red, and green atoms correspond to X, A, and B in $X_3[A_{0.5},B_{0.5}]$, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

1	2											13	14	15
3 Li [12]	4 Be [77]											5 B [86]	6 C [95]	7 N [100]
11 Na [11]	12 Mg [73]	3	4	5	6	7	8	9	10	11	12	13 Al [80]	14 Si [85]	15 P [90]
19 K [10]	20 Ca [16]	21 Sc [19]	22 Ti [51]	23 V [54]	24 Cr [57]	25 Mn [60]	26 Fe [61]	27 Co [64]	28 Ni [67]	29 Cu [72]	30 Zn [76]	31 Ga [81]	32 Ge [84]	33 As [89]
37 Rb [9]	38 Sr [15]	39 Y [25]	40 Zr [49]	41 Nb [53]	42 Mo [56]	43 Tc [59]	44 Ru [62]	45 Rh [65]	46 Pd [69]	47 Ag [71]	48 Cd [75]	49 In [79]	50 Sn [83]	51 Sb [88]
55 Cs [9]	56 Ba [14]	57 La* [33]	72 Hf [50]	73 Ta [52]	74 W [55]	75 Re [58]	76 Os [63]	77 Ir [66]	78 Pt [68]	79 Au [70]	80 Hg [74]	81 Tl [78]	82 Pb [82]	83 Bi [87]

6 — atomic number
C — element
[95] — Mendeleev number

Fig. 1. For each base element in $X_3[A_{0.5},B_{0.5}]$, there are 40 elements chosen for A and B, which includes 38 elements (highlighted in blue) chosen from the periodic table and the remaining two of three base elements X (highlighted in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

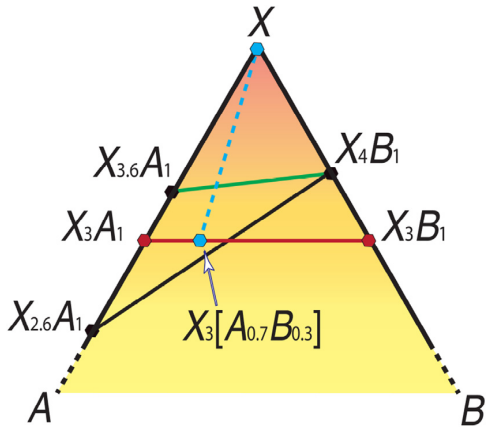


Fig. 3. The two-phase equilibrium screening criterion discussed in Sec. 2.3 (similar to Fig. 2 in Ref. [42]). If a tieline between the host matrix and the $L1_2$ precipitate phase (light blue dotted line) is intersected by the tie line for another phase (e.g., green line between $X_{3.6}A_1$ and X_4B_1) then the precipitate phase will not be in two-phase equilibrium with the host matrix for any concentration between X_3A_1 and X_3B_1. On the other hand, even if the line connecting X_3A_1 and X_3B_1 is intersected by another tie line (e.g., black line between $X_{2.6}A_1$ and X_4B_1), there may still be a concentration of the precipitate phase that can be in two-phase equilibrium with the host matrix, as show by the light-blue dotted line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Ref. [26].

The special quasirandom structure (SQS) [23] approach mimics the statistics of a random alloy in a small supercell [36]. Fig. 2 depicts the 32-atom SQS [25] that was used for all calculations in this work. It is an $L1_2$-based structure where X atoms (blue) are on the face centers of the conventional fcc cell and A (red), B (green) atoms on the corners.

2.2. Thermodynamic property calculations

The formation enthalpy (ΔH_f) is calculated for any ternary structure $X_3[A_{0.5},B_{0.5}]$ as

$$\Delta H_f = E(X_3[A_{0.5}, B_{0.5}]) - \sum_m E_m,$$

where $E(X_3[A_{0.5},B_{0.5}])$ is the total energy per atom of the SQS-32–$X_3[A_{0.5},B_{0.5}]$ structure, and $\sum_m E_m$ is the sum of total energies of the corresponding stable, pure concentration structures. A negative formation enthalpy characterizes a system that prefers an ordered configuration over decomposition into its pure constituents, while unstable systems have a positive formation enthalpy.

To approximate the phase diagram of a given alloy system, we consider the low-temperature limit in which the behavior of the system is dictated by the ground state [37,38]. In compositional space, the set of ground state configurations defines the mini-

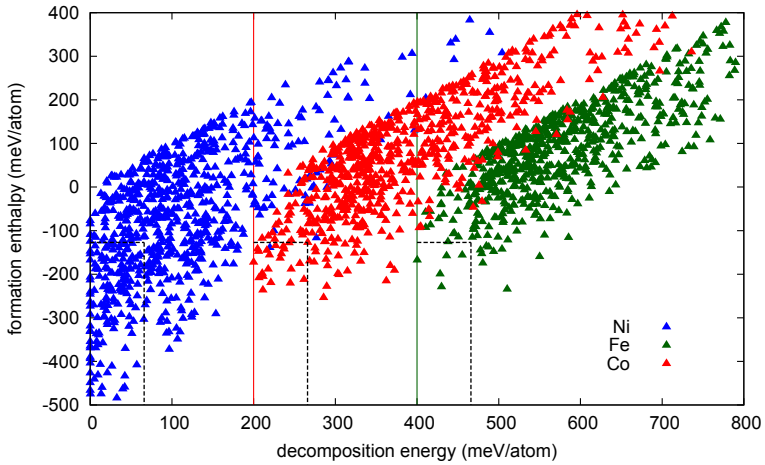


Fig. 4. Formation enthalpy vs. the decomposition energy for all 2224 ternary systems. Each triangle represents one $Ni_3/Co_3/Fe_3[A_{0.5},B_{0.5}]$ structure, where A and B are any two different elements in the periodic table from Fig. 1. Co-based and Fe-based systems are displaced on the x-axis by 200 meV and 400 meV, respectively, for clarity. Ni-based, Co-based, and Fe-based systems are marked in blue, red, and green triangles, respectively. Systems enclosed within dotted lines are the ones identified to be better than the $Co_3[Al_{0.5}, W_{0.5}]$ structure with respect to these properties. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

RESTAPI [27]. Each *ab-initio* calculation is performed using PAW potentials [28–30] within the generalized gradient approximation of Perdew, Burke, and Ernzerhof [31,32], as implemented in VASP [33,34]. The k-point meshes for sampling the Brillouin zone are constructed using the Monkhorst-Pack scheme [35]. A total number of at least 10,000 k-points per reciprocal atom are used, and spin polarization [26] is considered. The cutoff energy is chosen to be 1.4 times the default maximum value of the three elements in the respective ternary system. More details are available in

mum energy surface, also referred to as the lower-half convex hull. Compounds above the minimum energy surface are not stable, with the decomposition described by the facet directly below each. The energy gained from this decomposition is geometrically represented by the distance of the compound from the hull and quantifies the compound's tendency to decompose. We refer to this quantity as the decomposition energy.

While the minimum energy surface changes at finite temperature (favoring disordered structures), we expect the $T = 0$ K

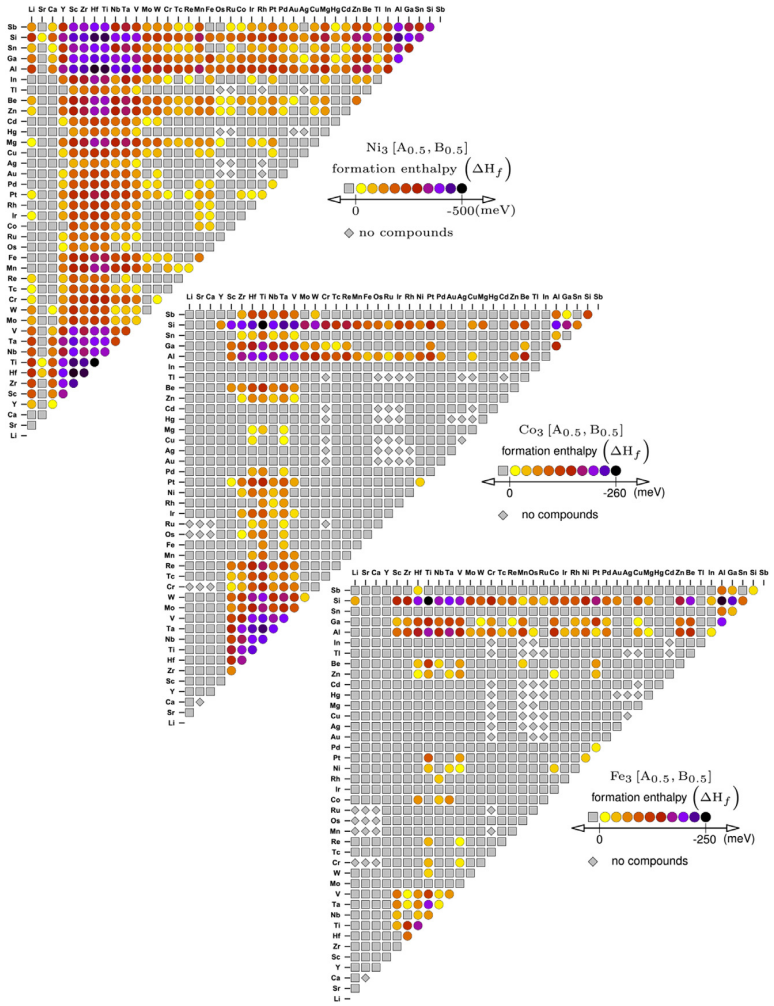


Fig. 5. All the elements are arranged as per the chemical scale (χ) introduced by Pettifor [44] in increasing order. Each diamond, square, and circle represents a ternary combination $X_3[A_{0.5},B_{0.5}]$ with X = Ni, Co, or Fe, and A,B specifying the elements indicated along the x and y-axes, respectively. A square indicates that the SQS-32 crystal structure has a positive formation enthalpy. A diamond indicates that there exists no stable binary or ternary compounds in the respective ternary system. A colored circle indicates that the SQS-32 structure has a negative formation enthalpy. The color contrast from yellow to black indicates decreasing formation enthalpy of the crystal structure in the ternary system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

decomposition energy to serve as a reasonable descriptor for relative stability. The ternary convex hulls and relevant calculations were performed[1] using the phase diagram module within AFLOW[7] (see Appendix for details).

We observe that ternary $L1_2$ phases in known superalloys have large negative formation enthalpies and appear near each other in Pettifor-like maps of the formation enthalpy and decomposition energy. Decomposition energy and formation enthalpy maps comprising all 2224 systems considered in this study are shown in Figs. 5 and 6. All those systems for which decomposition energy and formation enthalpy are less than that of $Co_3[Al_{0.5},W_{0.5}]$ are included in our list of potential candidates.

2.3. Coherency and two-phase equilibrium with the host

Because the strain energy cost is lower, compounds with smaller lattice mismatch between the $L1_2$ phase and the host matrix are more likely to form coherent precipitates. Relative lattice mismatch ($\Delta a/a_{\text{host}}$) is defined as the ratio of the difference between the lattice parameter of the host matrix and the precipitate compound, Δa, to the lattice parameter of host matrix, a_{host}. In this work, a relative lattice mismatch cutoff of no more than 5% is used to screen for potential superalloys.

Because precipitate strengthening is the key mechanism for superalloy performance, we apply another constraint requiring that the $L1_2$ precipitate phase be in two-phase equilibrium with the fcc

[1] We found in our calculations that the formation enthalpy of two compounds, namely Al_2Co and Al_2Fe with Be_2Zn structure (the prototype numbered 549 in AFLOW[8]), is anomalously low (< -1.8 eV/atom). Similar results with this Be_2Zn structure for other compounds were discussed previously by Taylor et al. [39]. They attribute the erroneous results to PAW-pseudopotentials distributed with VASP. The phase diagrams for systems with binary combinations (Al,Co) or (Al,Fe) are generated discarding the Be_2Zn structure in this work.

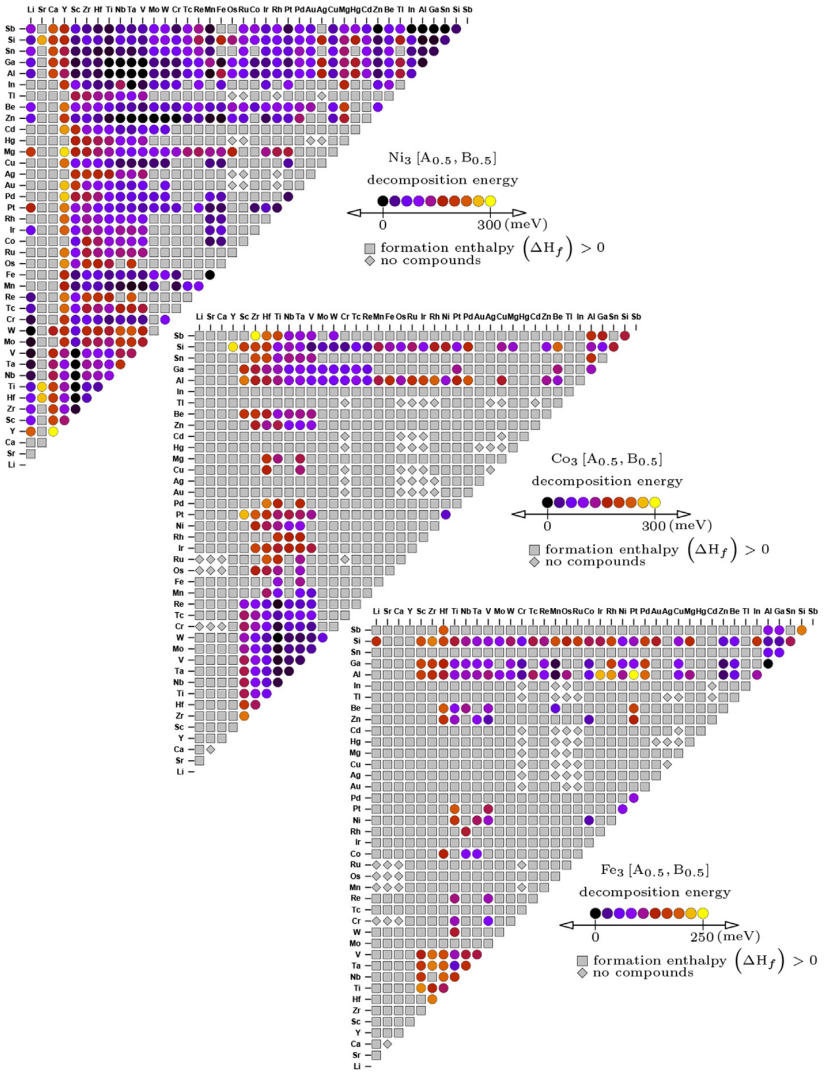


Fig. 6. All the elements are arranged as per the chemical scale (χ) introduced by Pettifor [44] in increasing order. Each diamond, square, and circle represents a ternary combination $X_3[A_{0.5},B_{0.5}]$ with X = Ni, Co, or Fe, and A,B are the elements indicated along the x and y-axes, respectively. A square indicates that the SQS-32 crystal structure has a positive formation enthalpy. A diamond indicates that there exists no stable binary or ternary compounds in the respective ternary system. The color contrast of the circles from yellow to black indicates increasing decomposition energy of the crystal structure in the ternary system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

host matrix.[2] As shown in Fig. 3, this constraint is satisfied if a tie-line can be drawn between the host matrix (100% X) and the $L1_2$ phase at any concentration ($X_3[A_x,B_{1-x}]$, $0 < x < 1$) without intersecting any other tieline. We allow for this variation in the concentration for the minority site (A_x,B_{1-x}) because stable $L1_2$ phases in experiment can vary over a wide concentration range [40,41]. Of the 179 systems with deeper formation enthalpy and smaller decomposition energy than Co-Al-W, 66 systems are eliminated using the two-phase equilibrium criterion.

[2] In cases where the formation enthalpy of the SQS structure is above the convex hull, we project it onto the convex hull and draw the tieline between the projected point and the host matrix to check the two-phase equilibrium criterion.

2.4. Bulk modulus calculations

The bulk modulus is determined from energy-volume data calculated for strains of −0.02 Å to +0.02 Å in steps of 0.01 Å applied to the unit cell, with at least five calculations for each system. The energy-volume data is fitted using the Murnaghan equation of state [43].

3. Results and analysis

3.1. Relative stability of SQS-32 and the distance to convex hull

Fig. 4 depicts the formation enthalpy (ΔH_f) vs. decomposition energy (E_d) for all 2224 SQS-32 ternary systems with composition

Table 1
Systems where the SQS structure computed in this work has a corresponding $L1_2$ phase reported in experiment. The experimental compounds are all close to the stoichiometry of the SQS structure, $X_{24}[A_4,B_4]$.

SQS	Exp.
$Al_{0.5}$ $Cr_{0.5}$ Ni_3	$Al_{0.8}$ $Cr_{0.2}$ Ni_3[50]
$Al_{0.5}$ $Cu_{0.5}$ Ni_3	Al_1 $Cu_{0.28}$ $Ni_{2.72}$[51]
$Al_{0.5}$ $Ga_{0.5}$ Ni_3	$Al_{0.5}$ $Ga_{0.5}$ Ni_3[52]
$Al_{0.5}$ $Hf_{0.5}$ Ni_3	$Al_{0.99}$ $Hf_{0.01}$ Ni_3[53]
$Al_{0.5}$ $Nb_{0.5}$ Ni_3	$Al_{0.65}$ $Nb_{0.35}$ Ni_3[54]
$Al_{0.5}$ Ni_3 $Pt_{0.5}$	Al_1 $Ni_{2.48}$ $Pt_{0.52}$[51]
$Al_{0.5}$ Ni_3 $Si_{0.5}$	$Al_{0.6}$ Ni_3 $Si_{0.4}$[51]
$Al_{0.5}$ Ni_3 $Sn_{0.5}$	$Al_{0.8}$ Ni_3 $Sn_{0.2}$[50]
$Al_{0.5}$ Ni_3 $Ta_{0.5}$	$Al_{0.76}$ Ni_3 $Ta_{0.24}$[41]
$Al_{0.5}$ Ni_3 $Ti_{0.5}$	Al_1 $Ni_{2.8}$ $Ti_{0.2}$[40]
$Al_{0.5}$ Ni_3 $V_{0.5}$	$Al_{0.28}$ Ni_3 $V_{0.2}$[50]
Co_3 $Ti_{0.5}$ $V_{0.5}$	Co_3 $Ti_{0.87}$ $V_{0.13}$[55]
$Ga_{0.5}$ Hf_4 Ni_3	$Ga_{0.88}$ $Hf_{0.12}$ Ni_3[51]
$Ga_{0.5}$ Nb_4 Ni_3	$Ga_{0.84}$ $Nb_{0.16}$ Ni_3[51]
$Ga_{0.5}$ Ni_3 $Sb_{0.5}$	$Ga_{0.92}$ Ni_3 $Sb_{0.08}$[51]
$Ga_{0.5}$ Ni_3 $Si_{0.5}$	$Ga_{0.4}$ Ni_3 $Si_{0.6}$[51]
$Ga_{0.5}$ Ni_3 $Sn_{0.5}$	$Ga_{0.84}$ Ni_3 $Sn_{0.16}$[51]
$Ga_{0.5}$ Ni_3 $Ta_{0.5}$	$Ga_{0.68}$ Ni_3 $Ta_{0.32}$[51]
$Ga_{0.5}$ Ni_3 $Ti_{0.5}$	$Ga_{0.84}$ Ni_3 $Ti_{0.16}$[51]
$Ga_{0.5}$ Ni_3 $V_{0.5}$	$Ga_{0.76}$ Ni_3 $V_{0.24}$[51]

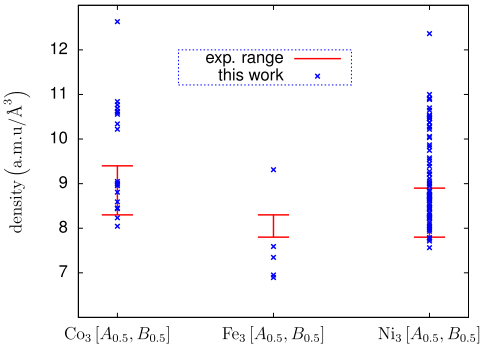


Fig. 7. A comparison between the density range for the theoretical calculations performed in this work and modern superalloys. Densities are computed for 102 ternary systems screened from the 2224 systems computed in this work. The red line shows the range of density for commercially-available superalloys at present.

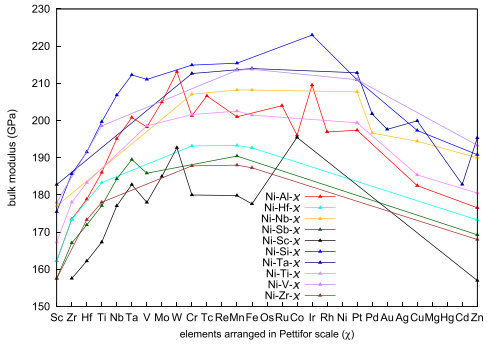


Fig. 8. The magnitude of the bulk modulus for Ni-A-x (A = Al, Hf, Nb, Sb, Sc, Si, Ta, Ti, V, and Zr) systems with the x-axis arranged according to the χ scale in Pettifor maps. In general, the systems display a maximum in the bulk modulus at or before Ni. Only systems with simultaneously lower E_d and ΔH_f than $Co_3[Al_{0.5},W_{0.5}]$ are plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

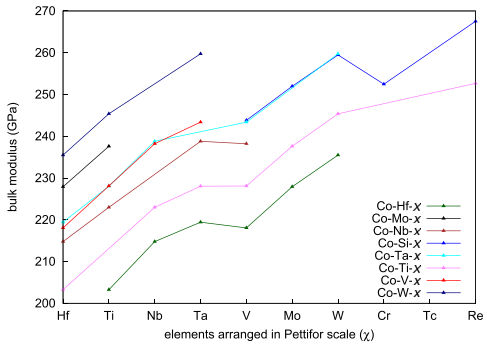


Fig. 9. The magnitude of the bulk modulus for Co-A-x (A = Hf, Mo, Nb, Si, Ta, Ti, V, and W) systems with the x-axis arranged according to the χ scale in Pettifor maps. In general, the magnitude of the bulk modulus increases with χ up to Re. Only systems with simultaneously lower E_d and ΔH_f than $Co_3[Al_{0.5}, W_{0.5}]$ are plotted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distinguished by color. It is found that 2111/2224 ternary systems are compound-forming. Each point on the plot represents one Ni_3/Co_3/$Fe_3(A,B)$ system, where A and B are any two different elements highlighted in Fig. 1. On average, Ni-based superalloys are thermodynamically more stable than Co- or Fe-based superalloys.

The SQS-32 structure in 179 ternary systems is found to be thermodynamically more stable and have lower formation enthalpy than the Co_3(Al, W) system. These systems are enclosed within dotted lines in Fig. 4. Out of these systems, 152 are Ni-based, 22 are Co-based, and 5 are Fe-based. Furthermore, 102 systems of these 179 are observed to be in two-phase equilibrium with the host matrix and have no more than 5% relative lattice mismatch with respect to the respective host lattice. Of these 102 systems, 37 have no reported phase diagrams in standard databases [4,5,24]. Of these systems, 33 are Ni-based, 3 are Co-based systems, and 1 is Fe-based.

The magnitude of ΔH_f is closely associated with the high temperature limit of an alloy. If a compound has a large negative formation enthalpy, it is more likely to withstand decomposition at higher temperatures. Fig. 4 shows that many Ni-based alloys are as low as −400 meV compared to −167 meV of the discovered Co_3(Al, W) superalloy [22].

Although the elemental form of Fe is bcc, fcc stabilizers (e.g., carbon, tungsten, or nickel) can be added in small amounts to stabilize the fcc structure. We have modeled Fe-based systems with $L1_2$ precipitate-forming potential by calculating fcc Fe, without explicitly including the effects of the stabilizing additions. Had we found promising Fe systems, this rough approximation would have needed refinement, but all of our promising candidates but one turned out to be Co- or Ni-based.

3.2. Formation enthalpy and decomposition energy maps

Recognizing that ternary $L1_2$ phases in known superalloys have large negative formation enthalpies and small decomposition energies, it is useful to identify chemical trends (via the Pettifor chemical scale) for these two quantities. We visualize these trends with Pettifor-like “formation enthalpy maps” and “decomposition energy maps” (Figs. 5 and 6). In the formation enthalpy maps, the formation enthalpy of every system computed in this work is displayed together, arranged in a grid ordered by the Pettifor scale [44] of the two minority components, A,B in $X_3[A_{0.5},B_{0.5}]$. In a similar

Table 2
Candidates for precipitate-forming systems that have no previously reported phase diagrams in standard databases [5,24,46,47]. These have a smaller decomposition energy and a lower formation enthalpy than the $Co_3(Al, W)$ superalloy. All are in stable two-phase equilibrium with the host matrix and have a relative lattice mismatch with the host matrix of less than or equal to 5%. Promising candidates (see Section 3.6) are boxed. '***' indicates that the quantity is not computed in this work.

System	Formation enthalpy [meV]	Decomposition energy [meV/atom]	Density [gm/cm^3]	Bulk modulus [GPa]	Relative lattice mismatch [%]
Al_4 Ni_{24} Rh_4	−189	49	8.71	197	−2
Au_4 Ni_{24} Ta_4	−142	46	12.17	198	−5
Be_4 Fe_4 Ni_{24}	−129	40	8.20	206	1
Be_4 Ga_4 Ni_{24}	−203	59	8.33	184	0
Be_4 Mn_4 Ni_{24}	−132	43	8.12	***	1
Be_4 Nb_4 Ni_{24}	−237	37	8.38	198	−1
Be_4 Ni_{24} Sb_4	−159	59	8.71	177	−2
Be_4 Ni_{24} Si_4	−298	48	7.78	201	1
Be_4 Ni_{24} Ta_4	−269	33	10.02	204	−1
Be_4 Ni_{24} Ti_4	−308	53	7.79	189	0
Be_4 Ni_{24} V_4	−225	21	8.07	203	1
Be_4 Ni_{24} W_4	−144	44	10.23	219	−1
$\boxed{Co_{24}Nb_4V_4}$	−156	19	9.05	238	−2
Co_4 Ni_{24} Sc_4	−166	55	8.04	169	−3
Co_{24} Re_4 Ti_4	−142	5	10.69	253	−2
$\boxed{Co_{24}Ta_4V_4}$	−189	18	10.62	243	−2
Fe_{24} Ga_4 Si_4	−200	28	7.59	***	−4
Ga_4 Ir_4 Ni_{24}	−129	27	11.00	209	−2
$\boxed{Hf_4Ni_{24}Si_4}$	−459	42	9.83	192	−3
In_4 Ni_{24} V_4	−165	14	8.91	182	−4
Ir_4 Ni_{24} Si_4	−184	55	10.54	223	−1
$\boxed{Mn_4Ni_{24}Sb_4}$	−151	8	9.06	184	−4
Nb_4 Ni_{24} Pd_4	−129	52	9.39	197	−4
Nb_4 Ni_{24} Pt_4	−172	48	10.89	208	−4
Nb_4 Ni_{24} Zn_4	−241	0	8.95	190	−3
Ni_{24} Pd_4 Ta_4	−160	51	10.92	202	−4
Ni_{24} Pt_4 Si_4	−228	39	10.46	211	−2
Ni_{24} Pt_4 Ta_4	−202	45	12.36	213	−4
Ni_{24} Pt_4 Ti_4	−250	58	10.38	199	−3
$\boxed{Ni_{24}Sb_4Si_4}$	−310	21	8.82	187	−3
$\boxed{Ni_{24}Sb_4Ti_4}$	−335	11	8.72	177	−5
Ni_{24} Sc_4 Zn_4	−241	39	7.97	157	−4
Ni_{24} Si_4 Sn_4	−303	26	8.76	185	−3
Ni_{24} Ta_4 Zn_4	−274	0	10.49	195	−3
Ni_{24} V_4 Zn_4	−213	0	8.66	193	−1
Ni_{24} W_4 Zn_4	−147	0	10.70	210	−2
Ni_{24} Zn_4 Zr_4	−261	48	8.61	168	−4

Pettifor scale grid fashion, the decomposition energy maps show the decomposition energy of every system computed in this work (Fig. 5). The "islands" of similarly colored compounds visible in these plots reveal distinct chemical trends. Many of the promising superalloy candidates identified in our study with no previously reported phase diagrams are found within these islands. In general, early *d*-block elements and *p*-block combinations dominate the list of favorable systems, which have both low formation enthalpies and low decomposition energies. For Ni-based alloys, favorable compounds mostly comprise of transition metals Y, Sc, Zr, Hf, Ti, Nb, Ta, and metalloids, including Ga, Si, and Sb. In the case of Co-based alloys, combinations of Zr, Hf, Ti, Nb, Ta, and Al define the majority of favorable compounds. Combinations of Al, Si, Hf, and Ti with Fe tend to produce some favorable compounds as well. On the other hand, combinations with Os, Ru, and Cr tend to yield unstable compounds for combinations with Ni, Co, and Fe.

3.3. Phase diagrams

Ternary phase diagrams at $T = 0$ K for all 2111 compound-forming systems have been plotted in this work using the data in the open-access materials properties database AFLOW[45]. Convex hulls constructed from a DFT database are only as reliable as the database is complete. To be robust, the database must include all possible structural prototypes. Our prototypes list includes essentially all known prototypes from the Pauling File [5,24] (a database of experimentally observed binary metallic phases) and binary and ternary intermetallic prototypes[3] in the ICSD [46,47]. Our prototypes list also includes binary and ternary hypothetical structures (enumerated as in Refs. [48,49]). Our convex hulls were constructed from more than 800 DFT calculations per system. In total, 271,000 calculations were used for the 2111 compound-forming systems, giving us a high degree of the confidence that the phase stability predictions and potential superalloy candidates listed in this work are reasonably likely to be stable experimentally. Further evidence of the robustness of the calculations is given in Table 1.

The ternary phase diagrams of all 2111 compound-forming systems are included in the Supplementary Material accompanying this work and are available online via http://aflow.org/superalloys. They were created with the phase diagram module within AFLOW. In almost all cases, the AFLOW convex hulls contain more phases than reported in the experimental databases. In some cases, this may indicate an opportunity for further experimental study, but it is likely that some of these DFT ground states are low temperature phases and are therefore kinetically inaccessible, which explains why they are not reported in experimental phase diagrams.

There are 66 systems which meet all our criteria discussed in

[3] Although entries with incomplete structural information or phases with partially occupied wyckoff positions obviously cannot be included.

Secs. 2.2 and 2.3 and for which there are published phase diagrams. In 20 of those systems, the predicted $L1_2$ phase is validated by an experimentally reported $L1_2$ phase at nearby concentrations. In 37 cases, the phase diagrams are incomplete in the region of interest. In the remaining eight cases, three have fcc solid solutions near our composition, three report disordered χ-like phases or unknown structures, one has a disordered $D0_{24}$ structure (closely related to $L1_2$ and a precipitate phase in some superalloys), and one reports the structure prototype $Mg_6Cu_{16}Si_7$.

3.4. Density of superalloys

Low density and high-temperature strength are two critical properties of superalloys for any application. For example, increased density can result in higher stress on mating components in aircraft gas turbines [21]. A comparison between the density range for theoretical calculations performed in this work and modern superalloys is listed in Fig. 7. Of the theoretical ternary combinations, there are 5 Ni-based alloys 2 Co-based and 4 Fe-based alloys with density less than the range of commercially-available superalloys. This certainly warrants further analysis of mechanical properties of these alloys, which may yield novel lightweight, high-strength superalloys.

3.5. Bulk modulus

For the aforementioned systems with simultaneously lower E_d and ΔH_f than the Co_3(Al, W) system, the bulk modulus is computed in this work. All the Co- and Fe-based alloys have a bulk modulus of at least 200 GPa. This is consistent with the observation that commercial Co-based alloys have better mechanical properties than many Ni-based alloys [1].

Figs. 8 and 9 depict the magnitude of the bulk modulus for Ni-*A*-*x* (*A* = Al, Hf, Nb, Sb, Sc, Si, Ta, Ti, V, and Zr) and Co-*A*-*x* (*A* = Hf, Mo, Nb, Si, Ta, Ti, V, and W) systems. *x* is the third element in the ternary system and arranged along the *x*-axis of the plot in increasing order of the Pettifor chemical scale (χ). The bulk modulus of ternary alloys of the form Ni-*A*-*x* reaches a maximum at or before Ni. In case of Co-*A*-*x* systems, the bulk modulus increases with increasing χ up to Re.

The magnitude of the bulk modulus suggests that Co-based superalloys are particularly resistant to compression compared to Ni-based superalloys. 68 ternary systems with simultaneously lower E_d and ΔH_f than $Co_3[Al_{0.5},W_{0.5}]$ have bulk moduli greater than 200 GPa.

3.6. Promising candidates

Table 2 lists the 37 systems that are predicted to have stable precipitate-forming $L1_2$ phases and for which there are no reported phase diagrams in standard databases [4,5,24]. Avoiding elements (i.e. Au, Be, Cd, Ga, Hg, Ir, In, Li, Os, Pd, Pt, Re, Rh, Ru, Sb, Sc, Tc, and Tl) that are toxic, expensive, or have low melting temperatures (which can result in difficulty incorporating them in alloy synthesis), we prioritize this list into a smaller set of six candidate superalloy systems. These are denoted by boxes in Table 2.

4. Conclusion

We used DFT calculations to search for new ternary systems with $L1_2$ precipitate-forming potential. We examined a total of 2224 different ternary systems comprising 41 different elements. The Pettifor-type formation enthalpy and decomposition energy maps (Figs. 5 and 6) introduced in this work reveal that combinations of early *d*-block and *p*-block elements tend to form stable superalloy systems with base-elements Ni, Co, and Fe. Ni-based superalloys tend to be thermodynamically more stable than Co- or Fe-based superalloys.

A total of 102 ternary systems are found to have lower formation enthalpy and decomposition energy than the recently discovered $Co_3[Al_{0.5},W_{0.5}]$ superalloy. All the systems are observed to be in two-phase equilibrium with the host matrix and have a lattice mismatch of less than or equal to 5% with the host matrix. Further analysis should be done for these systems with, e.g., cluster expansion [56–58] in the interest of experimental verification. Of these, 37 systems have no experimental phase diagram reported in literature. A comparison between the density range for our theoretical systems and modern superalloys reveal many candidate low-density superalloys. Co-based superalloys are observed to have a higher bulk modulus than Ni- and Fe-based alloys. Based on cost, experimental difficulty, and toxicity, we prioritize a shorter list of six promising superalloy systems (see Table 2).

Acknowledgments

The authors thank Eric Perim, Eric Gossett, M. Buongiorno Nardelli, M. Fornari, S. Butenko, and C. Toher for useful discussion. Funding from ONR (MURI N00014-13-1-0635). C. Oses acknowledges support from the National Science Foundation Graduate Research Fellowship under Grant No. DGF1106401. Calculations were performed at the *Duke University Center for Materials Genomics* and at the *BYU Fulton Supercomputing Lab*.

Appendix A. Phase diagram (Convex Hull) analysis

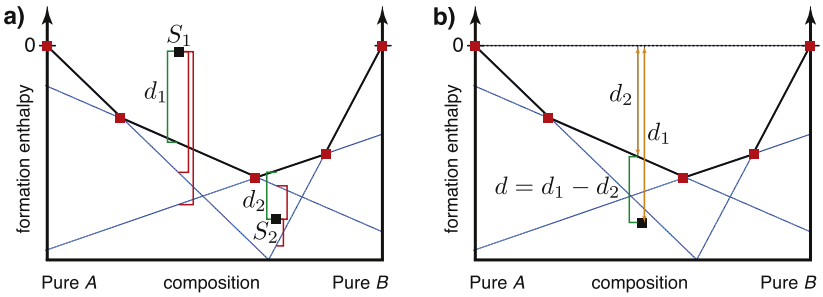


Figure A10. Distance to the $T = 0$ K convex hull algorithm. **a)** The correct distance (shown in green) for d_1 is the minimum distance of structure S_1 to all hyperplanes defining the convex hull. In case of structure S_2, the minimum distance is not d_2 (green line), an artifact of the hyperplane description for hull facets. **b)** Projecting the points to the zero energy line guarantees that all points will lie within the hull, thus enabling the use of minimization algorithm to calculate the correct distance. The distance to the hull d is given as the difference of the projected distance d_2 from the distance to the zero energy line d_1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We construct the $T = 0$ K convex hull using the phase diagram module within AFLOW[7]. Elements of this implementation were inspired by the QHULL algorithm [59]. For k-nary systems, it computes the distances to the hull with the following considerations. Let a facet of the convex hull, i.e., a hyperplane, be described by,

$$a_0 + \sum_{m=1}^{k} a_m x_m = 0. \tag{A.1}$$

Here $a_1,\ldots,a_k$ define the normals of the hyperplane while the constant a_0 uniquely defines the hyperplane in space. This is a simple extension of the familiar 3-D plane equation $Ax + By + Cz = D$. Let a k-nary structure have the coordinates denoted in k-dimensional space as $c_1,c_2,\ldots,c_k$, where $c_1,\ldots,c_{k-1}$, are the concentrations of the $k-1$ elements in a k-nary system and c_k is the formation enthalpy. Note that we neglect the concentration of the k^{th} element because it is implicit given the other $k-1$ concentrations. The distance d of the structure to a given facet of the convex hull is computed as follows,

$$d = c_n - (1/a_n)\left(- a_0 + \sum_{m=1}^{n-1} a_m c_m \right). \tag{A.2}$$

This equation is different from the nominal (shortest) distance between a plane and a point, which projects the point onto the plane along the normal vector. Instead, we want the distance that projects the point onto the plane along the energy axis.

The distance of the structure to the convex hull is the minimum of Eq. (A.2) computed for all facets of the convex hull. This minimization avoids a costly analysis of identifying the relevant facet, including the conversion of all facet vertices to barycentric coordinates. However, it is important to recognize that this minimization algorithm is only valid for compounds above the convex hull.

The correct (incorrect) distances of each structure to the convex hull is illustrated by the green (red) lines in Fig. A.10a. For structures within the convex hull, i.e., S_1, the minimum distance correctly matches the structure to the plane immediately below it. However, imagine we were interested in determining the importance/stability of a convex hull member. This property may be quantified by determining the distance of this structure from the bottom of a new pseudo-hull which does not contain the structure, such as what is illustrated by S_2. For such cases, we need a generalized distance to hull algorithm. The minimization algorithm alone would not identify the correct facet because the algorithm is dependent on the hyperplane description of the facet. Therefore, it is possible to find the imaginary extension of a distant facet to be closer to the compound than that of the correct facet. To avoid this problem, we generalize our algorithm by simply taking the projection of the point (compound) to the zero energy line, perform the minimization, and subtract the projected distance. This is illustrated in Fig. A.10b.

References

[1] J. Sato, T. Omori, K. Oikawa, I. Ohnuma, R. Kainuma, K. Ishida, Cobalt-base high-temperature alloys, Science 312 (2006) 90–91.
[2] L.L.C. MatWeb, Matweb Material Property Data, 2011. http://www.matweb.com.
[3] MatWeb, LLC, Matweb Material Property Data: http://www.matbase.com 2011.
[4] P. Villars, H. Okamoto, K. Cenzual, ASM Alloy Phase Diagram Database, 2006. http://www1.asminternational.org/AsmEnterprise/APD.
[5] P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, H. Putz, S. Iwata, The Pauling file, binaries edition, J. Alloys. Compd. 367 (2004) 293–297.
[6] S. Curtarolo, G.L.W. Hart, M. Buongiorno Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12 (2013) 191–201.
[7] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnátek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, AFLOW: an automatic framework for high-throughput materials discovery, Comp. Mat. Sci. 58 (2012) 218–226.
[8] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG: a distributed materials properties repository from high-throughput *ab initio* calculations, Comp. Mat. Sci. 58 (2012) 227–235.
[9] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, APL Mater 1 (2013) 011002.
[10] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R.S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A.M. Brockway, A. Aspuru-Guzik, The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the world community grid, J. Phys. Chem. Lett. 2 (2011) 2241–2251.
[11] M. Scheffler, C. Draxl, Computer Center of the Max-Planck Society, Garching, the NoMaD Repository, 2014. http://nomad-repository.eu.
[12] O. Levy, G.L.W. Hart, S. Curtarolo, Uncovering compounds by synergy of cluster expansion and high-throughput methods, J. Am. Chem. Soc. 132 (2010) 4830–4833.
[13] L.-F. Arsenault, O.A. von Lilienfeld, A.J. Millis, Machine Learning for Many-body Physics: Efficient Solution of Dynamical Mean-field Theory, arXiv:1506.08858, 2015.
[14] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (2015) 2326–2331.
[15] O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, Materials cartography: representing and mining materials space using structural and electronic fingerprints, Chem. Mater. 27 (2015) 735–743.
[16] J. Carrete, N. Mingo, S. Wang, S. Curtarolo, Nanograined half-Heusler semiconductors as advanced thermoelectrics: an ab initio high-throughput statistical study, Adv. Func. Mater. 24 (2014) 7427–7432.
[17] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (2015) 105503.
[18] O. Levy, M. Jahnátek, R.V. Chepulskii, G.L.W. Hart, S. Curtarolo, Ordered structures in rhenium binary alloys from first-principles calculations, J. Am. Chem. Soc. 133 (2011) 158–163.
[19] M. Jahnátek, O. Levy, G.L.W. Hart, L.J. Nelson, R.V. Chepulskii, J. Xue, S. Curtarolo, Ordered phases in ruthenium binary alloys from high-throughput first-principles calculations, Phys. Rev. B 84 (2011) 214110.
[20] O. Levy, G.L.W. Hart, S. Curtarolo, Structure maps for hcp metals from first-principles calculations, Phys. Rev. B 81 (2010) 174106.
[21] M.J. Donachie, S.J. Donachie, Superalloys: a Technical Guide, second ed., ASM International, 2002.
[22] J.E. Saal, C. Wolverton, Thermodynamic stability of Co-Al-W $L1_2\gamma'$, Acta Mater. 61 (2013) 2330–2338.
[23] A. Zunger, S.-H. Wei, L.G. Ferreira, J.E. Bernard, Special quasirandom structures, Phys. Rev. Lett. 65 (1990) 353–356.
[24] P. Villars, K. Cenzual, J.L.C. Daams, F. Hulliger, T.B. Massalski, H. Okamoto, K. Osaki, A. Prince, S. Iwata, Crystal Impact, Pauling File. Inorganic Materials Database and Design System,Binaries Edition, ASM International, Metal Park, OH, 2003.
[25] C. Jiang, Y. Du, Thermodynamic and mechanical stabilities of γ'-Ir_3(Al,W), J. Appl. Phys. 109 (2011).
[26] C.E. Calderon, J.J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M.J. Mehl, G.L.W. Hart, M. Buongiorno Nardelli, S. Curtarolo, The AFLOW standard for high-throughput materials science calculations, Comp. Mat. Sci. 108 (Part A) (2015) 233–238.
[27] R.H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. Buongiorno Nardelli, S. Curtarolo, A RESTful API for exchanging materials data in the AFLOWLIB.org consortium, Comp. Mater. Sci. 93 (2014) 178–192.
[28] G. Kresse, J. Hafner, Norm-conserving and ultrasoft pseudopotentials for first-row and transition-elements, J. Phys. Condens. Matter. 6 (1994) 8245–8257.
[29] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979.
[30] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758.
[31] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868.
[32] J.P. Perdew, K. Burke, M. Ernzerhof, Erratum: generalized gradient approximation made simple, Phys. Rev. Lett. 78 (1997) 1396.
[33] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comp. Mat. Sci. 6 (1996) 15.
[34] G. Kresse, J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169–11186.
[35] H.J. Monkhorst, J.D. Pack, Special points for Brillouin-zone integrations, Phys. Rev. B 13 (1976) 5188–5192.
[36] C. Jiang, First-principles study of ternary bcc alloys using special quasi-random structures, Acta Mater. 57 (2009) 4716–4726.

[37] S. Curtarolo, D. Morgan, G. Ceder, Accuracy of *ab initio* methods in predicting the crystal structures of metals: a review of 80 binary alloys, Calphad 29 (2005) 163–211.
[38] G.L.W. Hart, S. Curtarolo, T.B. Massalski, O. Levy, Comprehensive search for new phases and compounds in binary alloy systems based on platinum-group metals, using a computational first-principles approach, Phys. Rev. X 3 (2013) 041035.
[39] R.H. Taylor, S. Curtarolo, G.L.W. Hart, Guiding the experimental discovery of magnesium alloys, Phys. Rev. B 84 (2011) 084101.
[40] B. Huneau, P. Rogl, K. Zeng, R. Schmid-Fetzer, M. Bohn, J. Bauer, The ternary system Al-Ni-Ti part I: isothermal section at 900°C; experimental investigation and thermodynamic calculation, Intermetallics 7 (1999) 1337–1345.
[41] B.C. Giessen, N.J. Grant, New intermediate phases in transition metal systems, III, Acta Cryst. 18 (1965) 1080–1081.
[42] S. Kirklin, J.E. Saal, V.I. Hegde, C. Wolverton, High-throughput computational search for strengthening precipitates in alloys, Acta Mater. 102 (2016) 125–135.
[43] F.D. Murnaghan, The compressibility of media under extreme pressures, Proc. Natl. Acad. Sci. 30 (1944) 244–247.
[44] D.G. Pettifor, A chemical scale for crystal-structure maps, Sol. State Commun. 51 (1984) 31–34.
[45] W. Setyawan, S. Curtarolo, AflowLib: Ab-initio Electronic Structure Library Database, 2011. http://www.aflowlib.org.
[46] V.L. Karen, M. Hellenbrandt, Inorganic crystal structure database: new developments, Acta Cryst. A58 (2002) c367.
[47] I.D. Brown, S.C. Abrahams, M. Berndt, J. Faber, V.L. Karen, W.D.S. Motherwell, P. Villars, J.D. Westbrook, B. McMahon, Report of the working group on crystal phase identifiers, Acta Cryst. A61 (2005) 575–580.
[48] G.L.W. Hart, R.W. Forcade, Generating derivative structures: algorithm and applications, Phys. Rev. B 77 (2008) 224115.
[49] G.L.W. Hart, R.W. Forcade, Generating derivative structures from multilattices: algorithm and application to hcp alloys, Phys. Rev. B 80 (2009) 014120.
[50] I. Ul-Haq, J.G. Booth, Magnetic and structural properties of Ni_3 Al based alloys, J. Magn. Magn. Mater. 62 (1986) 256–268.
[51] Y. Mishima, S. Ochiai, T. Suzuki, Lattice parameters of Ni(γ), Ni_3 Al(γ') and Ni_3 Ga(γ') solid solutions with additions of transition and B-subgroup elements, Acta Metall. 33 (1985) 1161–1169.
[52] S. Ochiai, Y. Mishima, T. Suzuki, Lattice parameter data of nickel (gamma), ni sub 3 al (gamma prime) and ni sub 3 ga (gamma prime) solid solutions, Bull. Res. Lab. Precis. Mach. Electron. (1984) 15–28.
[53] P. Rao, K.S. Murthy, S. Suryanarayana, S. Naidu, Effect of ternary additions on the room temperature lattice parameter of Ni_3Al, Phys. Status Solidi A 133 (1992) 231–235.
[54] R.S. Mints, G.F. Belyaeva, Y.S. Malkov, Equilibrium diagram of the Ni_3Al-Ni_3Nb system, Russ. J. Inorg. Chem. 7 (1962) 1236–1239.
[55] Y. Liu, T. Takasugi, O. Izumi, Alloying behavior of Co_3Ti, Metall. Trans. A 17 (1986) 1433–1439.
[56] J.M. Sanchez, F. Ducastelle, D. Gratias, Generalized cluster description of multicomponent systems, Phys. A 128 (1984) 334–350.
[57] D. de Fontaine, Cluster approach to order-disorder transformations in alloys, in: H. Ehrenreich, D. Turnbull (Eds.), Solid State Physics, Volume 47, Wiley, New York, 1994, pp. 33–176.
[58] A. Zunger, First-principles statistical mechanics of semiconductor alloys and intermetallic compounds, in: A. Gonis, P. Turchi (Eds.), NATO Advanced Study Institute on Statics and Dynamics of Alloy Phase Transformations, 1994, pp. 361–419.
[59] C.B. Barber, D.P. Dobkin, H. Huhdanpaa, The quickhull algorithm for convex hulls, ACM Trans. Math. Soft 22 (1996) 469–483.

2.5 FOLLOWUP WORK: EFFECT OF QUATERNARY ADDITIONS TO AL-CO-W

Two of the six promising alloys predicted in the high-throughput work, namely Co-Nb-V and Co-Ta-V, were made experimentally by David Dunand's group at Northwestern [14]. They reported the existence of $L1_2$ precipitates (γ') in these two alloy systems. However, it was reported that these precipitates are not stable after 2 hrs of aging. The precipitate is found to be coarsened, dissolved, and transformed into lamellar C36-Co_3(Ta, V) and needle-shaped $D0_{19}$-Co_3(Nb, V) phases.

The experimental results reported by Prof. Dunand's group brings the question of how to stabilize the γ' precipitate in Co-Nb-V and Co-Ta-V alloys. From a computational perspective, a possible solution is to search for appropriate quaternary elemental additions to these alloys which can stabilize the precipitates. However, analyzing the stability of quaternary alloys by computing the quaternary convex hulls for each elemental addition is computationally expensive. A computationally cheaper, but less rigorous approach to this problem is to search for elemental additions which lower the formation enthalpy of the most stable γ' precipitate in the ternary alloy (e.g., Co-Nb-V or Co-Ta-V). The searching is done by first finding the most stable structure in the ternary phase and adding the quaternary element to that structure at varying concentrations and verifying if the new additional element decreases the formation enthalpy. As a test case for the method, we performed this faster approach on Co-Al-W alloy. In this regard, I mentored Hayden Oliver, an undergraduate student in our group through this project[2] to understand the effect of quaternary elements on Al-Co-W alloy.

As a first step, we needed to explore the ternary Co-Al-W system in detail and compute a *robust* convex-hull (with $\sim$ 200 000 structures—all possible ternaries up to 12-atom unit cells) to identify the most stable ternary crystal structure. In order to compute the convex hull, we employed a machine learning model based on Moment Tensor Potentials (MTP) [6]. The robust convex hull with formation enthalpy of over 200 000 crystal structures using MTP revealed 8-, 12-, and a 16-atom $L1_2$-like structures, closer to the convex hull. The 8-atom,12-atom and 16-atom $L1_2$-like structures which Hayden found are an interesting result as they are lower in the formation enthalpy in comparison to the random 32-atom SQS structure we used in the high-throughput work and are likely candidate structures of the experimentally observed alloy. A detailed description of these new crystal structures near the convex hull of Al-Co-W system will be discussed elsewhere[2].

Using these 12- and 16-atom unit cells as parent structures, we studied the effect of the addition of a quaternary element. We considered 10 possible elements (C, Cr, Fe, Mo, Nb, Ni, Si, Ta, Ti, and V) and studied the effect on formation enthalpy of parent structures at $\sim$6%, $\sim$8%, and $\sim$12% atomic concentrations[3]. Fig. 2.13 shows the effect of all 10 elements on the Al-Co-W system. We found that C, Si, Ta, Ti, and V help lower the formation enthalpy of parent structures. At $\sim$6%, only Si and C lower the enthalpy, whereas at $\sim$8% and $\sim$12% we see Ta and V lowering the enthalpy. Also, Si, C, and Ti lower the formation enthalpy further with increasing concentration.

[2] Hayden Oliver, Chandramouli Nyshadham, Carlos Leon, Brayden Bekker, and Gus. L. W. Hart, Investigating Co-Al-W using moment tensor potentials (paper in preparation)

[3] $\sim$6% and $\sim$8% is achieved by adding one quaternary element to a 16- and 12-atom unit cell parent structures. $\sim$12% is achieved by adding two quaternary elements to a 16-atom unit cell parent structure.

In principle, computing a quaternary convex hull is necessary to know if the quaternary precipitates are stable. However, the results from figure 2.13, are a good indication that C, Si, Ta, Ti, and V are possible quaternary additions that could stabilize the ternary precipitate phase. Such studies should be further undertaken on Co-Nb-V and Co-Ta-V systems.

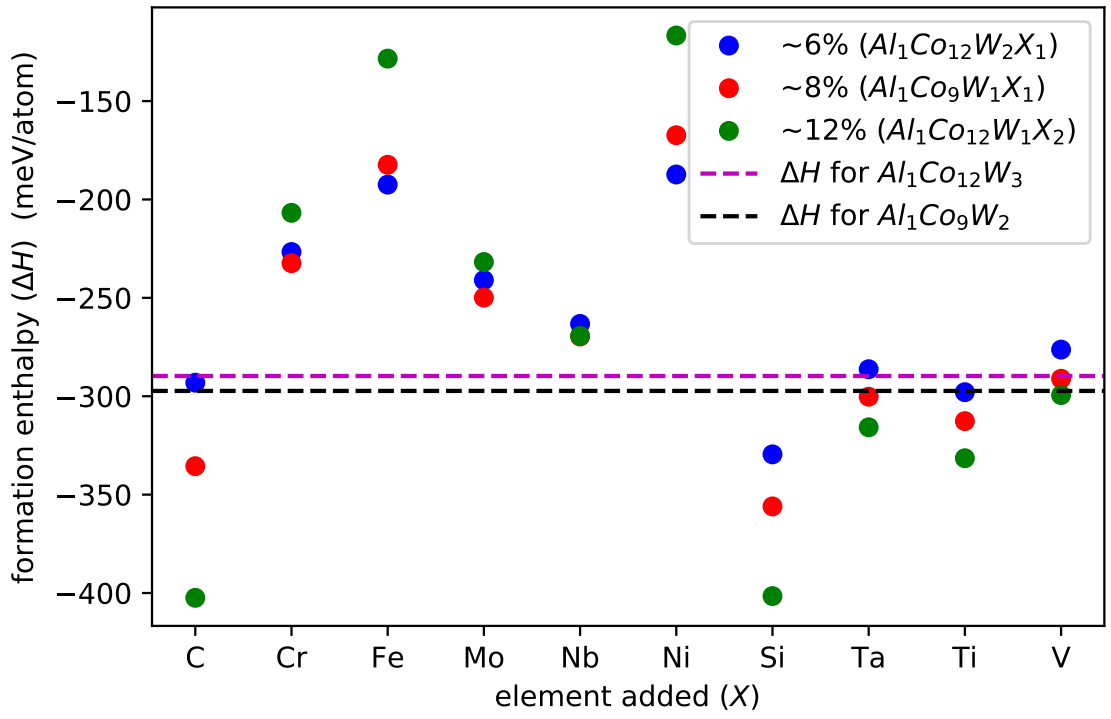


Figure 2.13 Change in the formation enthalpy by adding a quaternary element to Al-Co-W alloy. The parent structures are 12 and 16-atom unit cell and are close to (∼ 2 meV above) convex hull. We can see C, Si, Ta, Ti, and V lowering the formation enthalpy. At ∼6%, Si and C lower the enthalpy of Al-Co-W. At ∼8%, and ∼12% we see Ta and V lowering the enthalpy too. Also, Si, C, and Ti lower the formation enthalpy further with increasing concentration.

CHAPTER 3

Machine-learned models for materials prediction

3.1 MACHINE-LEARNING: NEW PARADIGM FOR MATERIALS PREDICTION

The limitation of DFT-based high-throughput methods is that the methodology is only as fast as DFT itself. Computational methods faster than DFT can allow screening of more possible materials within a short time. In this regard, material scientists resorted to replacing DFT with machine learning (ML) based models which scale linearly with system size without compromising the accuracy. ML is the new paradigm in computational materials science and can help us leverage the surge of materials databases in the last decade. The advantage of ML methods over traditional empirical potentials is that they are accurate, non-parametric, and are systematically improvable with data.

3.2 ACCELERATED HIGH-THROUGHPUT USING ML

The *idea* of an accelerated high-throughput approach (figure 1 in the journal article) is to use ML-based surrogate models to replace electronic structure calculations and screen the best candidate materials from a pool of possible hypothetical materials at a faster rate than using DFT data with high-throughput approach. The candidate materials screened using ML models are validated using first-principle calculations. Many ML-based surrogate models to DFT calculations have been proposed in the literature in the past decade [5–7, 22–28]. Many of the proposed ML models aim to relate structure with the property of materials. The significant differences in all these various ML models lie either in the descriptors they chose to represent the crystal structure or the machine learning algorithm used.

In the second project of this Ph.D. work, we aim to answer the following questions on state-of-the-art ML models in the literature.

- How applicable are current state-of-the-art models to predict properties of solids?
- Are different ML models consistent in their prediction?
- What are the limits of these ML models in predicting properties?
- Can we model multiple materials simultaneously using these models?

3.3 PROJECT

In order to answer these questions, we picked five ML models in the materials science literature. These five ML models consist of three state-of-the-art representations namely Many-body tensor representation [7] (MBTR), smooth overlap of atomic positions [5, 26] (SOAP) and moment tensor potentials [6] (MTP). We also used cluster expansion [38–41](CE), and a machine-learned representation using MBTR along with deep neural networks [31, 42] (DNN). Our goal was not to compare the performance of these different surrogate models, so we did not aim to minimize the error; instead, aimed to maintain an average speed/accuracy balance [28].

The dataset consists of 15950 unrelaxed DFT calculations, containing 10 different binary alloys comprising of 10 different

elements. We chose formation enthalpy as the property of interest to predict. Formation enthalpy is one of the essential theoretical properties in materials science. It indicates how likely a compound is to occur in nature. We compare the formation enthalpy of all compounds through a convex hull plot to predict if a phase will be stable in an alloy. For many metal alloys, the discrepancy in the formation enthalpy between any two competing structures may be as small as $\sim$3 meV/atom (about the limit of DFT accuracy making this problem a challenging one). Our goal is to know how accurately the five ML models can predict the formation enthalpy of binary alloys.

Our results show that all five ML models agree qualitatively on their prediction of formation enthalpy for all 10 materials, which implies that the prediction of ML models are independent of the details of the particular model used. We also show that the ML models are capable of simultaneously predicting the formation enthalpy for multiple materials. In other words, instead of building a separate model for each binary alloy, we can make one model of a 10 component alloy that can predict for each of the 10 binary alloys. We also found that for six of the ten materials, the prediction errors are higher than the required accuracy of $\sim$3 meV/atom. We found that materials with higher prediction errors correspond to a higher range of formation enthalpy in data. These errors indicate the limitation of these models, and further analysis is necessary to improve the accuracy of ML models.

The details of various ML models, materials used in this work, and results are in the journal article [28] attached in the following pages.

3.3.1 My contribution

My contribution to the publication of this project was as follows: conceiving the idea, generating the DFT dataset, performing the MBTR based calculations (MBTR+DNN, MBTR+KRR), interpreting results, Figures 1 and 2 in the paper, Figures 2 and 4 in supplementary material, writing a significant portion of the paper, and responding to the reviewers. Before starting the project, as a proof of concept, I generated a computationally cheaper dataset using LAMMPS package [43, 44] (a molecular dynamics and materials modeling program) and the ASE package [45]. Although in the published article the results for ML models other than MBTR+DNN is generated by other contributing authors, I also worked and have experience with all five ML models personally (for cluster expansion I used the UNCLE package [46]). Contribution of other authors is mentioned in the “AUTHOR CONTRIBUTIONS" section in the paper.

npj Computational Materials www.nature.com/npjcompumats

ARTICLE OPEN

Machine-learned multi-system surrogate models for materials prediction

Chandramouli Nyshadham[1], Matthias Rupp[2,7], Brayden Bekker[1], Alexander V. Shapeev[3], Tim Mueller[4], Conrad W. Rosenbrock[1], Gábor Csányi[5], David W. Wingate[6] and Gus L. W. Hart[1]

Surrogate machine-learning models are transforming computational materials science by predicting properties of materials with the accuracy of ab initio methods at a fraction of the computational cost. We demonstrate surrogate models that simultaneously interpolate energies of different materials on a dataset of 10 binary alloys (AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, and NbNi) with 10 different species and all possible fcc, bcc, and hcp structures up to eight atoms in the unit cell, 15,950 structures in total. We find that the deviation of prediction errors when increasing the number of simultaneously modeled alloys is <1 meV/atom. Several state-of-the-art materials representations and learning algorithms were found to qualitatively agree on the prediction errors of formation enthalpy with relative errors of <2.5% for all systems.

npj Computational Materials (2019)5:51 ; https://doi.org/10.1038/s41524-019-0189-9

INTRODUCTION

Advances in computational power and electronic structure methods have enabled large materials databases.[1–4] Using high-throughput approaches,[5] these databases have proven a useful tool to predict the properties of materials. However, given the combinatorial nature of materials space,[6,7] it is infeasible to compute properties for more than a tiny fraction of all possible materials using electronic structure methods such as density functional theory (DFT).[8,9] A potential answer to this challenge lies in a new paradigm: surrogate machine-learning models for accurate materials predictions.[10–12]

The key idea is to use machine learning to rapidly and accurately interpolate between reference simulations, effectively mapping the problem of numerically solving for the electronic structure of a material onto a statistical regression problem.[13] Such fast surrogate models could be used to filter the most suitable materials from a large pool of possible materials and then validate the found subset by electronic structure calculations. Such an "accelerated high-throughput" (AHT) approach (Fig. 1) could potentially increase the number of investigated materials by several orders of magnitude.

Traditionally, empirical interatomic potentials were used to reproduce macroscopic properties of materials faster than DFT. Well-known empirical interatomic potentials for periodic solids include Lennard–Jones potentials, the Stillinger–Weber potential and embedded-atom methods (EAM) for alloys. A problem with empirical interatomic potentials is that they are designed with a fixed functional form and cannot be systematically improved. In contrast, surrogate models which are empirical interatomic models based on machine learning systematically improve with additional data. This potential advantage over traditional potentials has resulted in the proposal of many machine-learned surrogate models for materials prediction.

We demonstrate the feasibility of machine-learned surrogate models for predicting enthalpies of formation of materials across composition, lattice types, and atomic configurations. Our findings were motivated toward knowing whether different surrogate models proposed in the literature are consistent in their predictions of formation enthalpy rather than comparing the performance of different surrogate models. We find that five combinations of state-of-the-art representations and regression methods (Table 1) all yield consistent predictions with errors of ~10 meV/atom or less depending on the system. We also find that when we combined the data from all 10 systems to build a single model, the combined model is essentially as good as the 10 individual models.

A surrogate machine-learning model replaces ab initio simulations by mapping a crystal structure to properties such as formation enthalpy, elastic constants, or band gaps, etc. Its utility lies in the fact that once the model is trained, properties of new materials can be predicted very quickly. The prediction time is either constant, or scales linearly with the number of atoms in the system, with a low pre-factor, typically in milliseconds.

The two major parts of a surrogate machine-learning model are the numerical representation of the input data[11,14] and the learning algorithm. We use the term "representation" for a set of features (as opposed to a collection of unrelated or only loosely related descriptors) that satisfies certain physical requirements[12,13,15,16] such as invariance to translation, rotation, permutation of atoms, uniqueness (representation is variant against transformations changing the property, as systems with identical representation but differing in the property would introduce errors[17]), differentiability, and computational efficiency. The role of

[1]Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602, USA; [2]Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany; [3]Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Building 3, Moscow 143026, Russia; [4]Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA; [5]Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK; [6]Computer Science Department, Brigham Young University, Provo, UT 84602, USA; [7]Present address: Citrine Informatics, 702 Marshall Street, Redwood City, CA 94063, USA
Correspondence: Gus L. W. Hart (gus.hart@byu.edu)

Received: 26 September 2018 Accepted: 27 March 2019
Published online: 18 April 2019

the representation is akin to that of a basis set in that the predicted property is expanded in terms of a set of reference structures.

To model materials, it is desirable that a representation enables accurate predictions and is able to handle multiple elements simultaneously. The materials community has proposed several representations[10–12,14,15,18–21] for crystal structures. Some do not fulfill the above properties exactly or are restricted, in practice, to materials with a single element. Consequently, surrogate models based on these representations are limited in their accuracy, due to the violation of any of the physical requirements mentioned above (e.g., for the sorted and eigenspectrum variants of the Coulomb matrix, continuity and uniqueness, respectively[16,17]).

We explore three state-of-the-art representations that fulfill above properties for construction of general surrogate models: many-body tensor representation[12] (MBTR), smooth overlap of atomic positions[10,15] (SOAP), and moment tensor potentials[11] (MTP). Each representation is employed as proposed and implemented by its authors, including the regression method: Kernel ridge regression[13] (KRR) for MBTR, Gaussian process regression[22] (GPR) for SOAP, and polynomial regression[11] for MTP. Since predictions (but not necessarily other properties) of the kernel-based KRR and GPR are identical, we will use the two terms interchangeably here. We also employed cluster expansion[23–26] (CE) and deep neural network[27,28] (DNN) models. Our purpose is not to compare the performance of these different surrogate models. Consequently, the models were not optimized to minimize the error; rather they were generated to maintain a typical speed/accuracy balance.

CE models have been used for three decades to efficiently model ground state energies of metal alloys, but require that the atomic structure can be mapped to site occupancies on a fixed lattice. They are therefore less suited to model different materials. In this work, we use them as a baseline and build a separate CE model for each alloy. The comparison is not between CE and other models regarding performance, but our intention is to see how consistent are these different models in predicting the formation enthalpy of materials.

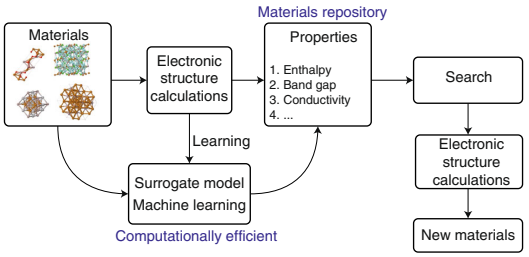


Fig. 1 The accelerated high-throughput approach. Candidate structures and properties are generated by surrogate machine-learning models based on reference electronic structure calculations in a materials repository. Selected structures are validated by electronic structure calculations, preventing false positive errors

DNNs are essentially recursively stacked layers of functions, a large number of layers being a major difference between DNNs and conventional neural networks. They have been used to predict energies[29–33] and to learn representations.[34,35] While DNNs can learn representations ("end-to-end learning", here from nuclear charges, atom positions and unit cell basis vectors to enthalpy of formation), this requires substantially more data than starting with a representation as input.[18–20] We, therefore, provide the DNN with MBTR as input. MBTR is a manually designed representation and works well with the Gaussian kernel. The idea of using MBTR along with DNN is to explore whether a representation-learning technique can improve upon a manually designed representation in conjunction with the standard Gaussian kernel (MBTR + KRR).

RESULTS AND DISCUSSION

Energy predictions for single alloys

Prediction errors for enthalpies of formation of each of the five surrogate models on each binary alloy subset of the data are presented in Fig. 2a. Prediction errors of all surrogate models agree qualitatively on all subsets of the data. We interpret this consistency to be indicative of the validity of the machine-learning approach to surrogate models of formation enthalpy of materials, independently of the parametrization details of the models.

For four binary systems (AgCu, AlMg, CoNi, CuNi) predictions errors are below 3 meV/atom. The prediction errors of all surrogate models on the remaining six systems (AlFe, AlNi, AlTi, CuFe, FeV, NbNi) are consistent, and it is not obvious as to why these systems are harder to learn. When generating the data, the same methodology and parameters were used for all alloys, and similar fitting procedures were employed for each surrogate model.

We point out that whenever the elements that constitute a binary alloy system belong to the same column of the periodic table or are close to each other in the periodic table in terms of atomic number, the surrogate models' predictions are good and vice versa. Indeed, together these numbers explain 80% of the variance in prediction errors (see supplementary material).

A complementary observation is that while absolute errors vary from alloy to alloy, relative errors (δ_{RMSE}), expressed as a percentage of the range of energies of an alloys' subset of the data, remains <2.5% for all systems (Fig. 2b).

Table 1. State-of-the-art surrogate machine-learning models investigated in this work

Abbrv.	Surrogate model	Description
CE	Cluster expansion[23–26] + Bayesian approach[26]	One of the early successful surrogate models developed in the materials community. A material's ground state energy is expanded as an Ising-type model with constant expansion coefficients.
MBTR +KRR	Many-body tensor representation[12] + kernel ridge regression	Materials are expanded in distributions of *k*-body terms stratified by chemical element species, using non-linear regression.
MBTR +DNN	Many-body tensor representation + deep neural network (DNN)[27,28]	MBTR is used as input for DNN to learn a new representation and predict using a parametric deep regression method.
SOAP +GP	Smooth overlap of atomic positions[15] + Gaussian process regression[22]	Atomic environments represented as smoothened Gaussian densities of neighboring atoms expanded in a spherical harmonics basis, using non-parametric regression.
MTP	Moment tensor potentials (MTP)[11] + polynomial regression	Atomic environments expanded in a tailored polynomial basis, computed via contractions of moment tensors.

General models trained on all alloys

We trained four of the five investigated surrogate models simultaneously on all 10 alloy systems and compared the mean absolute error (MAE) of these combined models with the average MAE when trained on each alloy system separately (Table 2; note that RMSE would differ from MAE due to its non-linear nature). The quantitative agreement indicates that the deviation of the prediction errors is <1 meV/atom when trained on multiple systems.

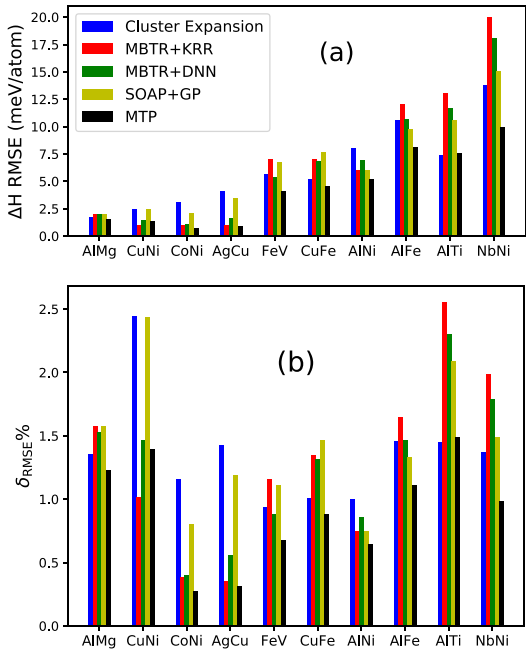


Fig. 2 Consistency in prediction errors of formation enthalpy of five machine-learning surrogate models on the DFT-10B dataset. **a** Root mean squared error (RMSE) of predicted enthalpies of formation of each surrogate model on each binary alloy subset in meV/atom (colored bars). RMSE for MTP results is computed using pure atom total energies obtained from DFT. The consistency of errors across models indicates the validity of machine-learning surrogate models to predict formation enthalpy of materials—prediction errors are similar, independent of the details of model parametrization. **b** Root mean squared error (RMSE) of predicted enthalpies of formation of each surrogate model on each binary alloy subset as a percentage of energy range. Note that relative errors are below 2.5% for all systems

Table 2. Performance of general models

	Mean absolute errors (meV/atom)	
Surrogate model	Average of separate models	Combined model
CE	4.7	4.8
MBTR + KRR	5.1	5.3
MBTR + DNN	5.1	4.6
SOAP + GP	4.5	–
MTP	3.1	3.4

Shown are mean absolute errors (MAE) of models trained on all 10 alloy systems simultaneously (right column) versus the average MAE of models trained on individual alloy systems. The combined fit using SOAP + GP was not performed in this work

For the cluster expansion, these results suggest that there is a single set of parameters for generating a prior probability distribution over effective cluster interaction (ECI) values (provided in the supporting information) that works well across a variety of chemistries and lattice types.

For CE, the representation is naturally tied to a particular lattice (e.g. fcc, bcc), making it difficult to train on multiple alloy systems with different lattices at the same time. Here we train a cluster expansion on all alloys by constraining all 30 systems to use a single set of hyperparameters for regularization (i.e. all use the same prior probability distribution of ECI values). The machine-learning surrogate models based on MBTR, SOAP, and MTP do not suffer from the problem of representation being tied to a particular lattice. They express energy as a continuous function of atomic positions and can be trained on multiple materials simultaneously.

We investigate simultaneous training of alloys in more detail for the MBTR + KRR model. Figure 3 presents deviations of the MAE of a single model trained on k alloy systems from the average MAE when the model is trained on each alloy system separately. In all of the possible $\sum_{k=1}^{10}\binom{10}{k} = 1023$ cases, the deviation is below 1 meV/atom. These deviations are on the order one would expect from minor differences in hyperparameter values. We conclude that prediction errors remain consistently unaffected when increasing the number of simultaneously modeled alloys.

In the case of MBTR + DNN model, we observe improvement in prediction errors on the combined model when compared with the average of separate models (Table 2 [see also Fig. 2 in supplementary material]). This suggests that it might be possible to learn element similarities between chemical element species using a DNN to improve learning rates further.[36]

Caveat emptor

Are reported errors reliable estimates of future performance in applications? It depends. We discuss the role of training and validation set composition as an example of the intricacies of statistical validation of machine-learning models.

In the limit of infinite independent and identically distributed data, one would simply sample a large enough validation set and measure prediction errors, with the law of large numbers ensuring the reliability of the estimates. Here, however, data are limited due to the costs of generating them via ab initio simulations, and are

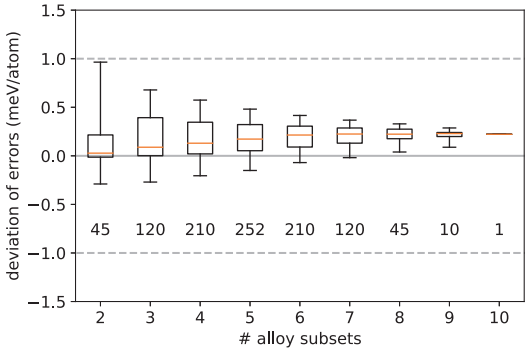


Fig. 3 Performance of MBTR + KRR model for multiple alloy systems. Shown are deviation of mean absolute error (MAE, vertical axis) of an MBTR + KRR surrogate model trained on k (horizontal axis) alloy systems simultaneously from the average MAE of k models trained on each alloy subsystem separately. Whiskers, boxes, horizontal line and numbers inside the plot show the range of values, quartiles, median and sample size, respectively. Difference in error between individual and combined models is always <1 meV/atom

neither independent nor identically distributed. In such a setting, part of the available data is used for validation, either in the form of a hold-out set (as in this work) or via cross-validation, suited for even smaller datasets.

Prediction errors in machine-learning models improve with data (otherwise it would not be machine *learning*). This implies that if only few training samples exist for a "subclass" of structures, prediction errors for similar structures will be high. For example, consider the number of atoms per unit cell in the 10 alloys dataset (DFT-10B) used in this work: There are only 11 structures for each alloy that have 1–2 atoms in the unit cell. Consequently, prediction errors are high for those structures (see Fig. 3 in supplementary material).

In addition to being sparse, smaller unit cells also have a different information content than the larger unit cells. Small unit cells are typically far away from the large unit cells and from each other. Each structure is a point in the representation space and interpolating between structures that are far apart is more prone to error than in regions where the data is tightly clustered (see Fig. 4 in supplementary material). Ideally, the data that the model is trained on would be uniformly distributed in the representation space. Because small unit cells are few in number and because they have a different information content, it is best to include them in the training set.

For combinatorial reasons, the number of possible structures increases strongly with the number of atoms in the unit cell (Table 3). This biases error statistics in two ways: As discussed, prediction errors will be lower for classes with more samples. At the same time, because these classes have more samples, they will contribute more to the measured errors, dominating averages such as the RMSE.

Figure 4 presents MBTR + KRR prediction errors (RMSE in meV/atom) for different but same-size splits of the data into training and validation sets. On the left, all structures with $|k|$ or fewer atoms in the unit cell are excluded from the training set (and therefore included in the validation set). This results in many high-error structures in the validation set, with the effect decreasing for smaller $|k|$. For $k = 0$, size does not influence the split. On the right, structures with $\leq k$ atoms are always included in the training set, resulting in fewer high-error structures in the validation set. The dashed line marks the value of $k = 2$ recommended in this work (see supplementary material).

Table 3. Size distribution in the DFT-10B dataset

Atoms/unit cell	1	2	3	4	5	6	7	8
No. of structures	4	7	12	48	56	210	208	1 050

Shown are the number of structures with k atoms in the unit cell, $k \leq 10$ (per alloy; multiply by 10 for the total dataset)

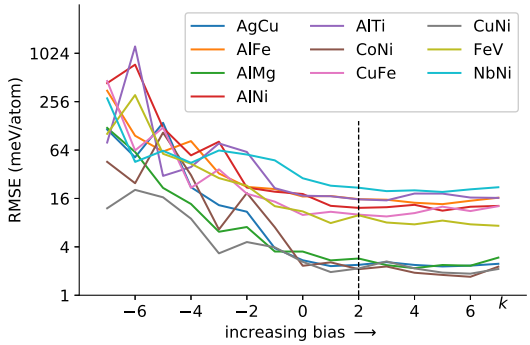


Fig. 4 Influence of biased training and validation sets. Shown are the root mean squared errors (meV/atom) as a function of training and validation set composition obtained using MBTR + KRR model. See main text for discussion

Retrospective errors reported in the literature should, therefore, be critically assessed. The design of such studies should report on "representative" validation sets instead of those tweaked to yield lowest possible errors. For combinatorial datasets, the smallest structures (those that can be considered to be outliers) should be included in the training set.[37]

We showed that it is possible to use machine learning to build a combined surrogate model that can simultaneously predict the enthalpy of formation of crystal structures across 10 different binary alloy systems, for three lattice types (fcc, bcc, hcp) and for structures not in their ground state. In this, we find that the concept of using machine learning to predict formation enthalpy of materials to be independent of the details of the used surrogate models as predictions of several state-of-the-art materials representations and learning algorithms were found to be in qualitative agreement. This observation also seems to be congruent with recent efforts toward a unifying mathematical framework for some of the used representations.[38]

The ability to use a single surrogate model for multiple systems simultaneously has the potential to simplify the use of surrogate models for exploration of materials spaces by avoiding the need to identify "homogeneous" subspaces and then building separate models for each of them. This also avoids problems such as discontinuities at the boundaries of separate models.

Is it possible to do better? Recent results suggest that it might be possible to exploit similarities between chemical element species to improve learning rates further.[36] This requires either to explicitly account for element similarities in the representations or to learn element similarities from the data, for example with a DNN. While such alchemical learning is outside of the scope of this work, we do observe an improvement in prediction errors for the general MBTR + DNN model (Table 2 [see also Fig. 2 in supplementary material]).

METHODS

Data

We created a dataset (DFT-10B) containing structures of the 10 binary alloys AgCu, AlFe, AlMg, AlNi, AlTi, CoNi, CuFe, CuNi, FeV, and NbNi. Each alloy system includes all possible unit cells with 1–8 atoms for face-centered cubic (fcc) and body-centered cubic (bcc) crystal types, and all possible unit cells with 2–8 atoms for the hexagonal close-packed (hcp) crystal type. This results in 631 fcc, 631 bcc, and 333 hcp structures, yielding $1595 \times 10 = 15{,}950$ unrelaxed structures in total. We refer to this dataset as DFT-10B in this work. The cell shape, volume, and atomic positions were not optimized and the calculations are all unrelaxed, for the sake of efficiency. The crystal structures were generated using the enumeration algorithm by Hart and Forcade.[39]

Lattice parameters for each crystal structure were set according to Vegard's law.[40,41] Total energies were computed using DFT with projector-augmented wave (PAW) potentials[42–44] within the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof[45] (PBE) as implemented in the Vienna Ab Initio Simulation Package[46,47] (VASP). The k-point meshes for sampling the Brillouin zone were constructed using generalized regular grids.[48,49] The details of the k-point density for all 10 alloys is mentioned in Table 1 of the supplementary material.

Models

All single-alloy surrogate models were trained using the same set of 1000 randomly selected crystal structures, including optimization of hyperparameters, and the prediction errors are reported on a hold-out test set of 595 different structures, never seen during training. The same set of

decorations are used as training and test sets for all binaries. Models trained on multiple alloys use the union of the individual alloy's splits. Parametrization details of all surrogate models used in this work can be found in the supplementary material.

DATA AVAILABILITY

The dataset (DFT-10B) generated and used for the current work is publicly available as BA10-18 (DFT-10B) at https://qmml.org/datasets.html.

ACKNOWLEDGEMENTS

C.N. is thankful to Kennedy Lincoln and Wiley Morgan for insightful discussions. C.N., B.B., C.R., and G.L.W.H. acknowledge the funding from ONR (MURI N00014-13-1-0635). M.R. acknowledges funding from the EU Horizon 2020 program Grant 676580, The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence. A.V.S. was supported by the Russian Science Foundation (Grant No 18-13-00479). T.M. acknowledges funding from the National Science Foundation under award number DMR-1352373 and computational resources provided by the Maryland Advanced Research Computing Center (MARCC).

AUTHOR CONTRIBUTIONS

C.N. conceived the idea, generated the dataset, ran the calculations of the MBTR-based models, interpreted the results, and wrote a significant portion of the paper. M. R. was responsible for dataset analysis, did the MBTR + KRR calculations, and also wrote a significant portion of the paper. B.B. helped generate the dataset and analyzed the MBTR + KRR calculations. A.V.S. performed the MTP calculations. T.M. performed all cluster expansion calculations. C.W.R. performed SOAP + GPR calculations. G.C. provided guidance and expertise in applying SOAP to our dataset. D.W.W provided his expertise for the MBTR + DNN model. G.L.W.H. contributed many ideas and critique to help guide the project and helped write the paper.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-019-0189-9).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Curtarolo, S. et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
2. Saal, J. E. et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *J. Miner. Met. Mater. Soc.* **65**, 1501–1509 (2013).
3. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
4. C. Draxl and M. Scheffler NOMAD: the FAIR concept for big-data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
5. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
6. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
7. Walsh, A. Inorganic materials: the quest for new functionality. *Nat. Chem.* **7**, 274 (2015).
8. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
9. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
10. Bartók, A. P., Payne, M. C. & Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
11. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale. Model. Simul.* **14**, 1153–1173 (2016).
12. Huo, H. and Rupp, M. Unified representation for machine learning of molecules and materials. *arXiv preprint* arXiv:1704.06439v3, 13754–13769 (2017).
13. Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quant. Chem.* **115**, 1058–1073 (2015).
14. Schütt, K. T. et al. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
15. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
16. von Lilienfeld, O. A., Ramakrishnan, R., Rupp, M. & Knoll, A. Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quant. Chem.* **115**, 1084–1093 (2015).
17. Moussa, J. E. Comment on fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **109**, 059801 (2012).
18. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
19. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
20. Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys. Condens. Matter* **26**, 183001 (2014).
21. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quant. Chem.* **115**, 1094–1101 (2015).
22. Rasmussen, C. & Williams, C. *Gaussian Processes for Machine Learning*. (MIT Press, Cambridge, 2006).
23. Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. Stat. Mech. Appl.* **128**, 334–350 (1984).
24. De Fontaine, D. in *Solid State Physics* (eds Ehrenreich, H. & Turnbull, D.) Vol. 47, 33–176 (Elsevier, 1994).
25. van de Walle, C. G. & Ceder, G. Automating first-principles phase diagram calculations. *J. Ph. Equilib.* **23**, 348–359 (2002).
26. Mueller, T. & Ceder, G. Bayesian approach to cluster expansions. *Phys. Rev. B* **80**, 024103 (2009).
27. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
28. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
29. Schütt, K. T. et al. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
30. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, 241715 (2018).
31. Mills, K., Spanner, M. & Tamblyn, I. Deep learning and the Schrödinger equation. *Phys. Rev. A* **96**, 042113 (2017).
32. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
33. Schütt, K. T. et al. Quantum-chemical insights from deep tensor neural networks. *Nat. Comm.* **8**, 13890 (2017).
34. Matlock, M. K., Le Dang, N. & Swamidass, S. J. Learning a local-variable model of aromatic and conjugated systems. *ACS Cent. Sci.* **4**, 52–62 (2018).
35. Gao, X. & Duan, L.-M. Efficient representation of quantum many-body states with deep neural networks. *Nat. Commun.* **8**, 662 (2017).
36. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
37. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
38. Willatt, M. J., Musil, F. & Ceriotti, M. Theory and practice of atom-density representations for machine learning. *arXiv preprint* arXiv:1807.00408 (2018).
39. Hart, G. L. W. & Forcade, R. W. Algorithm for generating derivative structures. *Phys. Rev. B* **77**, 224115 (2008).
40. Vegard, L. Die Konstitution der Mischkristalle und die Raumfüllung der Atome. *Z. Phys.* **5**, 17–26 (1921).
41. Denton, A. R. & Ashcroft, N. W. Vergard's law. *Phys. Rev. A* **43**, 3161 (1991).
42. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
43. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
44. Kresse, G. & Hafner, J. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *J. Phys. Condens. Matter* **6**, 8245–8257 (1994).
45. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
46. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
47. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).

48. Wisesa, P., McGill, K. A. & Mueller, T. Efficient generation of generalized Monkhorst-Pack grids through the use of informatics. *Phys. Rev. B* **93**, 155109 (2016).
49. Morgan, W. S., Jorgensen, J. J., Hess, B. C. & Hart, G. L. W. Efficiency of generalized regular *k*-point grids. *arXiv preprint* arXiv:1804.04741 (2018).

3.4 FOLLOWUP WORK: 45 ALLOYS MBTR MODELS, FIBS, AND CUR

We showed (in the previous paper) that the predictions of formation enthalpy by five different state-of-the-art models to be qualitatively the same and that the predictions are independent of the details of any particular model. The prediction errors on some materials were as low as 2 meV/atom, whereas some materials have greater than 10 meV/atom errors. High errors indicate that some material systems are harder to predict than others using these surrogate models. The interesting finding is that systems which are hard to predict are hard for all surrogate models. The agreement between all five models validate the application of machine learning techniques for predicting formation enthalpy of materials and that the prediction is independent of the details of the particular model used.

It is important to understand as to why these five ML models could not learn or predict the properties of some materials with high accuracy. In this follow up work, using MBTR+KRR (Many-Body Tensor Representation + Kernel Ridge Regression) model [7, 28], we tried to gain some insights into answering the following questions:

- Is there any correlation between constituent elements and prediction errors?
- How to pick additional training data to improve the prediction errors?

3.4.1 Many-body tensor representation with kernel ridge regression (MBTR+KRR) on 45 binary alloys

In this regard, as a preliminary step to answer the first question, we created a dataset (`DFT-45B`) of 45 binary alloys (all possible binary combinations of the 10 elements used in the previous work) constituting of 71775 unrelaxed DFT calculations. The dataset contains 10 different elements (Ag, Al, Co, Cu, Fe, Mg, Nb, Ni, Ti, and V) and all possible binary combinations ($\binom{10}{2} = 45$ in number) namely, AgAl, AgCo, AgCu, AgFe, AgMg, AgNb, AgNi, AgTi, AgV, AlCo, AlCu, AlFe, AlMg, AlNb, AlNi, AlTi, AlV, CoCu, CoFe, CoMg, CoNb, CoNi, CoTi, CoV, CuFe, CuMg, CuNb, CuNi,CuTi, CuV, FeMg, FeNb, FeNi, FeTi, FeV, MgNb, MgNi, MgTi, MgV, NbNi, NbTi, NbV, NiTi, NiV, and TiV.

We modeled all alloys using the MBTR+KRR model and found that several systems have high prediction errors (above $\sim$5 meV/atom) as shown in Fig. 3.5. The results indicate the importance of understanding and reducing prediction errors. From Fig. 3.5, we notice that low difference in column number (≤ 2) between constituent elements in the binary alloy (NbTi, CuNi, TiV, AgCu, CoNi, NbV AgNi, FeNi) are low in prediction errors. If the difference in the column number is significant (> 3) (CuNb, NbNi, AgNb, CuTi, AlNb, CuV, AgTi, MgNb), prediction errors are high (with some exceptions, e.g., CoFe). In systems with high prediction errors, the range of formation enthalpy is also observed to be higher than that of a system with low errors.

3.4.2 On improving the accuracy of a model

It is crucial to understand how to improve the prediction errors for systems with high errors. A general and simple solution is to add more training data. However, this begs the question of what is the least amount of training data we need to achieve a target accuracy. This smart selection of data directs us to an active learning [47] approach for building efficient models.

The active learning here refers to an intelligent selection of the least amount of new training data based on existing training data. From an information theory perspective, high prediction errors are a result of lack of information in the training set. Using active learning, we aim to add information not present in the training set, which can thus help us to improve ML models by eliminating training on redundant data.

Active learning is not a new concept and has been used in statistics to design *optimal experiments* [48]. There are many algorithms to implement active learning framework [47], but many of these algorithms are computationally expensive. In this follow-up work (using the kernel-based model, MBTR+KRR applied on NbNi system), we focused on implementing an active learning approach [49, 50] to analyze the best way to select new training samples.

ML models are efficient when test data is predicted using interpolation only. When predictions require extrapolation, the model often fails (has high prediction errors). An active learning framework aims to add new training samples that reduce the extrapolation or in other words, reduce the limitations of the model. If the ML model fails to predict on a test sample, that implies extrapolation. It is possible to estimate or quantify this extrapolation from the training data.

3.4.3 Active learning based ML models

From a mathematical perspective, in kernel-based ML models, data points in the input space are first mapped on to a higher-dimensional kernel space wherein a linear algorithm is applied to map to the target space [51] (see Fig. 3.6). In order to learn in an *active* manner or iterative supervised learning, we need to add new training data such that the total information content of the kernel matrix increases with the new addition of data.

Within the scope of kernel-based models such as MBTR+KRR, we analyzed the effect of three active learning algorithms on NbNi system. The first two algorithms are based on a matrix decomposition technique called CUR [49, 50], and the third algorithm called *frequency importance based sampling* (FIBS) was developed by me.

CUR based active learning algorithms: LTCUR and LSCUR

CUR [49,50] decomposition is a matrix decomposition technique, wherein the input matrix, $A \in \Re^{n\times d}$ is decomposed as a set of three matrices $C \in \Re^{n\times c}$, $U \in \Re^{c\times d}$, $R \in \Re^{d\times n}$, which when multiplied together approximate matrix A. The advantage of CUR over other decomposition methods such as SVD is that the rows of the matrix R and columns of the matrix C are expressed in terms of a small number of actual columns and actual rows of the data matrix and are thus more *interpretable* (see Fig. 3.7). The algorithms used to choose columns or rows can give us insight into what data points are more valuable than others from a given dataset.

We used two CUR algorithms, namely Linear-Time-CUR (LTCUR) [50] and Leverage-Score-CUR (LSCUR) [49] algorithms. The description of both the algorithms is mentioned elsewhere [49, 50]. LTCUR scales linearly and LSCUR scales on the order of $\mathcal{O}(n^3)$, where n is the number of points in the dataset.

Frequency importance based sampling

Apart from CUR based algorithms for choosing training data, we also developed and implemented another algorithm called frequency importance based sampling (FIBS). The algorithm for FIBS is as follows.

Algorithm 1:FIBS

Input: Matrix, $K \in \Re^{n \times d}$ with n samples and d features.
Output: FIBS score for each sample, $\{s_i\}_{i=1}^{i=n}$
for $i = 1$ to n **do**
$P = PDF(\{k_{i,j}\}_{j=1}^{j=d})$, where P is probability distribution, function
$m_i = k_{i,j} | P(k_{i,j}) = max(P_i)$
end
return $s_i = \frac{q_i}{\sum_{i=1}^{i=n} q_i} \forall i = 1, 2, \ldots, n$, where $q = 1 - PDF(\{m_i\}_{i=1}^{i=n})$

FIBS algorithm, when applied on a kernel matrix, captures the most important similarity of a sample with respect to all other samples. The frequency of the similarity occurrence determines the most important similarity; hence, the name *Frequency Importance Based Sampling* (FIBS). We note (from Algorithm 1) that the algorithmic complexity of FIBS is of the order of $\mathcal{O}(n)$, where n is the system size.

The goal of CUR algorithms (LTCUR, LSCUR) and FIBS are to rank each sample in the dataset according to the information content. We can see the usefulness of these algorithms through learning curves as an example. Given a data set and ML model, the different algorithms (LTCUR, LSCUR, and FIBS) show the best way to split the dataset into training and test sets with increasing training set sizes. If we pick the training data to be redundant, it causes the learning curve to be flat with high prediction errors. If we pick in a non-redundant fashion within a given dataset, the learning curve should monotonically decrease in error.

Figure 3.8 shows learning curves using LTCUR, LSCUR, and FIBS applied to NbNi dataset. We see that the error improves with all three algorithms. These algorithms identify whether a new training data point adds more information to the existing training data. Adding new information at each iterative active learning process improves ML models at a faster rate. Picking the best training data from a pool of many hypothetical materials reduces the DFT costs to train the ML model. We can see from Fig. 3.9 that FIBS is faster and as accurate as LSCUR algorithm and FIBS has the advantage of being parallelizable.

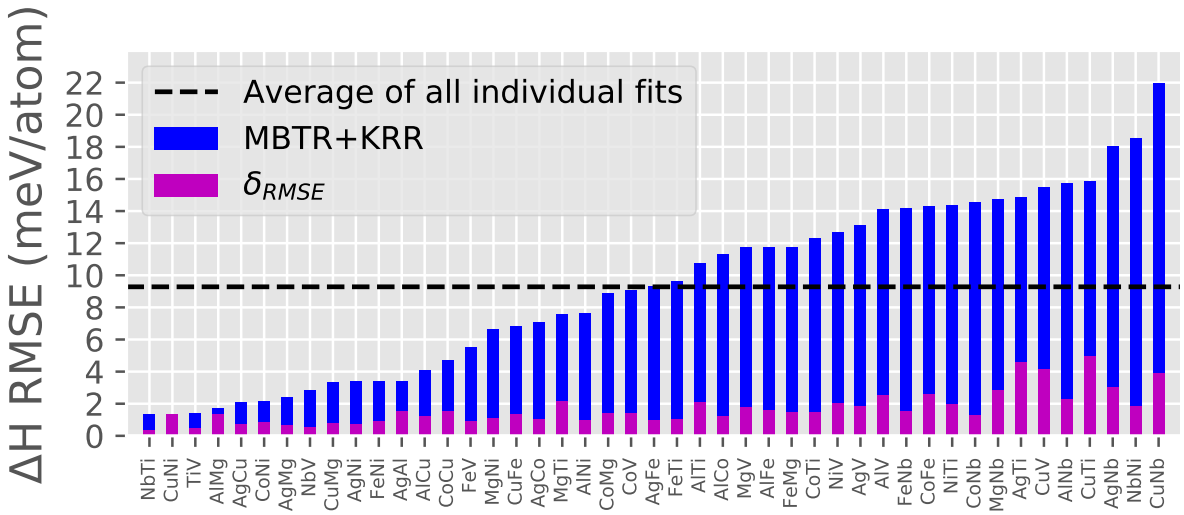


Figure 3.5 Root mean squared error (RMSE) of predicted enthalpies of formation of MBTR + KRR surrogate model on each of 45 binary alloys in meV/atom. The relative errors (δ_{RMSE}), express formation enthalpy as a percentage of the range of energies of individual alloy dataset.

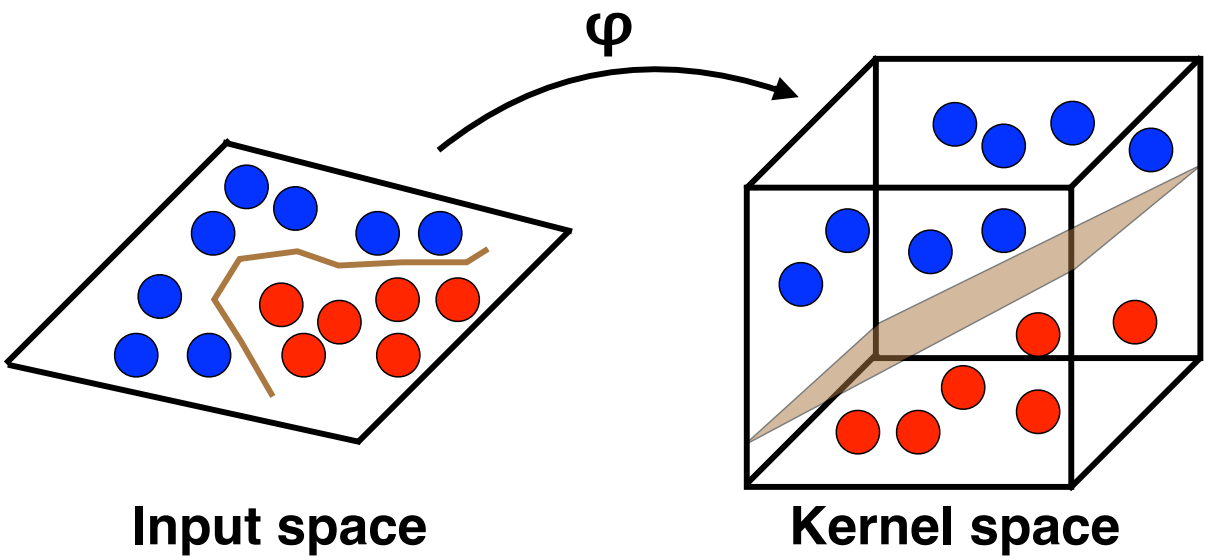


Figure 3.6 Mapping a set of data points (blue and red) from input space to kernel space using the kernel function ϕ. Mapping onto the higher dimensional space allows us to classify the blue and red data points, classified using non-linear function (brown line) in input space with a linear hyper plane (brown colored plane) in the kernel space.

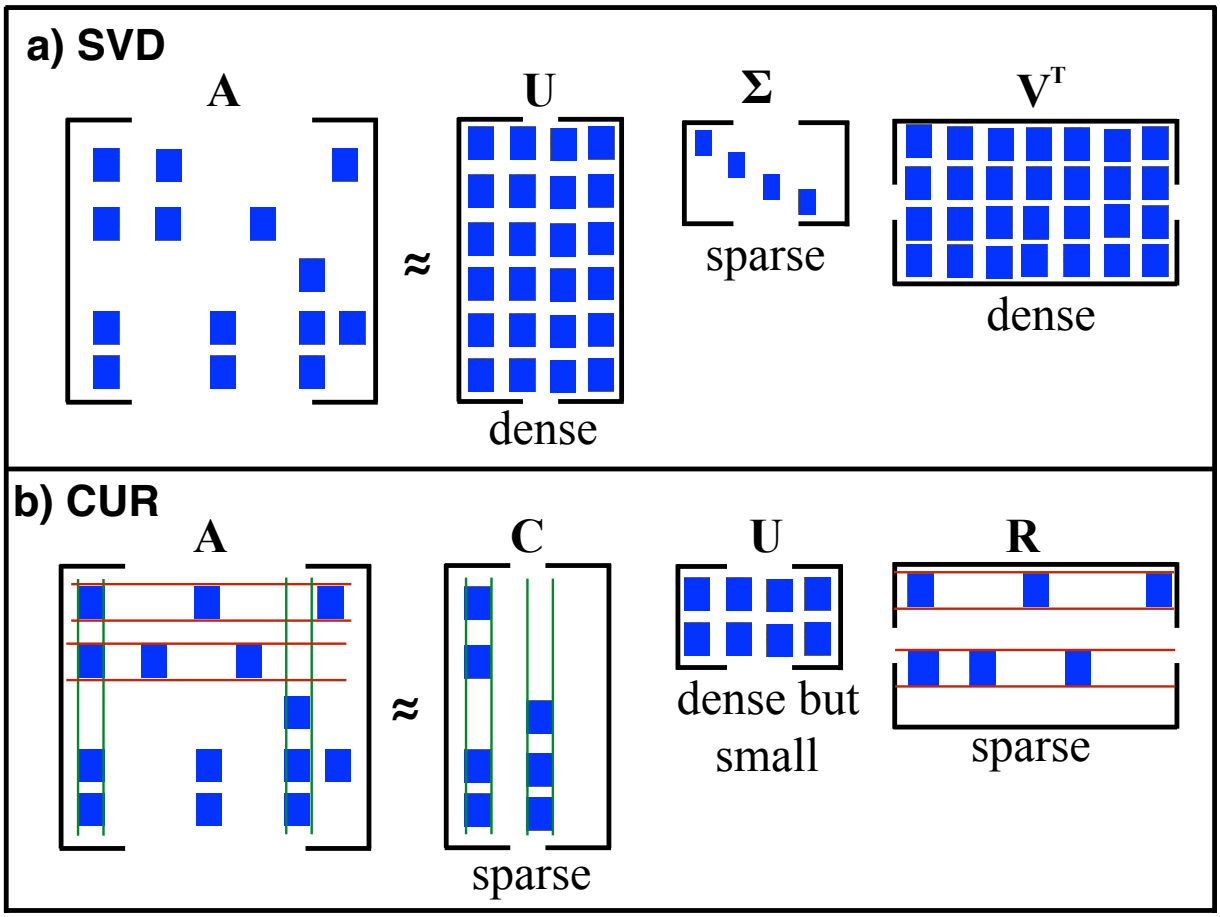


Figure 3.7 Graphical representation of a) singular value decomposition (SVD) and b) CUR decomposition of a matrix. In SVD, the matrix A is decomposed into three matrices namely U, Σ, and V^T. Here U and V^T are dense and big matrices with Σ being small and sparse. In CUR decomposition, the matrix A is decomposed into three matrices namely C, U, and R, where the matrices C and R are sparse and are expressed in terms of a small number of actual columns and actual rows of the data matrix and are thus more interpretable than SVD.

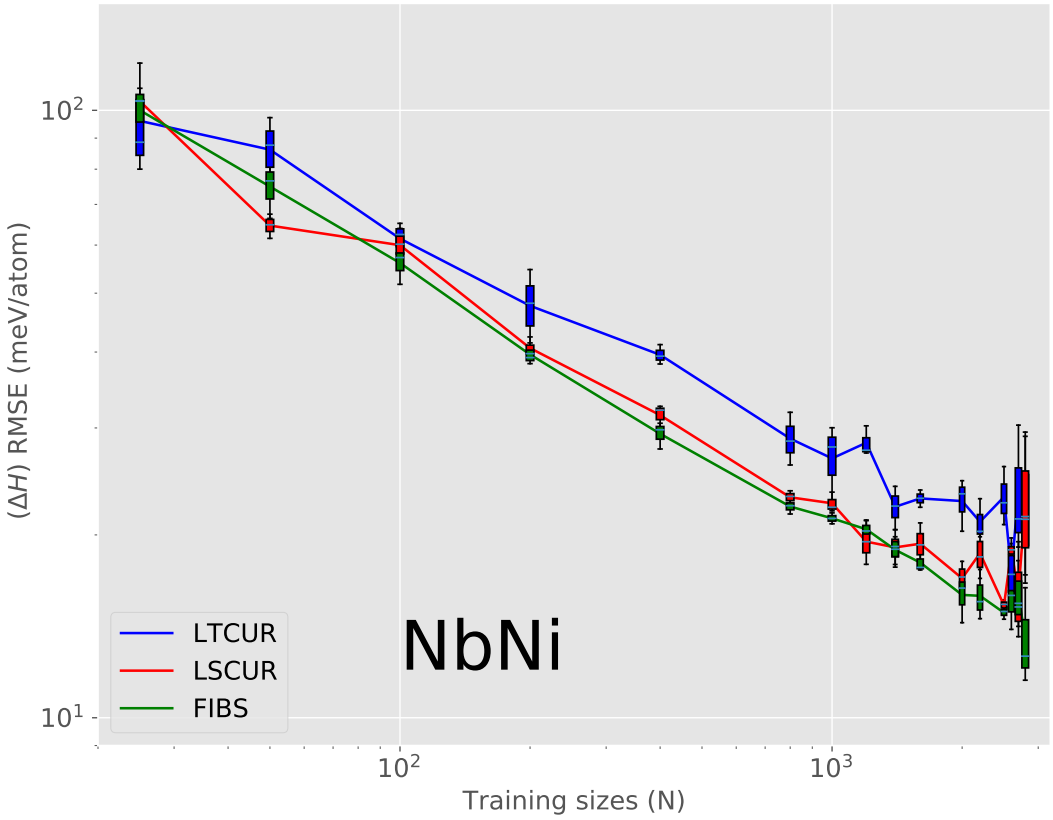


Figure 3.8 Root mean squared error (RMSE) of predicted enthalpies of formation of MBTR + KRR surrogate model on NbNi system in meV/atom using LTCUR, LSCUR and FIBS algorithms.

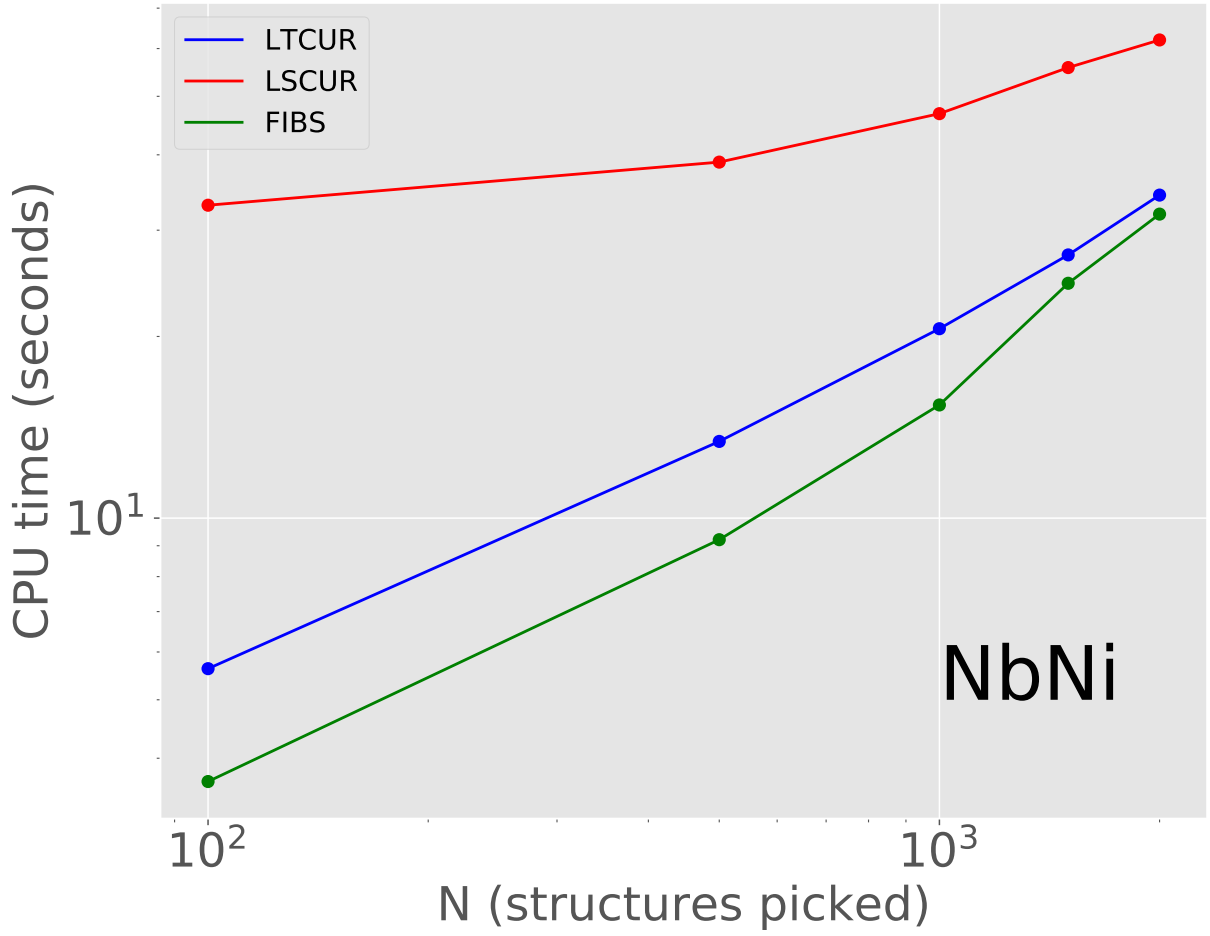


Figure 3.9 CPU time taken for LTCUR, LSCUR and FIBS algorithms for ranking and picking N training structures from a given dataset.

CHAPTER 4

Conclusion and future work

In this thesis work, two projects related to two computational approaches have been studied and used for materials prediction. The first approach is the high-throughput approach, a powerful technique to explore the combinatorial space of materials in a rational fashion for discovering new materials. The use of the high-throughput method in this work successfully led to the prediction of new superalloys. The second project deals with studying the consistency of various surrogate machine learning models to DFT for materials prediction. Five different ML models, including three state-of-the-art models, have been studied and used to predict formation enthalpy of 10 different binary alloy materials. We showed that all five ML models, despite their differences, are consistent and qualitatively agree in their prediction errors. We also showed that ML models are capable of simultaneously predicting the formation enthalpy of multiple materials.

The potential superalloy candidates reported in our high-throughput project led to further experimental studies by Prof. Dunand's group at Northwestern University. Their study of two Co-based superalloy candidates namely Co-Ta-V and Co-Nb-V showed the existence of γ' precipitates in these systems. However, the precipitates are found to be metastable. A further investigation of quaternary elemental additions to these ternary systems can help us to stabilize these metastable materials. In this regard as a proof of methodology, we investigated the effect of 10 different elements as quaternary additions in stabilizing the γ' phase in Al-Co-W system. We found that C, Si, Ta, Ti, and V help lower the formation enthalpy of the precipitate. Similar to the Al-Co-W system, the effect of quaternary additions to Co-Ta-V and Co-Nb-V systems should be carried out in the future. A more rigorous and faster way to understand the stability of superalloys is through the calculation of compositional phase diagrams based on machine-learned interatomic potentials [52].

The conclusions of the second project carried out in this dissertation work show the importance of active learning framework in materials prediction. Active learning framework can help us to select important training data for building efficient ML models. We studied three active learning algorithms, namely, LTCUR, LSCUR, and FIBS, for selecting the training data for building efficient ML model for NbNi system. Further investigation should be carried out to understand the efficiency of ML models in modeling various other properties, including finite temperature properties, and point defect energetics.

The properties of materials we have limit our technology. The main goal of materials scientists is to discover materials with enhanced properties. With the advancement of modern electronic structure calculations such as density functional theory (DFT) [8, 9] and supercomputers, it has become feasible to compute properties of materials with the accuracy of quantum mechanics in a reasonable time. Nevertheless, the possible number of materials to search for is innumerable, which leads to the challenging problem at present: to come up with ever-faster computational methods to scan all possible materials.

What will the future of materials science be? In my opinion, the surge of first-principles materials data in the last few years led to a new paradigm in the field of computational

materials science. Materials prediction has become more data-driven, and there are two main ML-based approaches actively used for discovering new materials. The first approach is materials informatics approach wherein ML techniques are used to extract new knowledge or for building predictive models out of existing materials data [16, 53–55]. However, materials data is not *big data* (too large or complex unstructured data, hard to be dealt with a traditional data-processing software) and one of the main challenges at present is to leverage the existing data using ML-based surrogate models to make materials data, into *big data*. The second approach aims to solve this problem by building quantum-accurate interatomic potentials and enriching the materials database in a computationally cheaper and faster way [6, 7, 26].

Many of these computational methods such as high-throughput and machine learning, initially started in the chemistry or molecular community and were later adapted by material scientists for materials problems. Currently, the chemistry community is progressing towards more data-driven tools, inverse design approach (given a desired property, finding the structure), and automated labs for molecular designs [56, 57]. Although it is quite challenging to implement such methods for solids, we are heading in that direction.

Materials prediction through inverse design is a hard problem as the material with desired properties might not be stable. However, over the next few decades, materials data may become *big*, and it will be possible to tap the full potential of powerful tools such as deep learning for potentially solving the problem of inverse design. Such an inverse-design approach using deep neural networks is currently pursued to designing new molecules and drugs in the chemical and pharmaceutical industries. There is an excellent opportunity in the future to use deep learning tools [58] for inverse design of future materials.

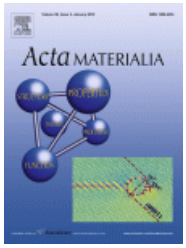
CHAPTER 5

Appendix

5.1 PAPER COPYRIGHT LICENSES

As the author of the Elsevier article (A computational high-throughput search for new ternary superalloys), I have the rights to redistribute the article (for a non-commercial purpose) in this thesis work. A copy from the journal's website regarding the permission for redistribution is shown in the following page.

The journal "npj computational materials" is open access and the article published in npj computational materials (Machine-learned multi-system surrogate models for materials prediction), included in this thesis work do not require any licenses or permissions for redistribution.



Title: A computational high-throughput search for new ternary superalloys

Author: Chandramouli Nyshadham,Corey Oses,Jacob E. Hansen,Ichiro Takeuchi,Stefano Curtarolo,Gus L.W. Hart

Publication: Acta Materialia

Publisher: Elsevier

Date: 1 January 2017

5.2 SUPPLEMENTARY INFORMATION FOR THE JOURNAL ARTICLES

5.2.1 A computational high-throughput search for new ternary superalloys

The supplementary information containing the convex hulls of 2224 ternary systems are available online at http://aflow.org/superalloys.

5.2.2 Machine-learned multi-system surrogate models for materials prediction

The supplementary information accompanying the work is attached in the following pages.

Machine-learned multi-system surrogate models for materials prediction

Chandramouli Nyshadham,[1] Matthias Rupp,[2,3] Brayden Bekker,[1] Alexander V. Shapeev,[4] Tim Mueller,[5] Conrad W. Rosenbrock,[1] Gábor Csányi,[6] David W. Wingate,[7] and Gus L. W. Hart[1]

[1] *Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602, USA*
[2] *Fritz Haber Institute of the Max Planck Society, Faradayweg 4–6, 14195 Berlin, Germany*
[3] *Present address: Citrine Informatics, 702 Marshall Street, Redwood City, CA 94063, USA*
[4] *Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Building 3, Moscow, 143026, Russia*
[5] *Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA*
[6] *Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*
[7] *Computer Science Department, Brigham Young University, Provo, UT 84602, USA*

SUPPLEMENTARY MATERIAL

A. Method details

1. Cluster expansion

To determine the cutoff distances for the cluster expansions and determine the initial parameters for the prior probability distributions, we used a length scale in which the edge of a bcc unit cell is 1 unit of length and assumed the hcp, bcc, and fcc crystal structures all had the same nearest-neighbor distance. The cutoff distances used to determine the set of clusters included in the expansion are as follows:

Number of sites in cluster	Maximum distance between sites
2	8
3	4
4	2
5	1.5
6	1.5

This resulted in a total of 791, 941, and 2870 distinct orbits of clusters in the bcc, fcc, and hcp expansions, respectively, including the empty cluster. These numbers were reduced after fitting by "trimming" the cluster expansions, in which cluster functions with very small ECI were removed from the expansion. To determine which clusters to remove, we used the fact that when the cluster functions are orthonormal, the expected squared error due to truncation, $E(\text{error}^2)$, is given by

$$E(\text{error}^2) = \sum_b V_b^2, \tag{1}$$

where V_b is the ECI for the b-th cluster function, the expectation of the squared error on the left is over all possible lattice decorations, and the sum on the right is over cluster functions excluded from the expansion. Thus removing an orbit of clusters with multiplicity m_b increases the expected squared error by $m_b V_b^2$. To trim clusters from the expansion with little loss of accuracy, we removed all orbits of cluster functions for which $\sqrt{m_b V_b^2} < 10^{-5}$ eV. The trimming procedure changed the final average root-mean-squared prediction errors on the training sets by less than 10^{-5} eV / atom and removed on average more than 70% of the ECIs in the expansions.

k	geometry	weighting	discretization	σ
2	`1/distance`	`identity^2`	$(0, 0.005, 90)$	2^{-17}
2	`1/distance`	`identity^2`	$(0, 0.005, 90)$	$2^{-4.5}$
3	`angle`	`1/dotdotdot`	$(-0.15, \pi/100, 100)$	2^{-14}

The ECIs for the cluster expansions were fit to the training data using the Bayesian approach with a multivariate Gaussian prior distribution[1]. The inverse of the covariance matrix for the prior, Λ, was diagonal, with elements given by

$$\lambda_{\alpha\alpha} = \left\{ \begin{array}{ll} 0 & \text{for } n_\alpha = 0 \\ e^{-\lambda_1} & \text{for } n_\alpha = 1 \\ e^{-\lambda_2} e^{-\lambda_3 r_\alpha} n_\alpha^{\lambda_4} & \text{for } n_\alpha > 1 \end{array} \right\}, \tag{2}$$

where n_α is the number of sites in cluster function α and r_α is the maximum distance between sites in Angstroms. The parameters λ_1, λ_2, λ_3, and λ_4 were initially set to 10, 10, 5, and 5 respectively then optimized by using a conjugate gradient algorithm to minimize the root mean square leave-one-out cross-validation error, an estimate of prediction error[2]. For the combined fit, in which a single set of regularization parameters were used for all 30 cluster expansions, the optimized values of $\lambda_1, \lambda_2, \lambda_3$, and λ_4 were 10.0, 20.8, 4.2, and 15.3 respectively.

2. MBTR+KRR

The Many-Body Tensor Representation (MBTR) numerically represents atomistic systems as distributions of many-body terms, such as atom counts, distances, and angles, stratified (separated) by chemical elements. For details please consult Ref.[3]. Kernel ridge regression[4] with a Gaussian kernel was employed throughout. In this work, we use the following parametrization:

We did not use 1-body terms as enthalpies of formation are the result of a linear operation in atom counts already. Values for the σ hyperparameter above refer to Fig. 3 in

k	geometry	weighting	discretization	σ
1	`atom count`	`1/identity`	$(0.5, 1, 25)$	10^{-4}
2	`1/distance`	`identity^2`	$(0.1, 0.005, 70)$	2^{-17}
3	`angle`	`1/dotdotdot`	$(0.1, 0.05, 140)$	$2^{-8.}$

manuscript, where we used fixed hyperparameter values (Gaussian kernel $\sigma = 2^7$, KRR regularization strength $\lambda = 2^{-20}$). For individual models, hyperparameters were optimized on a base-2 logarithmic grid.

3. MBTR+DNN

The mathematical details of the many-body tensor representation for the crystal structures are mentioned in ref.[3]. Each crystal structure is expanded in terms of distributions (k-body terms) of atom counts, (inverse) distances and angles. The Gaussian kernel with a variance (σ) of 11.3 was used for fitting. Each MBTR vector is 1450 long and was optimized using a grid search. The details of the weighting functions, smearing parameters for each k-body term are as follows,

MBTR+DNN model uses the same parameters as MBTR+KRR model for generating the representation. The only difference between the representations is that the k-body terms in MBTR+DNN model are stratified by all 10 elements instead of just two. This results in a representation vector which is 147100 long. The architecture of the convolution neural network used in this work is listed in the table below.

Layer type	Specifications
Fully connected layer	(Size: 2048)
Fully connected layer	(Size: 1024)
Reshaping data	(Size: 4 x 4 x 64)
Convolution transposed layer	(Kernel: 5 x 5, 64 filters)
Convolution layer	(Kernel: 3 x 3, 64 filters)
Max pooling layer	(Pool size: 2 x 2; stride: 2 x 2)
Convolution layer	(Kernel: 3 x 3, 32 filters)
Reshaping data	(Size: 1 x 1024)
Fully connected layer	(Size: 128)
Fully connected layer	(Size: 64)
Fully connected layer	(Size: 4)
Fully connected layer	(Output; size: 1)

The DNN code is implemented using the Tensorflow framework (software available from `www.tensorflow.org`). The models were trained with a mini-batch size of 50 and the RMSE error is used as the cost function for optimizing the weights of the network.

4. SOAP+GP

GAP fits were generated for each alloy system using a 2-body + SOAP approach. The standard deviation (SD, parameter δ) of the Gaussian process for the 2-body GAP is set to match the SD in energies of the training set. After fitting the 2-body potential, another SOAP GAP is fit with its SD set to match the remaining RMSE of the 2-body GAP relative to the DFT energies in the training set. The fits were performed using `teach_sparse` (software available from `www.libatoms.org`) with the following parameters for the 2-body GAP:

Parameter	Value
Cutoff	6.0 Å
Sparse points	10

and for SOAP, parameters were set to:

Parameter	Value
Cutoff	4.5 Å
Sparse points	500
$l_{\max}$	8
$n_{\max}$	8
ζ	2
σ_{atom}	5

As described above, δ is set using the standard deviation of the Gaussian Process based on the training set for the 2-body and SOAP fits respectively. The following table lists these values for each of the alloy systems.

Parameter	δ (2-body)	δ (SOAP)
AgCu	0.43	0.0126
AlFe	0.84	0.044
AlMg	0.43	0.0126
AlNi	0.396	0.0193
AlTi	0.78	0.03
CoNi	0.27	0.0337
CuFe	0.84	0.0446
CuNi	0.315	0.0207
FeV	0.184	0.0407
NbNi	0.9	0.05

For all alloy fits, the error hyperparameter σ was set to 1 meV for energies. Force and virial were not used in the fits. Because pure energies have a large effect on the predicted formation enthalpies, we increased the error hyperparameter to 10^{-4} for those training configurations that represented pure elements. This ensured accurate reproduction of the pure energies so that enthalpy errors closely match errors in total energy for configurations.

The parameter ϵ_0 was calculated for each isolated atom by including a padding of 10 Å around a single atom and

using the same pseudopotential as the bulk calculations discussed above. These energies were converged with respect to basis set size and used only the γ k-point.

5. MTP

MTP was introduced in Ref.[5] for single-component system and in Refs.[6,7] was extended to multicomponent systems. MTP partitions the predicted energy into contributions of environments of each atom. Around the central atom of an environment, the neighboring atoms form shells. In these shells, atoms are assigned fictitious weights depending on the distance to the central atom, their types, and the type of the central atom. These weights are free parameters fitted from data. An environment is described by moments of inertia of these shells. All possible contractions of one or more moment tensor to a scalar comprise an infinite sequence of basis functions. This sequence is truncated to yield a finite set of basis functions used in a particular MTP model. The contribution of an environment to the energy is, thus, a linear combination of basis function with coefficients which are also found from data. Refer to Ref.[7] for more details.

In this work for binary systems, we used an MTP with about 300 basis functions. The cutoff for atomic environments was 7 Å. The environments were described by five shells, and the dependence of the weight of a neighbor on the distance to the central atom of the environment was described by eight basis functions. Thus, the total number of parameters in a binary MTP is $5{\times}8{\times}2^2{+}300 \approx 450$ (the factor 2^2 follows from the fact that there can be two types of the central atom and two types of each neighboring atoms). For the 10-component MTP, we used six shells and 850 basis functions, totaling about $6 \times 8 \times 30 + 864 \approx 2300$ parameters, where the factor 30 is the number of interacting pairs of atoms.

B. Dataset details

Table I. *k-point density* Shown are the minimum and maximum values of k-point density across all structures for each of the alloys for computing the DFT total energies.

	Number of k-points /Å^3	
System	Maximum	Minimum
AgCu	550	516
AlFe	596	468
AlMg	635	478
AlNi	589	464
AlTi	535	399
CoNi	554	433
CuFe	568	444
CuNi	561	440
FeV	480	401
NbNi	516	472

C. Analysis of dataset and models

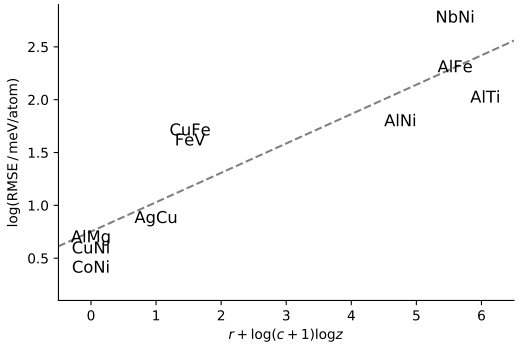


Figure 1. *Alchemical similarity explains prediction errors.*Shown are the logarithmized root mean squared error (RMSE; compare Fig. 2 as a function of an analytic expression in the difference in row r and column c of the periodic table as well as atomic number z of the two chemical element species of a binary alloy. $R^2 = 0.81$.

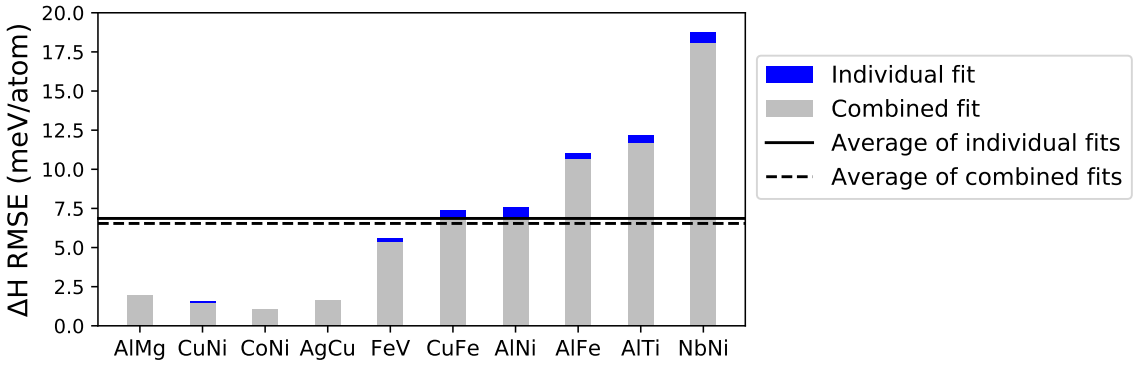


Figure 2. *Improvement of MBTR+DNN model on all alloys.* Shown are the root mean squared error (RMSE) when trained on each alloy separately (blue bars) and on all alloys simultaneously (grey bars).

v

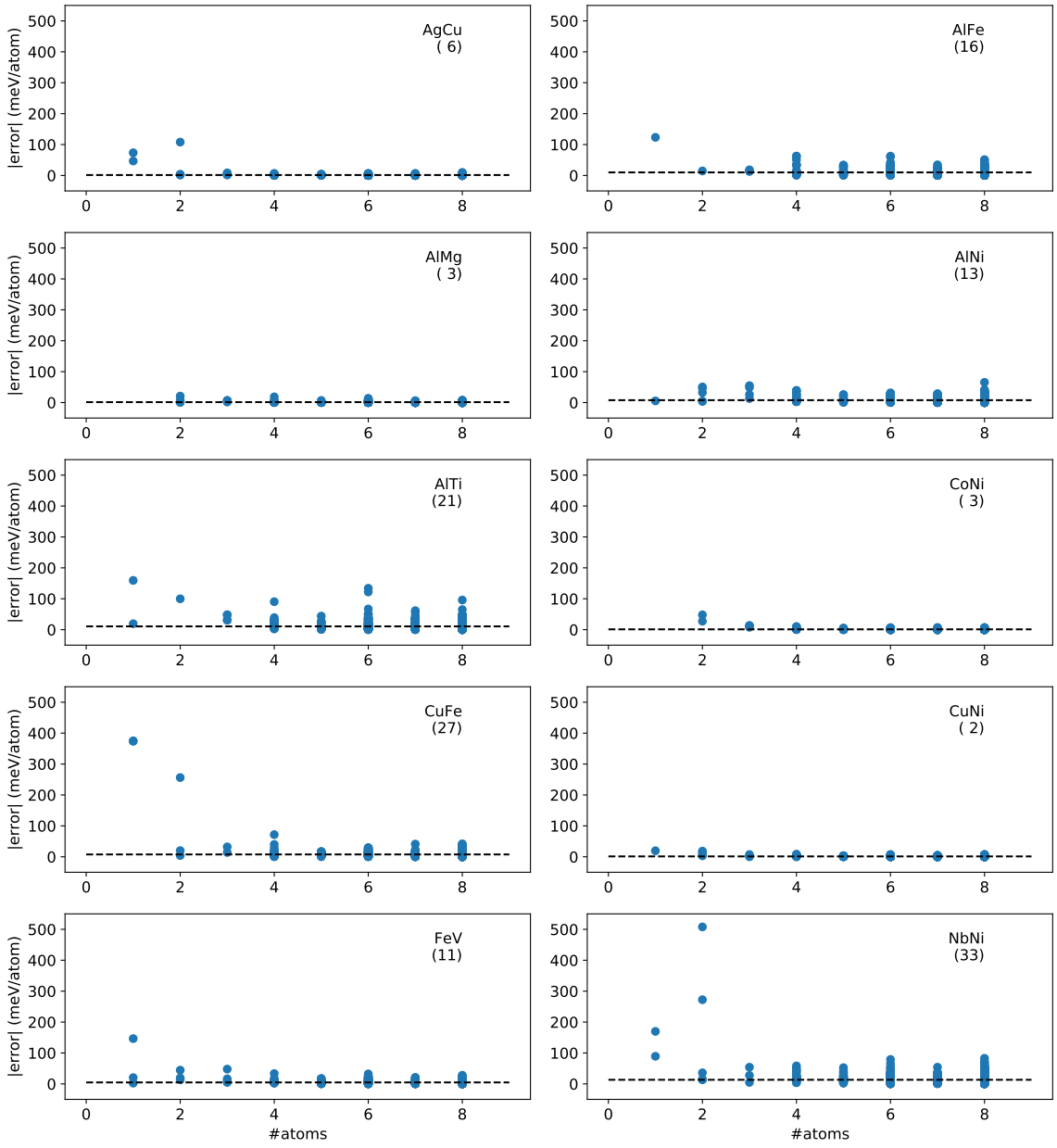


Figure 3. *Influence of unit cell size on errors.* Shown are the absolute errors (meV/atom) as a function of the number of atoms in the unit cell for a validation set of 595 randomly chosen structures using the MBTR+KRR model. The number in brackets and the dashed line indicate the root mean squared error (RMSE, meV/atom) and the median absolute error (meV/atom) on the same set. If small structures (one or two atoms in the unit cell) are not contained in the training data (that is, are shown in the plot) they tend to have larger errors, increasing overall RMSE as well. If all small structures are contained in the training data, the overall RMSE is low (AlMg, CoNi). Retraining models with small structures included in the training set improved RMSE in all cases, by an amount depending on how many structures were added.

Figure 4. Visualizing all 15950 structures (`DFT-10B`) using a t-SNE plot. Each structure in the higher-dimensional space (MBTR) is graphically represented on a 2D plane using t-SNE[8] method. We can observe that 1 or 2 atom unit cells are not representative of larger unit cells in the dataset and are away from other higher atom unit cells. This is a possible reason for high prediction errors when 1 or 2 atom cells are not included in the training set.

REFERENCES

[1] T. Mueller and G. Ceder, Phys. Rev. B **80**, 024103 (2009).
[2] C. G. van de Walle and G. Ceder, J. Ph. Equilib. **23**, 348 (2002).
[3] H. Huo and M. Rupp, arXiv , 1704.06439v3 (2017).
[4] M. Rupp, Int. J. Quant. Chem. **115**, 1058 (2015).
[5] A. V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).
[6] K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, The Journal of Chemical Physics **148**, 241727 (2018).
[7] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, arXiv preprint arXiv:1806.10567 (2018).
[8] L. v. d. Maaten and G. Hinton, Journal of machine learning research **9**, 2579 (2008).

5.3 METHODOLOGY FOR GENERATING DFT-45B DATASET

We created a dataset (DFT-45B) containing structures of the 45 binary alloys AgAl, AgCo, AgCu, AgFe, AgMg, AgNb, AgNi, AgTi, AgV, AlCo, AlCu, AlFe, AlMg, AlNb, AlNi, AlTi, AlV, CoCu, CoFe, CoMg, CoNb, CoNi, CoTi, CoV, CuFe, CuMg, CuNb, CuNi,CuTi, CuV, FeMg, FeNb, FeNi, FeTi, FeV, MgNb, MgNi, MgTi, MgV, NbNi, NbTi, NbV, NiTi, NiV, and TiV.

Each alloy system includes all possible unit cells with 1–8 atoms for face-centered cubic (fcc) and body-centered cubic (bcc) crystal types, and all possible unit cells with 2–8 atoms for the hexagonal close-packed (hcp) crystal type. This results in 631 fcc, 631 bcc and 333 hcp structures, yielding $1595 \cdot 45 = 71\,775$ unrelaxed structures in total. We refer to this dataset as DFT-45B in this work. The cell shape, volume, and atomic positions were not optimized and the calculations are all unrelaxed, for the sake of efficiency. The crystal structures were generated using the enumeration algorithm by Hart and Forcade [59].

Lattice parameters for each crystal structure were set according to Vegard's law. [60, 61] Total energies were computed using density functional theory (DFT) with projector-augmented wave (PAW) potentials [62–64] within the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof [65] (PBE) as implemented in the Vienna Ab Initio Simulation Package [66, 67] (VASP). The k-point meshes for sampling the Brillouin zone were constructed using generalized regular grids. [68, 69]

Bibliography

[1] A. Walsh, "Inorganic materials: The quest for new functionality," Nature chemistry **7,** 274 (2015).

[2] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal fragment descriptors for predicting properties of inorganic crystals," Nature communications **8,** 15679 (2017).

[3] S. Curtarolo *et al.*, "The high-throughput highway to computational materials design," Nat. Mater. **12,** 191–201 (2013).

[4] J. E. Saal *et al.*, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," J. Miner. Met. Mater. Soc. **65,** 1501–1509 (2013).

[5] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons," Phys. Rev. Lett. **104,** 136403 (2010).

[6] A. V. Shapeev, "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials," Multiscale Model. Simul. **14,** 1153–1173 (2016).

[7] H. Huo and M. Rupp, "Unified Representation for Machine Learning of Molecules and Materials," arXiv p. 1704.06439v3 (2017).

[8] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," Phys. Rev. **136,** B864–B871 (1964).

[9] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," Phys. Rev. **140,** A1133–A1138 (1965).

[10] O. Isayev *et al.*, "Universal fragment descriptors for predicting properties of inorganic crystals," Nature communications **8,** 15679 (2017).

[11] C. Nyshadham, C. Oses, J. E. Hansen, I. Takeuchi, S. Curtarolo, and G. L. Hart, "A computational high-throughput search for new ternary superalloys," Acta Materialia **122,** 438–447 (2017).

[12] R. C. Reed, *The superalloys: fundamentals and applications* (Cambridge university press, 2008).

[13] C. Nyshadham, J. E. Hansen, and G. L. Hart, "Superalloys compositions including at least one ternary intermetallic compound and applications thereof,", 2019, uS Patent App. 15/765,952.

[14] F. L. R. Tirado, J. P. Toinin, and D. C. Dunand, "$\gamma+\gamma'$ microstructures in the Co-Ta-V and Co-Nb-V ternary systems," Acta Materialia **151,** 137–148 (2018).

[15] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, "Error estimates for solid-state density-functional theory predictions: an overview by means of the ground-state elemental crystals," Critical Reviews in Solid State and Materials Sciences **39,** 1–24 (2014).

[16] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," npj Computational Materials **3,** 54 (2017).

[17] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach* (Malaysia; Pearson Education Limited,, 2016).

[18] S. Curtarolo *et al.*, "AFLOWLIB.ORG: A distributed materials properties repository from high-throughput *ab initio* calculations," Comput. Mater. Sci. **58,** 227–235 (2012).

[19] A. Jain *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," APL Mater. **1,** 011002 (2013).

[20] C. Draxl and M. Scheffler, "NOMAD: The FAIR Concept for Big-Data-Driven Materials Science," MRS Bull. to appear (2018).

[21] J. Sato, T. Omori, K. Oikawa, I. Ohnuma, R. Kainuma, and K. Ishida, "Cobalt-Base High-Temperature Alloys," Science **312,** 90–91 (2006).

[22] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," Phys. Rev. Lett. **98,** 146401 (2007).

[23] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134,** 074106 (2011).

[24] K. T. Schütt *et al.*, "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties," Phys. Rev. B **89,** 205118 (2014).

[25] J. Behler, "Representing potential energy surfaces by high-dimensional neural network potentials," J. Phys. Condens. Matter **26,** 183001 (2014).

[26] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87,** 184115 (2013).

[27] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal Structure Representations for Machine Learning Models of Formation Energies," Int. J. Quant. Chem. **115,** 1094–1101 (2015).

[28] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. Hart, "Machine-learned multi-system surrogate models for materials prediction," npj Computational Materials **5,** 51 (2019).

[29] B. Larson, "Toughness, NDT education resource center developed by the collaboration for ndt education," Center for Nondestructive Evaluation, Iowa State University, Ames, Iowa 50011 (2001), accessed: 2019-07-15.

[30] A. Höfler, "Deformation process in real crystal structures,", https://www.tec-science.com/material-science/ductility-of-metals/deformation-process-in-real-crystal-structures/, accessed: 2019-07-12.

[31] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks **61,** 85–117 (2015).

[32] J. E. Saal and C. Wolverton, "Thermodynamic stability of Co–Al–W $L1_2$ γ'," Acta Materialia **61,** 2330–2338 (2013).

[33] A. Zunger, S.-H. Wei, L. Ferreira, and J. E. Bernard, "Special quasirandom structures," Physical Review Letters **65,** 353 (1990).

[34] C. Jiang and Y. Du, "Thermodynamic and mechanical stabilities of γ'-Ir 3 (Al, W)," Journal of Applied Physics **109,** 023504 (2011).

[35] S. Curtarolo *et al.*, "AFLOW: an automatic framework for high-throughput materials discovery," Computational Materials Science **58,** 218–226 (2012).

[36] C. E. Calderon *et al.*, "The AFLOW standard for high-throughput materials science calculations," Computational Materials Science **108,** 233–238 (2015).

[37] C. Toher *et al.*, "The AFLOW fleet for materials discovery," Handbook of Materials Modeling: Methods: Theory and Modeling pp. 1–28 (2018).

[38] J. M. Sanchez, F. Ducastelle, and D. Gratias, "Generalized Cluster Description of Multicomponent Systems," Phys. Stat. Mech. Appl. **128,** 334–350 (1984).

[39] D. De Fontaine, "Cluster Approach to Order-Disorder Transformations in Alloys," in *Solid State Physics*, H. Ehrenreich and D. Turnbull, eds., (Elsevier, 1994), Vol. 47, pp. 33–176.

[40] C. G. van de Walle and G. Ceder, "Automating first-principles phase diagram calculations," J. Ph. Equilib. **23,** 348–359 (2002).

[41] T. Mueller and G. Ceder, "Bayesian approach to cluster expansions," Phys. Rev. B **80,** 024103 (2009).

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521,** 436–444 (2015).

[43] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," Journal of computational physics **117,** 1–19 (1995).

[44] "LAMMPS Molecular Dynamics Simulator,", https://lammps.sandia.gov/index.html, accessed: 2019-07-02.

[45] A. H. Larsen *et al.*, "The atomic simulation environment—a Python library for working with atoms," Journal of Physics: Condensed Matter **29,** 273002 (2017).

[46] D. Lerch, O. Wieckhorst, G. L. Hart, R. W. Forcade, and S. Müller, "UNCLE: a code for constructing cluster expansions for arbitrary lattices with minimal user-input," Modelling and Simulation in Materials Science and Engineering **17,** 055003 (2009).

[47] B. Settles, "Active learning literature survey," Technical report, University of Wisconsin-Madison Department of Computer Sciences (2009) .

[48] A. Atkinson, A. Donev, and R. Tobias, *Optimum experimental designs, with SAS* (Oxford University Press, 2007), Vol. 34.

[49] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," Proceedings of the National Academy of Sciences **106,** 697–702 (2009).

[50] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition," SIAM Journal on Computing **36,** 184–206 (2006).

[51] M. Rupp, "Machine learning for quantum mechanics in a nutshell," International Journal of Quantum Chemistry **115,** 1058–1073 (2015).

[52] C. W. Rosenbrock *et al.*, "Machine-learned Interatomic Potentials for Alloys and Alloy Phase Diagrams," arXiv preprint arXiv:1906.07816 (2019).

[53] K. Rajan, "Materials informatics," Materials Today **8,** 38–45 (2005).

[54] L. Ward and C. Wolverton, "Atomistic calculations and materials informatics: A review," Current Opinion in Solid State and Materials Science **21,** 167–176 (2017).

[55] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," Apl Materials **4,** 053208 (2016).

[56] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, and S. K. Saikin, "Autonomous Molecular Design: Then and Now," ACS applied materials & interfaces (2019).

[57] S. K. Saikin, C. Kreisbeck, D. Sheberla, and J. S. Becker, "Closed-loop discovery platform integration is needed for artificial intelligence to make an impact in drug discovery,", 2019.

[58] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).

[59] G. L. W. Hart and R. W. Forcade, "Algorithm for generating derivative structures," Phys. Rev. B **77,** 224115 (2008).

[60] L. Vegard, "Die Konstitution der Mischkristalle und die Raumfüllung der Atome," Z. Physik **5,** 17–26 (1921).

[61] A. R. Denton and N. W. Ashcroft, "Vergard's law," Phys. Rev. A **43,** 3161 (1991).

[62] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59,** 1758 (1999).

[63] P. E. Blöchl, "Projector augmented-wave method," Phys. Rev. B **50,** 17953 (1994).

[64] G. Kresse and J. Hafner, "Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements," J. Phys. Condens. Matter **6,** 8245–8257 (1994).

[65] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," Phys. Rev. Lett. **77,** 3865–3868 (1996).

[66] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6,** 15–50 (1996).

[67] G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54,** 11169 (1996).

[68] P. Wisesa, K. A. McGill, and T. Mueller, "Efficient generation of generalized Monkhorst-Pack grids through the use of informatics," Phys. Rev. B **93,** 155109 (2016).

[69] W. S. Morgan, J. J. Jorgensen, B. C. Hess, and G. L. W. Hart, "Efficiency of Generalized Regular *k*-point Grids," arXiv p. 1804.04741 (2018).