A Study of the Utility of a Machine-Learning Approach Applied to the Prediction of Site Occupancy and New Members of the Half-Heusler Family

by

Alexander Stanislav Gzyl

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Chemistry
University of Alberta

Abstract

Predicting the formation and structures of non-molecular inorganic compounds has long been a fundamental goal in solid state chemistry. In this thesis, machine learning approaches have been applied to confront this challenge, focusing in particular on the large family of half-Heusler compounds because they exhibit many useful materials properties but have not always been structurally well characterized. Two specific problems have been tackled: assigning the correct site distributions in existing half-Heusler compounds, and predicting the formation of new half-Heusler compounds.

The site preference within the structures of half-Heusler compounds have been evaluated through a machine-learning approach. A support-vector machine algorithm was applied to develop a model which was trained on 179 experimentally reported structures and 23 descriptors based solely on the chemical composition. The model gave excellent performance with sensitivity of 93%, specificity of 96% and accuracy of 95%. As an illustration of data sanitization, two compounds (GdPtSb, HoPdBi) flagged by the model to have potentially incorrect site assignments were resynthesized and structurally characterized. The predictions of the correct site assignments from the machine-learning model were confirmed by single-crystal and powder X-ray diffraction analysis. These site assignments also correspond to the lowest total energy configurations as revealed from first-principles calculations.

A machine-learning ensemble was used to predict new half-Heusler compounds. Compositions were selected for synthesis if they were also adopted by a full-Heusler compound counterpart. The model gave excellent performance with sensitivity of 90.0%, Specificity of 98.0%, and accuracy of 97.7%. Perturbations in site occupancy (e.g., vacancies, disorder) led to changes in crystal symmetry. Synthetic minority oversampling (SMOTE) and ensemble methods have been combined and applied for the first time to a materials science problem, and the performance of this approach has been evaluated

Preface

This thesis summarizes the work that was performed in the Mar research group in the Department of Chemistry at the University of Alberta from September 2016 to July 2019. My contributions are summarized below.

Chapter 2 of this thesis has been published as A. S. Gzyl, A. O. Oliynyk, L. A. Adutwum, A. Mar, "Solving the Coloring Problem in Half-Heusler Structures: Machine-Learning Predictions and Experimental Validation," *Inorg. Chem*. **2019**, *58*, 9280–9289. My contribution included synthesis, data collection, characterization, first-principles calculations, and development of the machine-learning model; I wrote an initial draft of the manuscript. A. O. Oliynyk mentored me and helped with analysis, characterization, and manuscript preparation. L. A. Adutwum mentored me and taught me how to use the machine-learning tools he developed. A. Mar edited, composed, and submitted the manuscript.

Chapter 3 is a portion of a manuscript tentatively titled "Using Machine Learning and DFT as Analytical Tools to Understand Heusler Phases" and authored by A. S. Gzyl, A. O. Oliynyk, L. A. Adutwum, and A. Mar. My contribution included synthesis, data collection, characterization, first-principles calculations, and development of the machine-learning model; I wrote an initial draft of the manuscript. A. O. Oliynyk performed first-principles calculations and helped with analysis, characterization, and manuscript preparation. A. Mar is in the process of editing the manuscript.

Acknowledgements

I owe a debt of gratitude to my supervisor Dr. Mar for taking me into his research group and introducing me to solid state chemistry and applied machine learning. Before joining his group, neither was familiar to me and both are now subjects that I am passionate about. I would like to express my appreciation for his supervising style which fosters creativity and scientific exploration.

I would like to express gratitude to the people in the trenches with me doing great research. I cherish the comradery and support from my groupmates Vidyanshu, Dundappa, Yuqiao, Guillaume, Kate, Manon, Harshil, Marissa, Florian, Jan, Ebru, and Abishek. I would like to express my sincere appreciation to Anton, Lawrence, and Mansura for taking the time to mentor me through my studies.

I wish to express gratitude to my committee members, Dr. Veinot and Dr. Hanna, for guiding me through my graduate studies. I would like to thank the ATUMS program and all its members for providing me with many opportunities to further my research, to grow, and to share my work. I would like to thank Dr. Tom Nilges and his group (Claudia, Anna, Markus, Felix, Patrick, and Annabelle) for hosting me in Munich for three months.

I would like to give a special thank you to my family, my best friend Dong, and my partner Maria del Sagrario for their continued support since the beginning of my journey. Without them, I would not be the person that I am today.

This research was funded by the University of Alberta, the Natural Sciences and Engineering Research Council of Canada, NSERC CREATE-IRTG, ATUMS, and FES.

Table of Contents

Chapter 1 Introduction 1
1.1. Heusler compounds 1
1.2. Synthesis of intermetallic compounds 2
1.3. X-ray diffraction 4
1.3.1. Single-crystal diffraction 5
1.3.2. Powder diffraction 7
1.4. Scanning electron microscopy 7
1.5. Band structure calculations 8
1.6. Machine learning 9
1.6.1. Feature selection and principal component analysis 11
1.6.2. Partial least-squares discriminant analysis 15
1.6.3. Synthetic minority oversampling and *k*-nearest neighbours 15
1.6.4. Support vector machine 17
1.7. Research motivation 18
1.8. References 19
Chapter 2 Solving the Colouring Problem in Half-Heusler Structures: Machine-Learning Predictions and Experimental Validation 23
2.1. Introduction 23
2.2. Experimental Section 30
2.2.1. Machine-Learning Model 30
2.2.2. Synthesis of Half-Heusler Compounds 32
2.2.3. First-Principles Calculations 33
2.3. Results and Discussion 34
2.3.1. Machine-Learning Predictions 34
2.3.2. Factors Affecting Site Distributions in Half-Heusler Compounds 42
2.3.3. Total Energy Calculation and Charge Density Analysis for GdPtSb and HoPdBi 44
2.4. Conclusions 46
2.5. References 48

Chapter 3 Predicting New Half-Heusler Compounds 55
3.1. Introduction 55
3.2. Experimental 61
3.2.1. Machine-Learning Model 61
3.3. Results and Discussion 64
3.3.1. Machine learning 64
3.4. Conclusion 72
3.5. References 73
Chapter 4 Conclusions 77
4.1. Machine learning approach to data sanitization 77
4.2. Machine learning approach to materials discovery 78
4.3. Future work 80
Bibliography 81
Appendix 1 Supplementary Data for Chapter 2 94

List of Tables

Table 2-1. Accuracy in predictions of 4*c* site preference in half-Heusler compounds…….25

Table 3-1. Elemental properties used to generate features……65

Table 3-2. Arithmetic expression used to generate features……..67

Table 3-3. Comparison of model performance……..68

Table 3-4. Full-Heusler counterpart probability……70

Table 3-5. Highest probability for half-Heusler structures……71

Table 4-1. Comparison of model performance……..79

Table S-1. Elemental properties……94

Table S-2. Arithmetic operations applied to elemental properties……95

Table S-3. Probabilities for correctness of site distributions in half-Heusler compounds……95

List of Figures

Figure 1-1. Solid state synthesis methods (a) standard resistance furnace (b) arc melter……3

Figure 1-2. Generation of X-rays and the emission spectrum for a Cu X-ray tube…….….5

Figure 1-3. (a) Laue and (b) Bragg conditions for X-ray diffraction……6

Figure 1-4. Band dispersion, density of states, and crystal orbital Hamilton population curves……..9

Figure 1-5. Projection of points onto a line of best fit including Pythagorean relationship…….12

Figure 1-6. Projection of a point in PCA space consisting of two principal components……..…13

Figure 1-7. Non-overlapping confidence ellipses in PCA space……14

Figure 1-8. *k*-nearest neighbours' algorithm…….....16

Figure 1-9. Choice of support vectors gives different decision barrier widths…….....17

Figure 1-10. The projection of data using a gaussian kernel function……..18

Figure 2-1. Merging of (a) NaCl-type and ZnS-type substructures results in (b) half-Heusler structure containing tetrahedral (at 4*a* and 4*b*) and cubic sites (at 4*c*), which can be occupied by a variety of elements as summarized in (c)…….24

Figure 2-2. Simulated powder XRD patterns for LiAuSb with the 4*c* site being occupied by Li, Au, or Sb atoms……...28

Figure 2-3. Probability for correctness of site distributions (a) before and (b) after CR-FS procedure applied……..35

Figure 2-4. VIP scores for descriptors used in the machine-learning model before (blue bars) and after CR-FS (orange bars). For each elemental property, six arithmetic operations were applied…….37

Figure 2-5. Rietveld refinements for MnIrGa with Ga in 4*c*, MnPtSn with Sn in 4*c*, and MnPdSb and Sb in 4*c*. Impurities were included in the peak profiles……..38

Figure 2-6. Probability of element occupying 4*c* site in GdPtSb and HoPdBi……40

Figure 2-7. Rietveld refinements for (a) GdPtSb and (b) HoPdBi with different site distributions (the element highlighted in red in the formula is placed in the 4*c* site)……41

Figure 2-8. Bader charge analysis for (a) GdPtSb and (b) HoPdBi in alternative site distributions, with total energies calculated from first principles and machine-learning probabilities indicated……45

Figure 2-9. DOS and –COHP curves for (a) GdPtSb and (b) HoPdBi in alternative site distributions, with machine-learning probabilities indicated……..46

Figure 3-1. Relationships between CsCl, full-Heusler, and half-Heusler structures……..……..58

Figure 3-2. Compositional overlap between full- and half-Heusler compounds……..59

Figure 3-3. Occurrence of elements in half-Heusler structure type……..60

Figure 3-4. Machine-learning workflow……...62

Figure 3-5. The process of SMOTE: (1) initial unbalanced dataset, (2) generation of synthetic samples using the minority class, and (3) evaluation of synthetic samples in the dataset using KNN……63

Figure 3-6. Ensemble of models used for classification…….…64

Figure S-1. Highlighted samples on prediction probability figure…….104

Chapter 1 Introduction

1.1. Heusler compounds

One of the largest families of intermetallic compounds (combinations of metals having definite composition and ordered structures) are Heusler compounds. Discovered in 1903 by Fritz Heusler, Cu_2MnAl was the first example in this family.[1] It was remarkable at the time, when magnetic phenomena were still not well understood, because it behaves like a ferromagnet even though none of the constituent elements are ferromagnetic. Eventually, this behaviour was explained as arising from magnetic interactions between Mn d states mediated through conduction electrons (known as RKKY interactions).[2] Today, hundreds of Heusler compounds are known which exhibit diverse properties useful for many applications, including memory-shape alloys, superconductors, topological insulators, spintronics, and thermoelectric materials.

The wide range of properties found in Heusler compounds can be traced to the flexibility in their structures and compositions. The crystal structure of Cu_2MnAl, known as the Heusler (or full-Heusler) structure, consists of atoms in three sites (8*c*, 4*b*, 4*a*) in the centrosymmetric space group $Fm\bar{3}m$.[3] The coordination environment around each site can be described as cubic. The family of Heusler compounds includes not only the original Heusler structure, but also other closely related derivatives such as half-Heusler and inverse Heusler structures. The sites in the crystal structures can be occupied by many metallic elements in the periodic table ranging from the alkali metals to the pnicogens, and sometimes even some decidedly non-metallic elements such as O, Br, and Te. Although Heusler compounds are well studied, the focus of this thesis is to try a new approach in understanding and predicting their structures.

1.2. Synthesis of intermetallic compounds

Like many other inorganic solid-state compounds, intermetallic compounds can be prepared by directly combining the elemental components in a sealed container (typically a fused-silica ampoule that is evacuated) and heating them at high temperatures in a furnace. If all reactants remain as solids even at these high temperatures, the mechanism for forming these compounds is assumed to proceed through atomic diffusion.[4] Starting materials can come in various forms such as powders, ingots, or foils. However, reaction rates can be increased by grinding reactants or pelletizing them together increases the contact area between reactants and decreases the path lengths for diffusion.[5] For atoms to move from one position to another, bonds are broken, the structure is distorted, and new bonds are formed. Structural reorganization can be minimized if reactants and products have similar structures. Diffusion rates are affected not only by temperature but other factors such as atomic sizes and packing.[6] Ideally, melting can be achieved but reactions can still proceed below the solidus; a rule of thumb (called Tammann's rule) to ensure reasonable rates is that the reaction temperature should be at least two-thirds of the lowest melting point among the reactants.[7] As a reaction proceeds, a product layer forms and acts as a barrier between reactants. Subsequent regrinding and reheating of the sample is then needed to ensure complete reaction.

Standard resistance furnaces are often unable to reach high enough temperatures for metals to react together, as is the case for many samples synthesized in this work. In this case, arc melting is a useful alternative approach (Figure 1-1.).

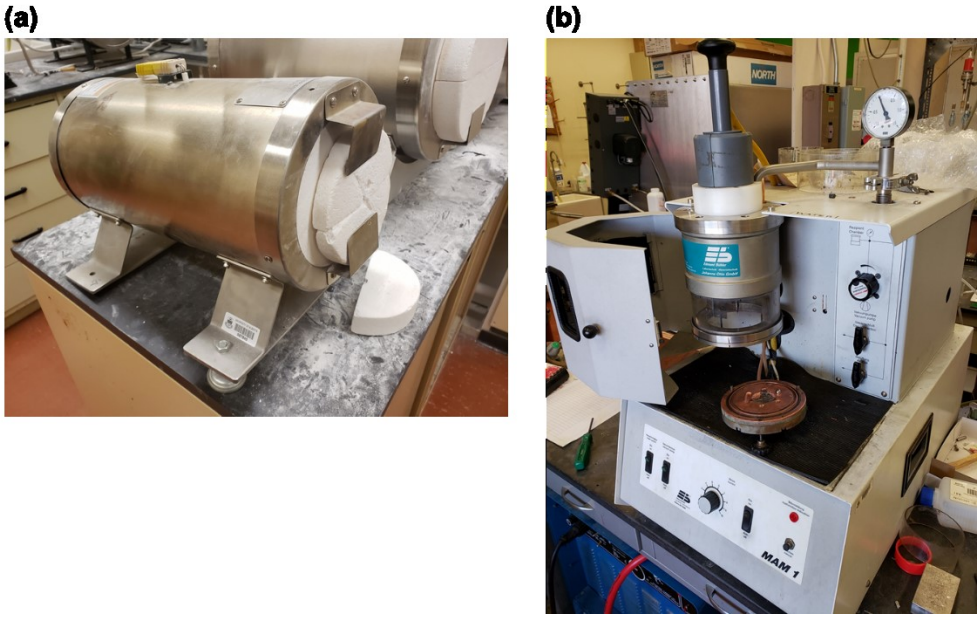


Figure 1-1. Solid state synthesis methods: (a) standard resistance furnace and (b) arc melter.

The starting materials are pressed into pellets and placed within an arc furnace, which is evacuated and back-filled with argon gas. An electric arc is generated by a high potential difference, and is directed to a tungsten getter which is first melted to remove remaining traces of oxygen. Then the pellets are arc-melted, and usually they are flipped over and arc-melted again to ensure homogeneity. Metals with low boiling points (<1500 °C) may volatilize and their loss must be compensated by pre-adding a slight excess, or minimized by using large pieces (foils or ingots). The samples are then further annealed in a furnace and quenched in cold water. For characterization, half of the sample is ground for powder X-ray diffraction and the remaining half is examined for presence of crystals.

1.3. X-ray diffraction

The primary means of characterizing the structures of solid state compounds is through X-ray diffraction. Within the spectrum of electromagnetic radiation, hard X-rays have wavelengths of 0.3 to 2 Å which are comparable to distances between atoms in a crystal.[8] When they strike a crystal, these X-rays are elastically scattered and undergo constructive interference if Bragg's law is satisfied. The directions of the scattered waves depend on the lattice (a set of points such that each has the same environment) and their intensities depend on the basis (the set of atoms associated with each point). X-ray diffraction experiments can be conducted on single crystals or powders (which contain randomly oriented crystallites).

X-rays are produced when electrons are accelerated across an electrical potential and strike a metal anode (typically Cu for powder diffraction or Mo for single-crystal diffraction). An X-ray spectrum consists of a broad background (caused by inelastic scattering processes) superimposed by a few sharp and intense peaks. The background is caused by inelastic scattering processes in which the electron loses kinetic energy in variable amounts. The sharp peaks are caused by elastic scattering processes in which core electrons from the K shell are ejected and the holes are filled by relaxation of higher energy electrons from the L and M shells (Figure 1-2). The less intense K_β line (resulting from the M to K transition) appearing at shorter wavelengths is typically filtered out, while the more intense K_α line (resulting from the L to K transition) appearing at higher wavelengths is selected for the X-ray diffraction experiment. Actually, the K_α line is split into a doublet because of small energy differences arising from spin-orbit coupling.

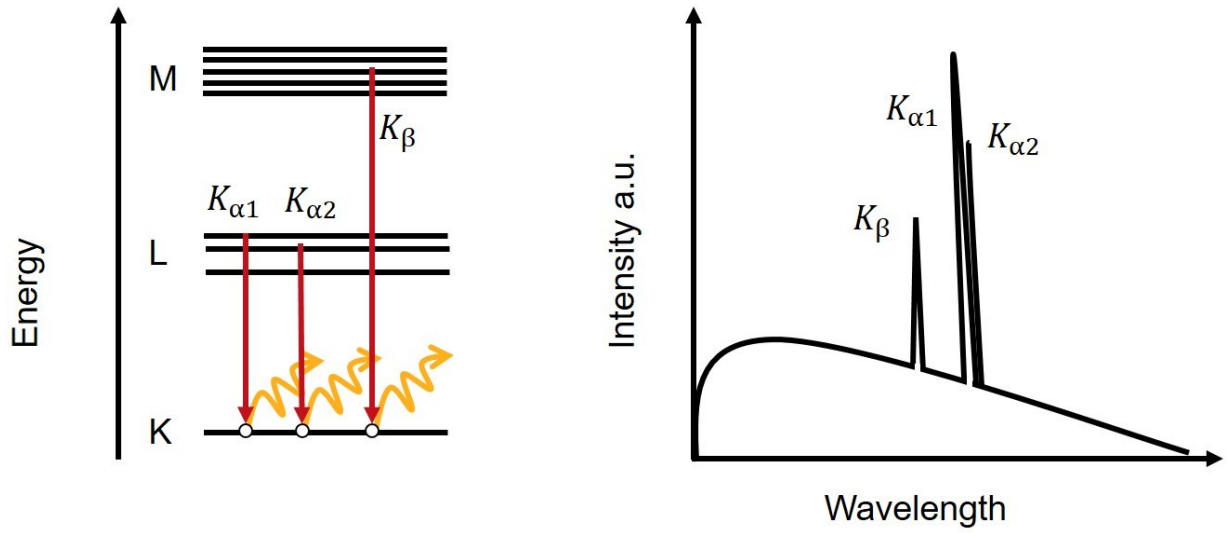


Figure 1-2. Generation of X-rays and the emission spectrum for a Cu X-ray tube.

1.3.1. Single-crystal diffraction

A crystal structure is defined by a lattice which describes the long-range periodicity. The lattice points are outlined by a unit cell, which is the smallest repeat unit, with lengths a, b, c and angles α, β, γ. In 3D, seven crystal systems are possible and when centring operations are included, 14 Bravais lattices are generated. A space group describes the set of point and translational symmetry operations possible within a crystal structure.

The conditions for X-ray diffraction can be described in two ways (Figure 1-3). According to the Laue equations, constructive interference takes place at the intersection of Laue cones along which scattered X-rays emanate, whereas according to Bragg's law, constructive interference takes place when the path difference between X-rays reflected from lattice planes is equal to an integral multiple of wavelengths.[9]

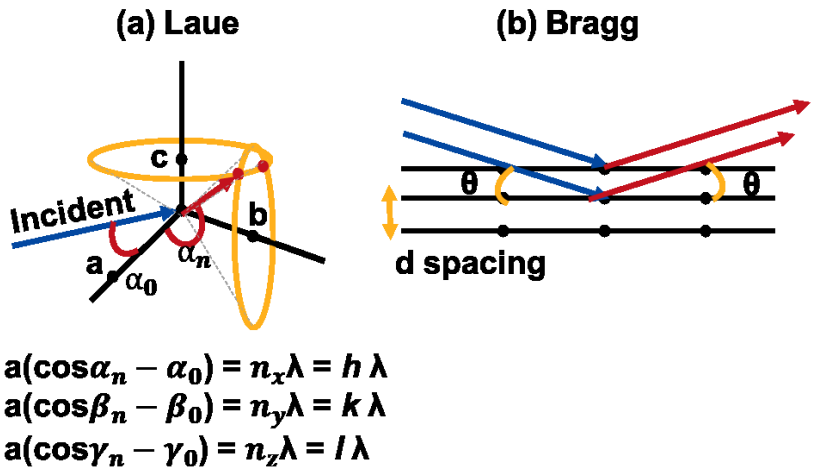


Figure 1-3. (a) Laue and (b) Bragg conditions for X-ray diffraction.

Suitable crystals are typically 0.01 to 0.5 mm in dimension and are mounted on a goniometer of a diffractometer. In this work, a Bruker PLATFORM diffractometer is used which is equipped with a SMART APEX II CCD area detector and a Mo K_α X-ray source. Frames of intensity data are collected, usually for 10 to 30 seconds. The intensities I_{hkl} are proportional to the square of the amplitude of the structure factor, F_{hkl}, which is the superposition of scattered waves for a given set of lattice planes (*hkl*) and depends on the scattering factors, displacement parameters, type, and location of atoms. The electron density function of a crystal can be obtained by a Fourier transform of the set of structure factors. Although the amplitudes of the structure factors can be obtained experimentally from the intensities, there is no information about their phases. Instead, structure solution proceeds through guessing these phases. The calculated structure factors from a proposed model are compared with the experimental structure factors, and the model is refined until the differences between these structure factors are minimized (as indicated by a low residual index).

1.3.2. Powder diffraction

X-ray diffraction patterns can also be collected on powder samples, which is prepared by fine grinding. Because a powder contains small crystallites oriented randomly, diffraction rings are observed instead of discrete spots, and intensities for symmetry-equivalent reflections cannot be resolved. Structure determination is more difficult than in single-crystal X-ray diffraction. However, typically an experimental powder pattern can be compared with a simulated pattern to see if a desired phase has been formed or to identify multiple phases in a sample. In this work, an Inel powder diffractometer is used which is equipped with a curved position-sensitive detector, allowing for simultaneous collection of intensities over an angular range of 0 to 120° in 2θ.

1.4. Scanning electron microscopy

The surface topology and composition of solids can be examined by scanning electron microscopy. Within a high-vacuum environment, electrons are accelerated (typically over a voltage of 20 kV) and when they undergo several types of interactions when they strike a sample. Secondary electrons emitted from the surface through inelastic scattering processes are used to probe the surface topology. Energy-dispersive X-ray (EDX) spectroscopy is used to determine the composition of the sample. When the incident electron beam ejects core electrons, higher-level electrons relax to the hole and X-rays characteristic of the element are emitted. The amount of a given element can be detected to roughly ~2 wt. %, provided that they are heavier than Na. Lighter elements are harder to detect because of the possibility of reabsorption of low-energy X-rays by the sample.

1.5. Band structure calculations

A crystalline solid contains a periodic arrangement of atoms whose orbitals can overlap to form energy levels that come so close together that they are described as quasicontinuous bands rather than discrete levels. To solve the Schrödinger equation, a few assumptions are made. First, electron-electron interactions are neglected and the Kohn-Sham one-electron equation is solved instead of dealing with the many-electron case.[10] Second, by taking advantage of the translational invariance of a periodic solid, Bloch's theorem expresses the wavefunctions as a linear combination of atomic orbitals:

$$\Psi(k,r) = \sum_j \exp\{ikr_j\}\phi_j$$

where r_j is the position of the atom j in the unit cell, ϕ_j is the orbital of the atom, and k is a wavevector in reciprocal space and measures the momentum of an electron. It suffices to plot the energies of these wavefunctions within the first Brillouin zone, which is the Wigner-Seitz cell (constructed by identifying midpoints to the nearest neighbour reciprocal lattice points, so as to portray symmetry more clearly around the reciprocal lattice origin) in reciprocal space. In 1D, for example, the Brillouin zone is bounded by $-\frac{\pi}{a} \leq k \leq \frac{\pi}{a}$ where a is the unit cell length.

A band dispersion diagram (E vs. k) plots the energies along high-symmetry points within the Brillouin zone. Another useful plot is the density of states (DOS) curve, which is inversely proportional to the slope of the bands. It is possible to extract the atomic projections to the DOS to determine which elements contribute to the bands at a particular energy. More information can be obtained by a plot of the crystal orbital Hamiltonian population (COHP) which can reveal the nature of bonding interactions (Figure 1-4). COHP values are derived from the density of states matrix weighted by the Hamiltonian matrix to describes the interaction between two orbitals belonging to neighbouring atoms.[11,12]

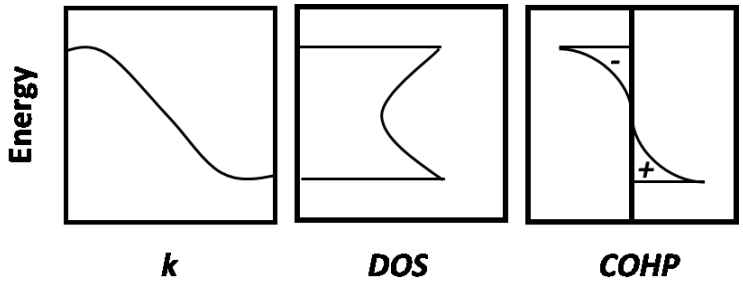


Figure 1-4. Band dispersion, density of states, and crystal orbital Hamilton population curves.

Calculations in this work were performed with the tight-binding linear muffin tin orbital (TB-LMTO) program package.[13,14] The tight-binding approximation treats the electrons as if they are tightly bound to the atom and the interactions are limited to neighbouring atoms. The form of the potential energy function experienced by electrons is assumed to be spherically symmetric around atoms but constant in the interstitial sites, resembling the shape of a muffin tin. The atomic-spheres approximation involves expanding the muffin-tin radii so that they are allowed to overlap slightly with the aim of filling up all space. This approximation is appropriate if the structure contains close-packed atoms, as normally found in intermetallic compounds. By default, the program uses the local density approximation which approximates the exchange-correlation energy by locally treating the electron density as a homogenous electron gas.

1.6. Machine learning

The term "machine learning" was first coined by Arthur Samuel in 1959 when he was studying how a program learns to play the game of checkers.[15] Later, Tom Mitchell remarked that "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and

performance measure *P*, if its performance at tasks *T*, as measured by *P*, improves with experience *E*."[16] The practical application of machine learning in materials science is an exciting area which is becoming more widespread, and its utility is being recognized by experimentalists to guide the high-throughput discovery of materials. In particular, it is more efficient and practical to use machine-learning algorithms rather than ab initio calculations to establish correlations in complex, many-body systems. Moreover, the digitization of scientific information has enabled chemists with the opportunity to leverage the analysis of large data as a new tool to complement computational work and intuition.

The application of machine learning to material science is a developing field with its utility and associated challenges still being illuminated. In addition to inorganic materials discovery, machine learning can also be used for high-throughput property prediction, optimization of experimental conditions, automation of experimental procedures, materials tuning, microstructure optimization, and even materials characterization.[17-22] Many of these applications require the description of materials in a machine interpretable way and there have been many efforts to design a fingerprint that is chemically interpretable.[23-25] The line between performance and interpretability is one we have to straddle in this field and the hope is that with a good fingerprint we can have both. The performance is also influenced by the data and algorithms due to the fact that we need enough good data for our prediction to be reliable and each algorithm will handle the data differently. The development of new machine learning architecture itself is a burgeoning field and what kind of architecture is best suited for our application is an ongoing question in the field.[26] In regards to the data aspect, although information is being digitized and stored, its organization, access and quality are still problems.[27]

In this work, we examine tasks *T* involving classification where other tools would be too costly in terms of computation, time, and effort. The general procedure works by first extracting and describing information. For example, the compositions of a series of compounds can be extracted from a database and a simplified description can be formulated. The experience *E* comes in the form of features which is anticipated to help distinguish between a set of classes. Because it is not obvious at the outset which features will do this optimally, both chemically and non-chemically intuitive features are considered. These features are selected based on the evaluation of the performance *P*, with the features selected that result in the best performance. One metric used to evaluate performance *P* is the *accuracy*, which is defined as the sum of the true positives and true negatives over the total number of samples.

1.6.1. Feature selection and principal component analysis

To develop a machine-learning model, feature selection is a necessary first step. What are features and how are they generated? Features are used to train machine-learning models. They can be extracted through an unsupervised learning algorithm, they can be generated by hand by processing data for information, or they can be proposed based on one's intuition. In chemistry, it makes sense to generate features based on the properties of elements, such as atomic radii or melting points. These properties can be combined to generate new ones. For example, taking the ratio of two properties of elements would be sensible for radii (to mimic Pauling's radius ratio rules) but perplexing for boiling points. However, the idea of the procedure of feature selection is to avoid bias.

Feature selection accomplishes several purposes. Pragmatically, it helps us interpret the model by identifying what features correlate well with the dataset. Because the input matrix is smaller, the learning times are faster. It also prevents overfitting. Having a description of the data

that is too fine-grained or having too many features increases the number of dimensions on which the data are projected, leading to an increase in sparsity and causing difficulties in finding a good decision boundary (because even the in-class separation can become greater as more detail is added). The algorithm may learn peculiarities associated with the dataset and may come up with an unreasonable decision barrier, which leads to poor generalizability of the model and overfitting.

In a procedure called cluster resolution feature selection, the discriminating power is improved by maximizing the distances of the confidence ellipses bounding two clusters.[28,29] The features are first ranked by their Fisher ratio, which is the ratio of in-class to between-class variance. In a binary classification problem, a good feature maximally separates classes and minimizes in-class variance. A principal component analysis (PCA) model is created starting from a subset of features with the highest Fisher ratios. Higher-dimensional data are projected onto 2D PCA space. The data are autoscaled to fit around the origin and are projected on a random line passing through the origin (Figure 1-5).[30] The distance between the data points and the origin is fixed, and we seek to maximize the sum of the squared distances of the projected points to the origin. The line of best fit (the "principal component") is a linear combination of the variables, and the weights in this linear combination are called loading scores.

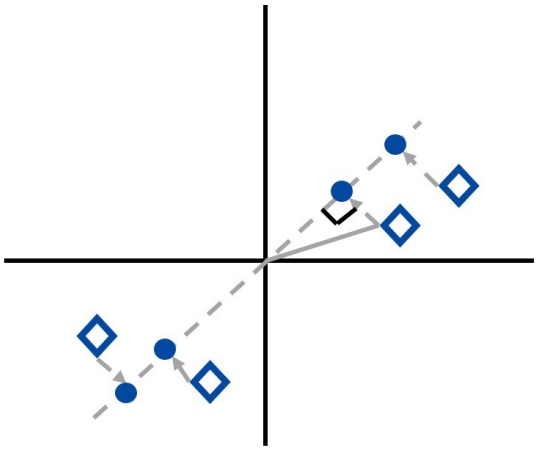


Figure 1-5. Projection of points onto a line of best fit including Pythagorean relationship.

The line is scaled by calculated its unit vector (called the eigenvector). The eigenvalue is then calculated, which determines how much variation is accounted for by each principal component. Further principal components are found by finding a line orthogonal to the previous principal component and repeating the procedure. The points can then be projected onto a lower-dimensional PCA space (Figure 1-6).

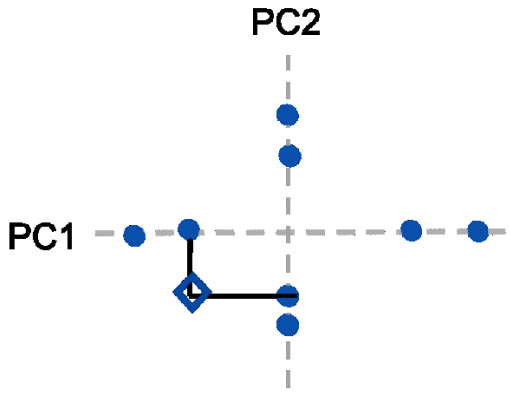


Figure 1-6. Projection of a point in PCA space consisting of two principal components.

In a binary classification problem, there are two possible classes, by definition. Samples belonging to one class are projected onto PCA space. The length of the confidence ellipse is generating using Hotelling's value, and the eigenvalues and directions of each principal component are given by the loading vectors. Points are distributed along the circumference of the ellipse. After both ellipses have been computed, the Euclidean distance between each point on both ellipses is calculated. The shortest distance is compared to half the distance between each point on the circumference. If it is shorter, then the ellipses are considered to be overlapping. The confidence limit is then changed until there is no overlap (Figure 1-7).

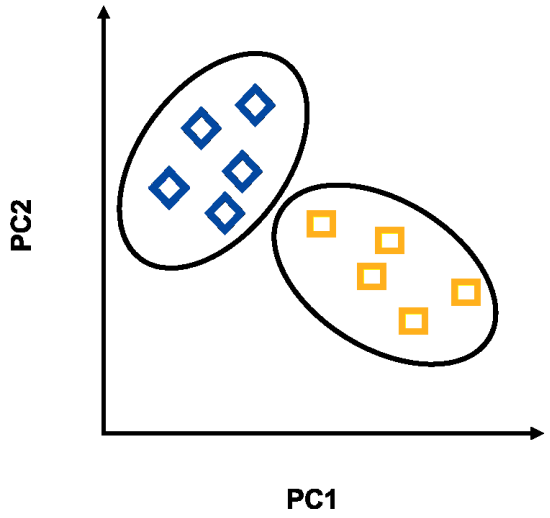


Figure 1-7. Non-overlapping confidence ellipses in PCA space.

The procedure is repeated in a stepwise fashion using backwards elimination and forward selection. In backwards elimination, a variable is removed and the performance of the model is evaluated. If the model improves, that variable is dropped, but if it gets worse, it is retained. Conversely, in forward selection, a variable is added and if the model improves, that variable is included, but if it gets worse, it is removed. This procedure attempts to circumvent the nesting problem of variables being eliminated too early. The combination of features is permuted for several generations and the survival rate of each variable is monitored. Models are generated at different survival rate thresholds and evaluated using the Matthew's correlation coefficient on the validation set. The subset of features that produces the best model is kept. The aim is to select features that make the model generalizable, rather than those that simply explain trends in the training set.

1.6.2. Partial least-squares discriminant analysis

Partial least-squares discriminant analysis can be used for dimensional reduction, feature selection, and classification. We consider here only the case of the base algorithm applied to a binary classification problem, in which a function is sought to discriminate two classes.[31] Each matrix X and Y is decomposed into matrices:

$$X = Tp + E \qquad Y = Tq + F \qquad w = X'y$$

$$t = \frac{Xw}{\sqrt{\sum w^2}} \qquad p = \frac{t'X}{\sqrt{\sum t^2}} \qquad q = \frac{y't}{\sqrt{\sum t^2}} \qquad b = w(pw)^{-1}q$$

where T is the score matrix, p is the x loading, q is the y loading, and E and F are residuals (also called errors or noise terms). A weight vector w is estimated which maximizes the covariance between X and Y. The covariance gives information about the strength of correlation between X and Y. This weight vector is then used to calculate the X score, which in turn is used to calculate the X and Y loadings. The X score is a projection of the training samples onto a new axis (called the PLS components), similar to how PCA is used to project data onto lower-dimensional space. The loadings are coefficients that related the variable (X or Y) and the PLS component. Subsequent PLS components are created from the residuals E and F, a new weight vector is calculated, and the process is repeated. For each PLS component, a regression coefficient b is calculated using the loadings and weight vectors. These coefficients are then collected in a matrix and can be used to predict Y from an input X. What makes this a classifier is that the Y variable is categorical (containing class 1 and class 0).

1.6.3. Synthetic minority oversampling and *k*-nearest neighbours

The *k*-nearest neighbours (KNN) algorithm classifies a data point on the basis of its surrounding neighbours (Figure 1-8). After the number of neighbours k to include is established,

the Euclidean distance between every point and the *k*th closest points are calculated. If *k* is too low, the neighbourhood is represented poorly, but if it is too high, the algorithm quickly becomes computationally expensive. A typical guideline to choose *k* is to start with the square root of the total number of samples, and round to the nearest odd integer to avoid any ties. Then *k* is varied to see how the performance of the model changes.

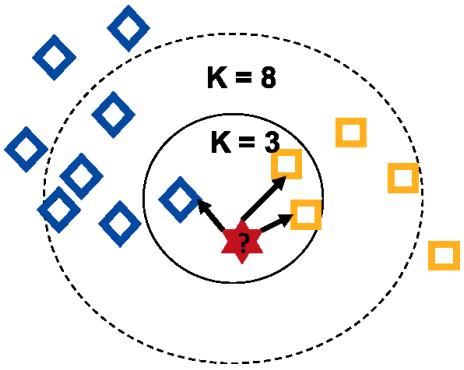


Figure 1-8. *k*-nearest neighbours' algorithm.

An algorithm that utilizes (KNN) is the synthetic minority oversampling technique (SMOTE), which is used to adjust the class distribution in unbalanced data sets, with the goal of improving the performance of the model by increasing the number of samples within the minority class (which also tends to be the class one wishes to predict).[32] Before SMOTE is applied, the data set must be partitioned into a calibration and a test set. SMOTE will only be used to balance the calibration set. In our case, we have a data set that contains two classes, with one being the minority. SMOTE then applies the *k*-nearest neighbour's algorithm to the minority class. In other words, for a sample in the minority class, it finds the *k* neighbours that are also part of the same minority class. Then a point is placed between them and this point is called a synthetic sample.

This is done for each sample in the minority class, resulting in the generation of a group of synthetic samples that are similar to the minority class. Each synthetic sample is evaluated by using the nearest neighbour algorithm once again, but the data set now includes the minority class, the majority class, and the synthetic samples. The synthetic samples are kept based on the criteria that their nearest neighbours consist of mainly samples belonging to the minority class. This voting threshold can be adjusted to make the algorithm stricter. For example, suppose we want 90% of the neighbours of the synthetic samples to consist of the minority class. This would give samples that are very similar to the original minority class, but will also reduce the number of synthetic samples generated. In the end, the classes may not be completely balanced but at least they are more balanced than before.

1.6.4. Support vector machine

The basic idea of a support vector machine is to find a hyperplane or decision barrier that best separates classes. Support vectors are a subset of points in each class that influence the decision barrier and form the margin of these classes (Figure 1-9).

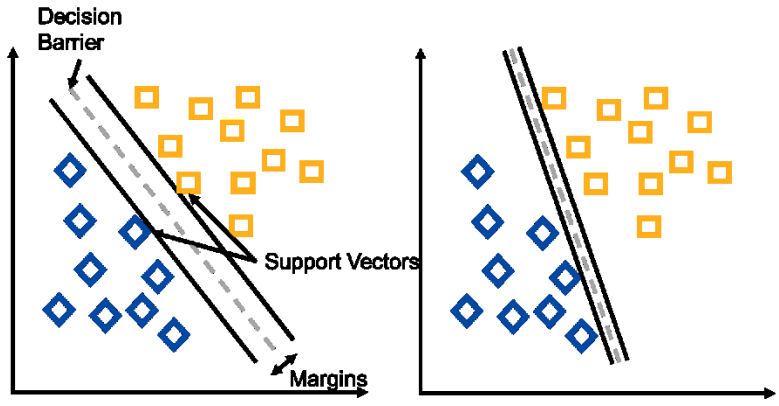


Figure 1-9. Choice of support vectors gives different decision barrier widths.

A set of support vectors is sought which maximizes the margin width. If the data are not linearly separable, two techniques can be employed. [33] The first technique is called the kernel trick, where a kernel function (Gaussian in this case) is used to project the data into higher dimensional space where it then becomes separable (Figure 1-10).

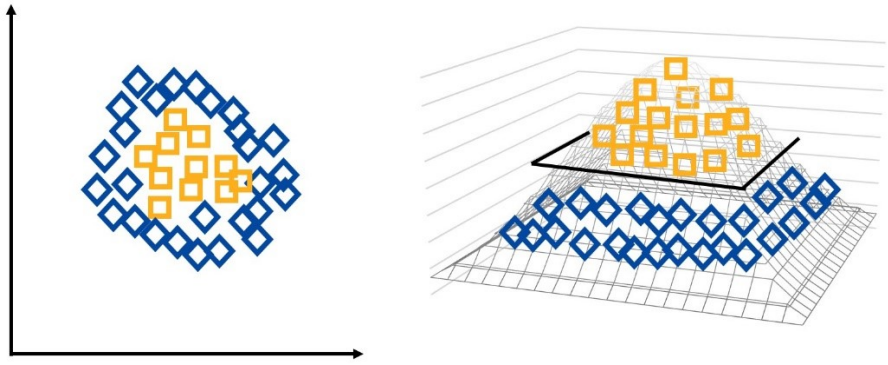


Figure 1-10. Projection of data using a Gaussian kernel function.

The second technique is called the soft margin approach where points are allowed to be on the wrong side of the margin and a penalty is added. The penalty depends on how far on the wrong side of the margin the point is located. Support vectors are then sought that minimize the penalty and also maximize the margin width.

1.7. Research motivation

Heusler compounds have attracted attention for their diverse physical properties, which can be tuned by chemical substitution or introducing deficiencies into their structures. Nevertheless, the uncomfortable fact is that the structures of many Heusler compounds have not been properly characterized. Out of 1371 reports of Heusler compounds, only 637 have had complete crystal structures determined, and of these, only 78 involved single-crystal X-ray

diffraction. Many of the subtle features about site occupation simply cannot be determined reliably from powder X-ray diffraction. The major goal of this thesis is thus to predict the correct assignment of atomic positions in Heusler compounds with the aid of machine-learning tools. This problem would be too time-consuming if every existing Heusler compound needed to be resynthesized and recharacterized experimentally, or if total energy calculations were to be performed. A secondary motivation is to illustrate the effectiveness of machine-learning techniques and to examine how conclusions can be drawn from good data. Throughout this work, predictions were experimentally validated whenever feasible, and the importance of empirical evidence is emphasized.

1.8. References

1) Heusler, F.; Stark, W.; Haupt, E. Über die synthese ferromagnetischer Manganlegierungen. *Verh. Deut. Phys. Ges* **1903**, *144*, 340–223.
2) Kübler, J.; Williams, A.R.; Sommers, C.B. Formation and coupling of magnetic moments in Heusler alloys. *Phys. Rev. B.* **1983**, *28*, 1745–1755.
3) Bradley, A. J.; Rodgers, J. W. The Crystal Structure of the Heusler Alloys. *Proc. R. Soc. London, Ser. A* **1934**, *144*, 340–359.
4) Khawam, A.; Flanagan, D.R. Solid-state kinetic models: basics and mathematical fundamentals. *J. Phys. Chem. B.* ***2006****, 110*, 17315–17328.
5) Vlack, V. *Materials science for engineers 6th ed*., Addison-Wesley: Reading, Massachusetts, **1975**.
6) Borg, R.J.; Dienes, G.J. *An introduction to solid state diffusion,* Academic press, **1988.**
7) Merkle, R.; Maier, J. On the Tammann rule. *Z. Anorg. Allg. Chem.* **2005**, *631*, 1163–1166.

8) West, A. R. Basic Solid State Chemistry, 2nd ed., Wiley: New York, **1999**.

9) Massa, W. Crystal Structure Determination, 2nd ed., Springer–Verlag: Berlin, **2004**

10) Dronkowski, R. *Computational chemistry of solid-state materials*, Wiley-VCH, Weinheim, **2005**.

11) Dronkowski, R.; Blöchl, P.E. Crystal orbital Hamiltonian populations (COHP) Energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *J. Phys. Chem.* **1993**, *97*, 8617–8624.

12) Deringer, V.L; Tchougreeff, A.L.; Dronskowski, R. Crystal orbital Hamiltonian population analysis as projected from plane-wave basis sets. *J. Phys. Chem. A* **2011**, *115*, 5461–5466.

13) Skriver, H.L. *The LMTO method,* Springer-Verlag Berlin Heidelberg, **1984.**

14) Tank, R.; Jepsen, O.; Burkhardt, A.; Andersen, O.K.; *TB-LMTO-ASA program*, version 4.7; Max Planck Institut für Festkörperforschung: Stuttgart, Germany, **1998.**

15) Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229.

16) Mitchell, T.M. *Machine learning*, Mcgraw-Hill, **1997**.

17) Tehrani, A.M.; Oliynyk, A.O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T.D.; Brgoch, J. Machine learning directed search for ultraincompressable, superhard materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844–9853.

18) Zhang, H.; Moon, S.K.; Ngo, T.H. Hybrid machine learning method to determine the optimal operating process window in aerosol jet 3d printing. *ACS Appl. Mater. Interfaces* **2019**, *11*, 17994–18003.

19) Rashidi, M.; Wolkow, R.A. Autonomous scanning probe microscopy in situ tip conditioning through machine learning. *ACS Nano* **2018**, *12*, 5185–5189.

20) Hou, Z.; Takagiwa, Y.; Shinohara, Y.; Xu, Y.; Tsuda, K. Machine-learning-assisted development and theoretical consideration for the $Al_2Fe_3Si_3$ thermoelectric material. *ACS Appl. Mater. Interfaces* **2019**, *11*, 11545–11554.

21) Liu, R.; Kumar, A.; Chen, Z.; Agrawal, A.; Sundararaghavan, V.; Choudhary, A. A predictive machine learning approach for microstructure optimization and materials design. *Sci. Rep.* **2015**, 11551.

22) Hong, S.; Nomura, K.; Krishnamoorthy, A.; Rajak, P.; Sheng, C.; Kalia, R.K.; Nakano, A.; Vashishta, P. Defect healing in layered materials: a machine learning-assisted characterization of MoS2 crystal phases. *J. Phys. Chem. Lett.* **2019**, *10*, 2739–2744.

23) Batra, R.; Tran, H.D.; Kim, C.; Chapman, J.; Chen, L.; Chandrasekaran, A.; Ramprasad, R. A general atomic neighbourhood fingerprint for machine learning based methods *J. Phys. Chem. C.* **2019**, *123*, 15859–15866.

24) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564–3572.

25) Xie, T.; Grossman, J.C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys.Rev. Lett.* **2018**, 145301.

26) Lengeling-Sanchez, B.; Outeiral, C.; Guimaraes, G.L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *Chemrxiv.* **2017**, Preprint.

27) Hill J., Mannodi-Kanakkithodi A., Ramprasad R., Meredig B. *Materials Data Infrastructure and Materials Informatics*. In: Shin D., Saal J. (eds) *Computational Materials System Design*. Springer: Cham **2018**.

28) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

29) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

30) Abdi, H.; Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010,** *2*, 433–459.

31) Lee, L.C.; Liong, C.Y.; Jemain, A.A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst,* **2018**, *143*, 3526–3539.

32) Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

33) Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Transactions on intelligent systems and technology,* **2011**, *2*, 1–27.

Chapter 2 Solving the Colouring Problem in Half-Heusler Structures: Machine-Learning Predictions and Experimental Validation

2.1. Introduction

One of the most fundamental problems in solid state chemistry is establishing how different types of atoms are distributed within various sites of a crystal structure. A classic example often mentioned in inorganic chemistry textbooks is the occurrence of normal vs. inverse spinels AB_2O_4, in which the differing occupation of octahedral and tetrahedral sites by the metal cations *A* and *B* can be rationalized by crystal field stabilization energies that depend on d-electron configurations.[1] Because crystal structures have a formal relationship to the mathematical concept of a graph, determining site preferences of atoms within a solid has been described as a case of a "colouring problem" in graph theory.[2,3]

The colouring problem is particularly pertinent for the large family of intermetallics known as half-Heusler compounds, which have the equiatomic composition *ABC* and display diverse properties useful for many applications (e.g., thermoelectrics, spintronics, topological insulators).[4–9] Their structure can be described in various ways. One helpful perspective is to merge the NaCl-type (rocksalt) and ZnS-type (zincblende) structures (Figure 1a) to result in the cubic structure (space group $F\bar{4}3m$) of half-Heusler compounds, also called the MgAgAs-type structure (Figure 1b).[10] The structure features three sites at 4*a* (0, 0, 0), 4*b* (½, ½, ½), and 4*c* (¼,¼,¼). The coordination geometries around these sites, as defined by the nearest-neighbour environments, are tetrahedral for 4*a* and 4*b*, and cubic for 4*c*.[11] The 4*a* and 4*b* sites are equivalent in the sense that if their occupation by the elemental components is swapped, the structure remains

unchanged.[12] Then, the colouring problem in half-Heusler compounds *ABC* can be expressed succinctly as, "Which of the components, *A*, *B*, or *C*, prefers to occupy the 4*c* site?"

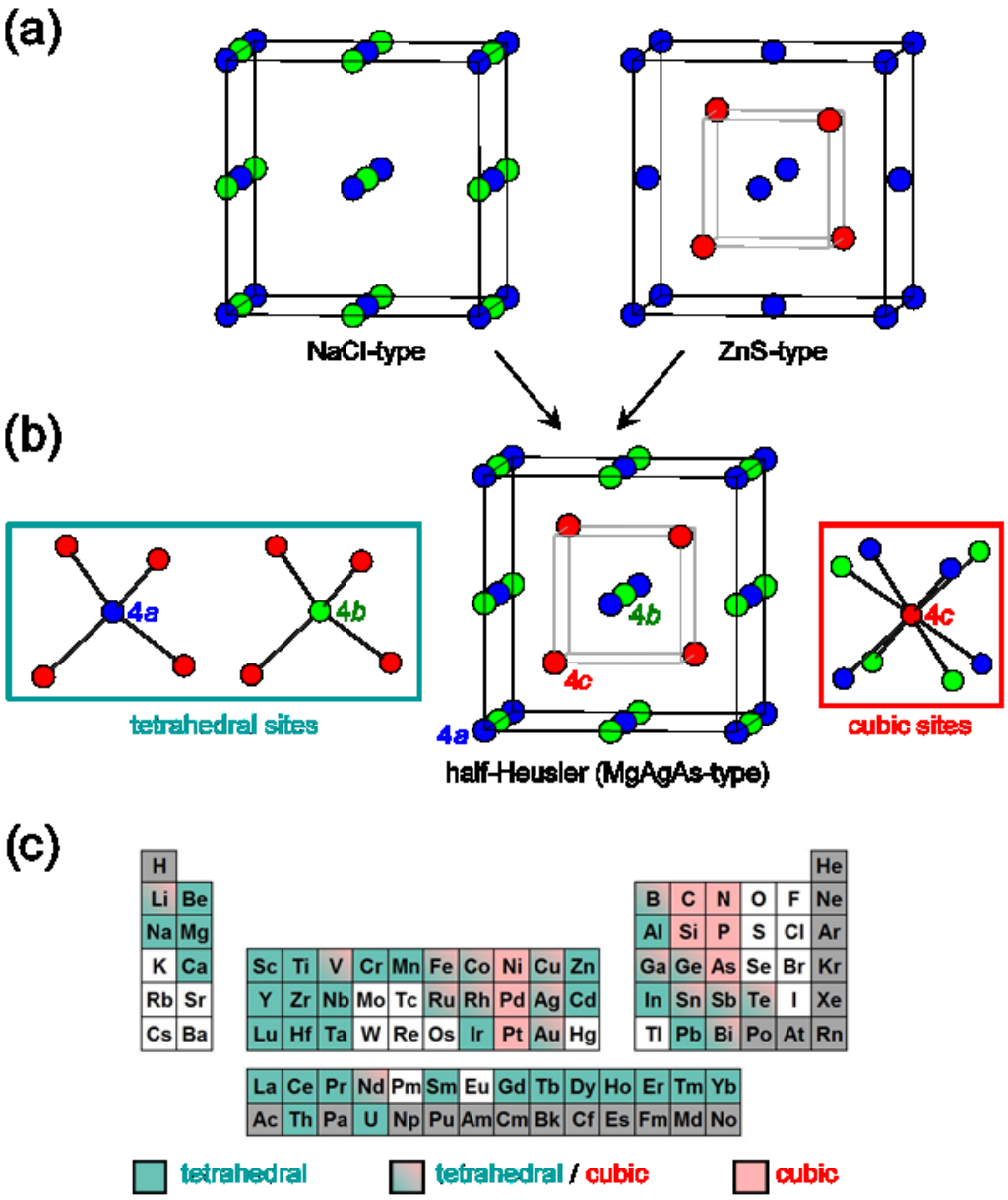


Figure 2-1. Merging of (a) NaCl-type and ZnS-type substructures results in (b) half-Heusler structure containing tetrahedral (at 4*a* and 4*b*) and cubic sites (at 4*c*), which can be occupied by a variety of elements as summarized in (c).

A simple rule has been proposed to predict these site preferences: The NaCl-type substructure, which entails greater ionic bonding character, should consist of atoms with a larger difference in electronegativity, whereas the ZnS-type substructure, which entails greater covalent character, should consist of atoms with a smaller difference in electronegativity.[4] Put another way, this means that the 4*a* and 4*b* sites should contain the most and least electronegative atoms and the 4*c* site should contain the atoms with intermediate electronegativity. How well does this rule work? For MgAgAs itself, the eponymous representative of this structure type, the NaCl-type substructure is predicted to consist of Mg and As atoms, and the ZnS-type substructure of Ag and As atoms; i.e., the 4*c* site should be occupied by Ag atoms. The prediction is not borne out by experiment, which shows that the 4*c* site is actually occupied by As atoms.[10] In fact, analysis of all half-Heusler structures reported to date in Pearson's Crystal Data shows that the situation is not so clear cut.[13] A wide variety of elements can occupy these sites; some elements are found exclusively in tetrahedral, some exclusively in cubic, and some in both sites (Figure 1c). Moreover, depending on the choice of electronegativity scale, the accuracy of these predictions ranges from poor (using the Gordy scale)[14] to modest (using the Allred-Rochow scale)[15] (Table 1).

Table 2-1. Accuracy in Predictions of 4*c* Site Preference in Half-Heusler Compounds [a]

electronegativity scale	accuracy
Gordy	0.313
Pauling	0.375
Mulliken	0.492
Martynov-Batsanov	0.525
Allred-Rochow	0.659

[a] Based on analysis of all reported half-Heusler compounds in Pearson's Crystal Data.

Of course, the rule of electronegativity differences can be extended by including other factors, such as size or electron configurations, which certainly influence crystal structures. More rigorous quantum calculations have been made to evaluate the relative importance of ionic vs. covalent bonding character in dictating the site preferences, but these studies are necessarily restricted to limited sets of compounds.[16–18] It is also inevitable that ambiguous cases will arise when electronegativities or sizes are similar.

Establishing the correct site distribution is essential because any conclusions about structure-property relationships, and thereby efforts to improve on the properties of half-Heusler or other compounds, rest on the premise that the crystal structure is accurate. Unfortunately, there have been many reports in the literature in which site distributions have been assumed without independent corroboration, or computational studies have been performed on hypothetical or experimentally unconfirmed structures. Site distributions are often not definitive in the literature or assigned inconsistently within databases; for example, it is unclear if the 4*c* site in MnPdSb is occupied by Pd or Sb atoms.[13,19–22] The electronic structure can change drastically if different site distributions are invoked; for example, Ni*M*Sn and Ni*M*Sb (*M* = Ti, Zr, Hf) have been computed to be narrow-gap semiconductors, zero-gap semimetals, or metals, depending on which structural model is chosen.[23]

Most experimental structural studies of half-Heusler compounds are based on powder X-ray diffraction (XRD) data because, with no refinable positional parameters, there is no strong incentive to perform single-crystal XRD experiments. Of the 720 entries of half-Heusler compounds listed in Pearson's Crystal Data, only 42 (6%) were examined using single-crystal data, and of these, 9 (1%) included a careful evaluation of site distributions, while the remainder were derived by analogy to related compounds.[13] As an illustration of the ambiguities that can

arise, the powder XRD patterns of LiAuSb are simulated with the 4*c* site being occupied by Li, Au, or Sb atoms (Figure 2).[24] Light elements like Li are undetectable in the presence of heavily scattering elements like Au and Sb, and the XRD patterns for the models with Au or Sb atoms in the 4*c* site become indistinguishable. This does not mean that the latter two models are equally viable; rather, the implication is that the X-ray diffraction method is agnostic about which model is likely to be correct. To resolve the quandary, alternative experimental techniques (such as neutron diffraction) or computational support may prove helpful.

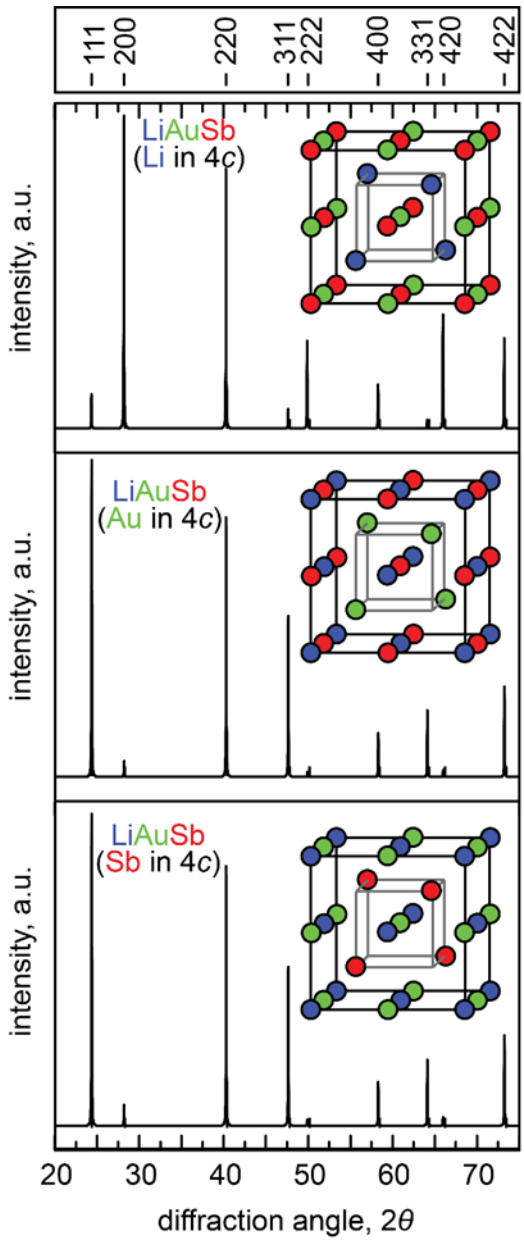


Figure 2-2. Simulated powder XRD patterns for LiAuSb with the 4*c* site being occupied by Li, Au, or Sb atoms.

High-throughput first-principles calculations have been performed on many half-Heusler compounds; these include studies that target hypothetical members, whose stabilities are evaluated by identifying energy minima relative to other competing phases.[25–28] As an alternative approach, we have been interested in applying machine-learning techniques to accelerate materials discovery, especially of intermetallic compounds.[29,30] Both methods have their own advantages and disadvantages. In first-principles calculations, by definition, no prior empirical information is required, but various approximations must be made to manage the time and cost of these calculations. In machine-learning methods, a model can be rapidly trained using empirical data, but the success depends on the quality of these data. Recently, Legrain et al. have compared these approaches as applied to the question of existence vs. nonexistence of half-Heusler compounds.[31] They demonstrated that predictions from different first-principles calculations are not fully consistent, but a machine-learning model based solely on chemical composition can perform very well and serve as a complement by highlighting unrecognized factors that influence stability. The ultimate arbiter, however, is experimental evidence to test the veracity of these predictions.

Either approach relies on identifying the correct site distributions within the structure of half-Heusler compounds. In first-principles calculations, this requires a search for the lowest energy candidate among models with different permutations of atoms within sites; in machine-learning methods, this requires an evaluation of the entries in crystallographic databases. However, as indicated above, it is not clear how reliable these crystallographic data really are, given the challenges in the diffraction analysis, and it is not clear how well the electronegativity difference rule really works, if the reported structures have not been independently verified. We set out several goals in the present study: (1) to develop and evaluate a machine-learning model that accurately predicts which component occupies the 4*c* site in the structures of half-Heusler

compounds *ABC*, (2) to apply this model as a data-sanitizing tool so that potentially incorrect structures reported in databases can be flagged, and (3) to test predictions in ambiguous cases (in which the electronegativity difference rule is unhelpful) by experimental synthesis and characterization.

2.2. Experimental Section

2.2.1. Machine-Learning Model

The problem at hand is to predict site distributions in half-Heusler compounds. In machine learning, this can be formulated as a classification problem, which we tackle by supervised learning using a training set of data. Crystallographic data were extracted for all half-Heusler compounds (MgAgAs-type structure, space group $F\bar{4}3m$) reported in Pearson's Crystal Data,[13] subject to the constraints that they do not contain hydrogen, noble gases, or elements with $Z > 83$, and that they exhibit fully ordered structures with no deviations from the ideal composition *ABC*. In total, there were 179 such compounds. For each compound, three structural variants were generated in which the 4*c* site is occupied by *A*, *B*, or *C*. The data set thus consists of 537 samples, categorized into "Class 1" containing 179 entries having the experimentally reported site distributions and "Class 2" containing 358 entries having alternative site distributions. (We use the term "sample" to be synonymous with "entry," "example," or "data point" in a machine-learning data set.)

Descriptors for the machine-learning model were derived from 43 properties of each element encompassing size, electronegativity, number of electrons, and others (Table S1 in Supporting Information). Atoms that occupy the tetrahedral 4*a* and 4*b* sites are interchangeable, and we anticipate that the structural preferences will be influenced strongly by how these atoms

can be discriminated from the ones that occupy the cubic 4*c* site. The elemental properties were thus combined through 6 arithmetic operations that express differences or ratios between values for the atoms occupying 4*a*/4*b* vs. 4*c* sites (Table S2 in Supporting Information), giving a total of 258 features or variables to be potentially used in the machine-learning model.

A machine-learning pipeline was developed using the PLS_Toolbox software (version 8.0.1),[32] implemented through the MATLAB (2018a release) interface.[33] The data were preprocessed by autoscaling (mean-centering and scaling to unit variance) and normalization to the sum of absolute values. Two-thirds of the data were assigned to a training set, and one-third to a validation set. A support vector machine (SVM) classifier algorithm was applied with a Gaussian radial basis function. The SVM classification was carried out with a venetian-blind cross-validation with a 10-fold data split. An important step in building the pipeline was to apply cluster-resolution feature selection (CR-FS), in which the choice of features is optimized through a systematic procedure involving backward elimination and forward selection.[34,35] The features are ranked according to their Fisher ratios, which are the ratios of between-class and in-class variabilities, and thus measure their discriminating ability. In backward elimination, a feature is successively removed starting from the lowest-ranked ones, and if the model improves, that feature is rejected. Conversely, in forward selection, a feature is successively added starting from the highest-ranked ones, and if the model improves, that feature is retained. Features were also scrubbed if values were missing or gave division-by-zero errors. Ten different iterations with 100 rounds of feature selection were performed and 1000 different models were generated. In the end, the best performing model was generated by identifying 23 features which were the most common ones and had the highest survival rates among the 1000 models. The survival rate was obtained

by noting which features survive after 100 rounds of feature selection; the 23 features survived 99 out of 100 times and were present across all 10 iterations of the procedure.

2.2.2. Synthesis of Half-Heusler Compounds

Two sets of compounds were synthesized anew, to test the predictions of the machine-learning model. In the first set, MnIrGa, MnPtSn, and MnPdSb were prepared to illustrate the importance of the CR-FS procedure in constructing a reliable machine-learning model. In the second set, GdPtSb and HoPdBi were flagged as among the most likely candidates to have incorrect site distributions as reported in Pearson's Crystal Data.

Starting materials were freshly filed Gd and Ho pieces (99.9%, Hefa), Mn powder (Alfa-Aesar, 99.95%), Ir powder (Cerac, 99.9%), Pd powder (Alfa-Aesar, 99.95%), Pt sponge (99.9%), Sn drops (Anachemia, 99.9%), Sb powder (Cerac, 99.995%), and Bi powder (Aldrich, 99.99+%). The elements were combined in equimolar ratios with a total mass of 0.3 g and pressed into pellets. These pellets were arc-melted three times on a copper hearth under an argon atmosphere, with the ingots being flipped each time, in an Edmund Bühler MAM-1 arc melter. The ingots were then placed in fused-silica tubes which were evacuated and sealed. The ingots were annealed at 1273 K for 7 d, and then quenched in cold water. The products were finely ground and analyzed by powder XRD on an Inel diffractometer equipped with a curved position-sensitive detector (CPS 120) and a Cu $K\alpha_1$ radiation source operated at 40 kV and 20 mA. If secondary phases were detected, the products were reground and pressed into pellets, and the arc-melting and annealing steps were repeated.

For all compounds synthesized, Rietveld refinements were performed on the powder XRD data using the TOPAS Academic software package.[36] Additionally, for HoPdBi, single-crystals were available which were confirmed to have an equiatomic composition by energy-dispersive X-

ray (EDX) analysis carried out on a JEOL JSM-6010LA InTouchScope scanning electron microscope. Single-crystal XRD data for HoPdBi were acquired at room temperature on a Bruker PLATFORM diffractometer equipped with a SMART APEX II CCD detector and a graphite-monochromated Mo $K\alpha$ radiation source, using ω scans at 6 different ϕ angles with a frame width of 0.3º and an exposure time of 15 s per frame. Face-indexed numerical absorption corrections were applied. Structure solution and refinement were carried out with use of the SHELXTL (version 6.12) program package.[37] The cubic space group $F4\bar{3}m$ was chosen on the basis of Laue symmetry, intensity statistics, and systematic absences. Structure refinement proceeded in a straightforward fashion and resulted in excellent agreement factors with no significant residual density observed.

2.2.3. First-Principles Calculations

To determine the total energies of GdPtSb and HoPdBi adopting structures with different site distributions, electronic structure calculations were performed using the Vienna ab initio simulation (VASP) package, within the Perdew-Burke-Ernzerhof generalized gradient approximation and with projector-augmented wave potentials applied.[38–40] A cut-off energy of 500 eV for the pseudopotential basis set were used. For the density of states (DOS) calculation, a Γ-centred *k*-point mesh of 12 × 12 × 12 was used. The criterion for energy convergence was set to 1×10^{-8} eV. To visualize the valence electron and charge distributions, the electron localization function (ELF) was calculated and a Bader charge analysis was performed.[41,42]

Bonding characteristics were evaluated through an energy-resolved crystal orbital Hamilton populations (COHP), which were extracted from the electronic structures calculated from the tight-binding linear muffin-tin orbital program with the atomic spheres approximation (TB-LMTO-ASA, version 4.7).[43,44] The basis sets included Ho 6s/(6p)/5d/4f, Gd 6s/(6p)/5d/4f,

Pd 5s/5p/4d/(4f), Pt 6s/6p/5d/(5f), Sb 5s/5p/(4d)/(4f), and Bi 6s/6p/(6d)/(5f) orbitals, with the orbitals shown in parentheses being downfolded. Integrations in reciprocal space were carried out using a *k*-point mesh of 8 × 8 × 8 with an improved tetrahedron method.

2.3. Results and Discussion

2.3.1. Machine-Learning Predictions

A machine-learning model to predict site distributions in half-Heusler compounds *ABC* has been built by means of an SVM algorithm. The model was trained on crystallographic entries appearing in Pearson's Crystal Data[13] and using descriptors solely based on elemental properties. For each compound, three site distributions are possible, depending on which element (*A*, *B*, *C*) occupies the cubic 4*c* site. The data consisted of 537 samples (179 compounds with 3 site distributions), labeled as Class 1 for those with the experimentally reported site distributions and Class 2 for those with alternative site distributions. The results of the machine-learning model are summarized graphically as plots of predicted probability for the correctness of the site distribution vs. sample number (Figure 3). The numerical values of these probabilities for individual samples are also listed (Table S3 in Supporting Information).

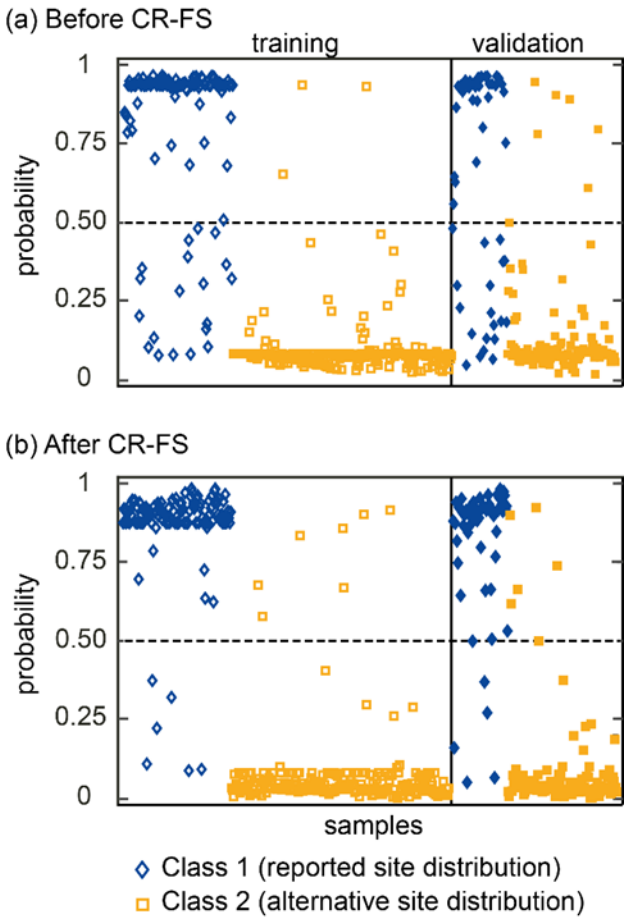


Figure 2-3. Probability for correctness of site distributions (a) before and (b) after CR-FS procedure applied.

Samples having probabilities close to 1 (as expected for Class 1 samples, showing agreement with the reported site distributions) are starkly delineated from those having probability close to 0 (as expected for Class 2 samples, with alternative site distributions). Overall, the machine-learning model has excellent predictive ability, achieving an accuracy of 95%, which is a significant improvement over the best accuracy of 66% attained when the electronegativity difference rule is applied (using the Allred-Rochow scale). Other statistical performance measures for the machine-learning model are impressive (Table 2).

A systematic procedure for feature selection called CR-FS has been applied, which aims to optimize the model with the best choice of descriptors. Before CR-FS, there are a small number of misclassified samples that fall beyond the decision barrier (horizontal line at a probability of 0.50 in Figure 3a). Class 1 samples with low probabilities are potentially false negatives and the corresponding Class 2 samples with high probabilities are potentially false positives. These situations may suggest that the machine-learning model requires further improvement, or that the site distributions as reported in the literature are incorrect and one of the other two alternative site distributions are likely to be the correct one instead. We have demonstrated previously that other machine-learning classification models for predicting crystal structures benefit from this type of careful feature selection.[45,46] The descriptors were combinations of various elemental properties of the components *A*, *B*, and *C*. An estimate of their potential importance in the machine-learning model can be quantified by their Variable Importance in Projection (VIP) scores (Figure 4).[47] However, a high VIP score (blue bars in Figure 4) alone does not guarantee that a given descriptor will be retained, because it is a combination of many descriptors, not just individual ones, that matters in achieving the highest quality of a model. Applying the CR-FS procedure, which aims to select the best set of descriptors in an unbiased manner, results in only 23 descriptors that were retained in the final model (orange bars in Figure 4). Compared to the model without CR-FS applied (Figure 3a), where there is a smattering of misclassified samples, the probabilities of Class 1 vs. Class 2 samples are much more cleanly separated after CR-FS (Figure 3b) and the performance of the model improves (Table 2).

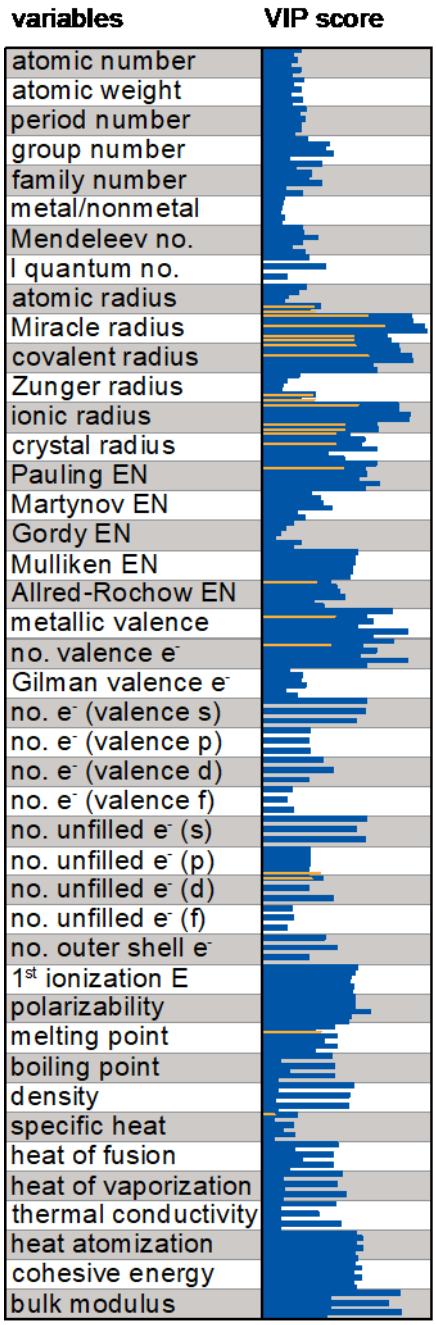


Figure 2-4. VIP scores for descriptors used in the machine-learning model before (blue bars) and after CR-FS (orange bars). For each elemental property, six arithmetic operations were applied.

Of course, there is a possibility that the literature reports could have been incorrect. To verify if the inconsistency originates from the model or from the experimental observations (or

both), we have identified three samples with low probabilities of being in Class 1 (before CR-FS) that merit further investigation: MnIrGa with Ga in 4*c* (probability of 0.127), MnPtSn with Sn in 4*c* (probability of 0.043), and MnPdSb with Sb in 4*c* (probability of 0.069) (Figure S1 in Supporting Information). In nearly all previous reports of these compounds, the site occupations were not explicitly deduced but rather were assigned by assumption. Therefore, we have resynthesized these compounds. Rietveld refinements of their powder XRD patterns confirm the revised site occupancies as indicated (Figure 5).

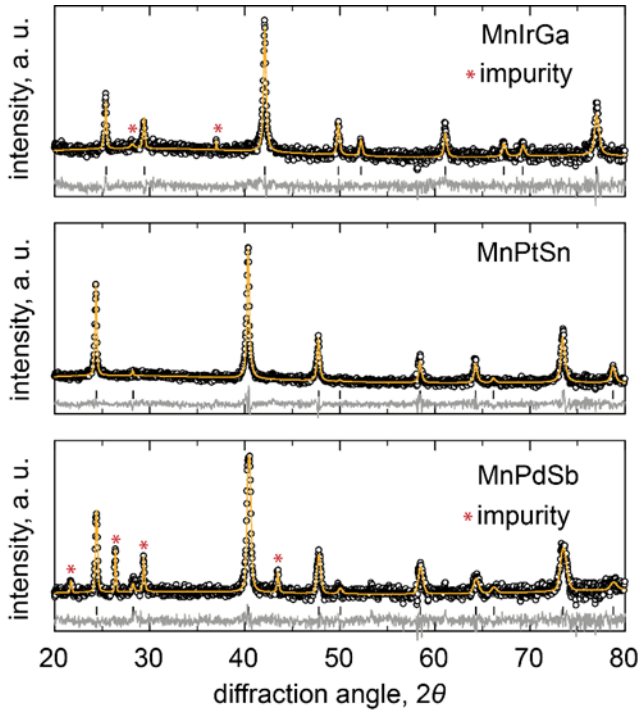


Figure 2-5. Rietveld refinements for MnIrGa with Ga in 4*c*, MnPtSn with Sn in 4*c*, and MnPdSb and Sb in 4*c*. Impurities were included in the peak profiles.

After CR-FS, these compounds are now correctly classified and have high probabilities of being in Class 1: MnIrGa with Ga in 4*c* (probability of 0.881), MnPtSn with Sn in 4*c* (probability of 0.881), MnPdSb with Sb in 4*c* (probability of 0.680) (Figure S1 in Supporting Information). Given that these site distributions agree with the electronegativity difference rule (here, the p-block component is the one of intermediate electronegativity while the precious metal component is the most electronegative), a skeptic may wonder if the machine-learning model offers any new insights.

To counter such a possible objection, we have examined the handful of misclassified samples remaining even after CR-FS (Figure S1 in Supporting Information). For example, it can be proposed that GdPtSb with Sb in 4*c* and HoPdBi with Bi in 4*c* are quite reasonable site distributions, by analogy to the compounds just discussed above. However, the machine-learning model suggests that these assignments are likely incorrect, and one of the alternative distributions in which the precious metal atoms enter the 4*c* site is a much more viable candidate (Figure 6).

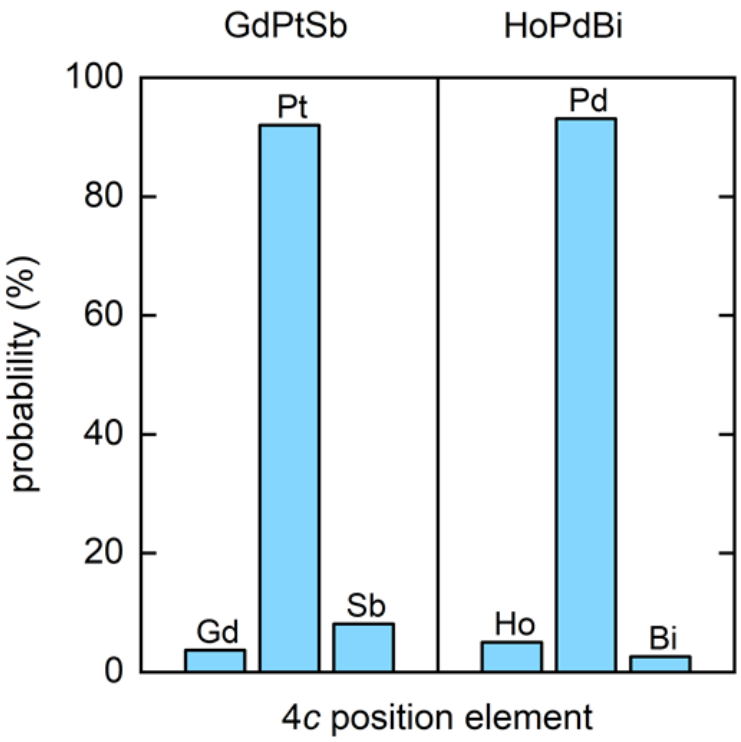


Figure 2-6. Probability of element occupying 4*c* site in GdPtSb and HoPdBi.

As before, we have resynthesized these compounds and carried out Rietveld refinements of their powder XRD patterns (Figure 7). Single crystals of HoPdBi were also available to allow a full structure determination, giving unambiguous proof for Pd occupying the 4*c* site. Crystal data for both the powder and single-crystal refinements are listed (Table 3). The revised CIFs have been submitted to CCDC. In general, inspection of the outliers indicates that pnictides and Li-containing compounds tend to be misclassified.

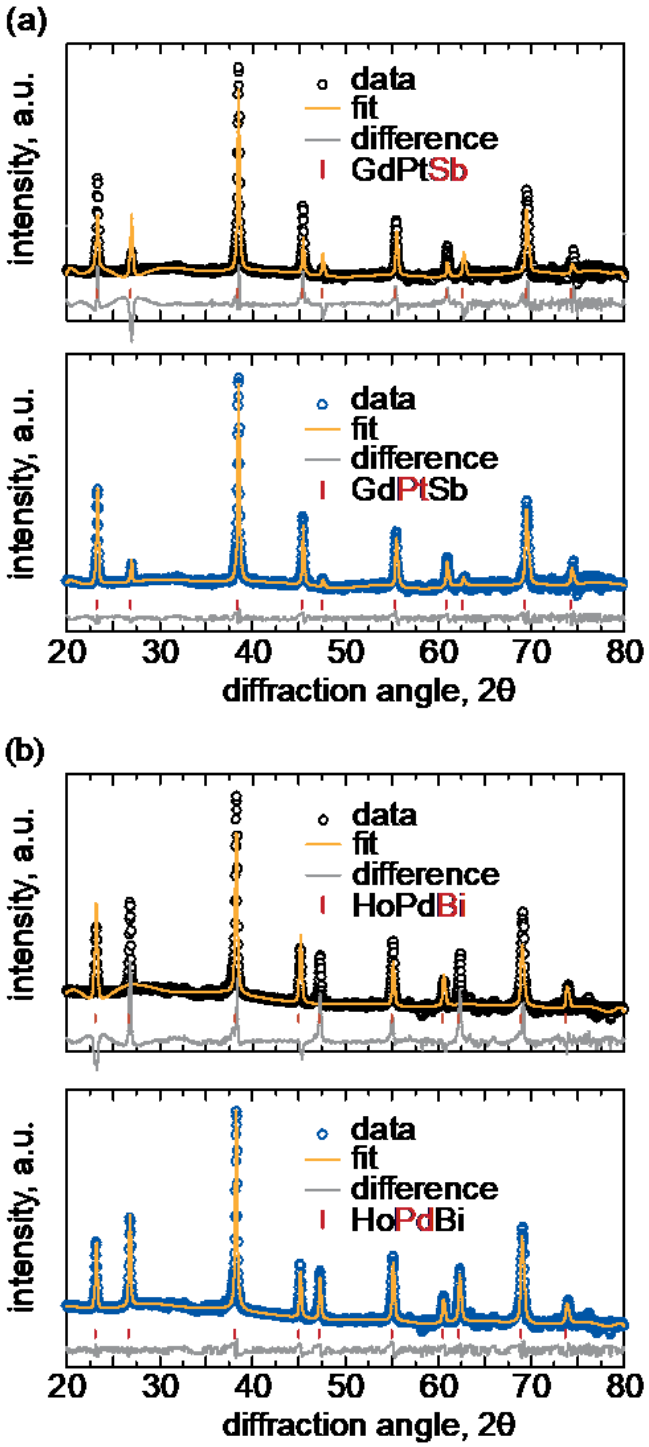


Figure 2-7. Rietveld refinements for (a) GdPtSb and (b) HoPdBi with different site distributions (the element highlighted in red in the formula is placed in the 4*c* site).

2.3.2. Factors Affecting Site Distributions in Half-Heusler Compounds

In traditional explanations for rationalizing the site distributions in half-Heusler compounds *ABC*, the electronegativity difference rule is sometimes augmented by guidelines invoking size differences or special cases of preferred occupation by certain elements.[4] The drawbacks to this approach are that continuous refinement of these rules is needed to explain new cases and that it is not clear what electronegativity or size scales are to be used, among many possible. In the machine-learning model, the most appropriate combination of these factors is chosen impartially. It is informative to take a closer look at the descriptors that influence the site distributions (Figure 4). Out of the 23 descriptors used in the final model, 15 are related to various radii and only 2 to electronegativity scales, but this does not imply that size factors are more important than electronegativity factors. Recall that these descriptors were built through 6 types of arithmetic operations applied to elemental properties. The greater number of size descriptors in the machine-learning model may simply reflect a more complicated mathematical dependence, but we offer another possible interpretation later. No one radius scale predominates, but rather a combination of six scales (atomic, covalent, ionic, crystal, Miracle, and Zunger pseudopotential radii) is required.[48] The inclusion of Miracle radii, which are derived from metallic glasses, is especially interesting because it reflects the metallic bonding character in half-Heusler compounds. Indeed, we propose that such radii should be used more frequently in relation to intermetallic compounds generally. Most of the remaining features (metallic valence, numbers of electrons) useful for classification pertain to distinguishing between transition-metal and p-block metalloids.

It is reassuring that the machine-learning model captures electronegativity as an influential factor for classification. In particular, the ratio of Pauling electronegativities and the difference in Allred-Rochow electronegativities between the atoms in the 4*c* vs. non-4*c* sites are highlighted,

essentially reproducing the electronegativity difference rule. The combination of these two electronegativity scales is very helpful in cases where there are discrepancies in their values. For example, in the cases of GdPtSb and HoPdBi mentioned above, these two scales actually give different predictions for the site occupancy of the 4*c* site. In the Pauling scale, Pd (2.20) and Pt (2.28) are more electronegative than Sb (2.05) and Bi (2.02), but in the Allred-Rochow scale, Pd (1.35) and Pt (1.44) are less electronegative than Sb (1.82) and Bi (1.67)![15,49,50] The experimental confirmation that Pd and Pt are the ones that really occupy the 4*c* site, conforming to the Allred-Rochow scale, vindicates the machine-learning approach which takes the best combination of features to account for the entire set of compounds. We caution that the apparent dominance of size over electronegativity factors cannot be interpreted too literally. Rather, what we can take away from this analysis is that size acts as an effective proxy for many other physical and quantum features, including electronegativity, for this classification problem. For example, Allred-Rochow electronegativities are calculated based on effective nuclear charges (which can be related to pseudopotential radii) and covalent radii.

It is interesting to compare how well the machine-learning models perform if they are built on electronegativity descriptors alone, or radius descriptors alone (Table 2). Keeping in mind that the number of descriptors is not the same, we note that both such models perform similarly, achieving an accuracy of 87–88% for the validation set, but this is still not as good as the full-feature model (accuracy of 95%). The improvement in accuracy is meaningful and can be traced to two major reasons. First, unclear cases (in which the probability of a correct site distribution falls in an ambiguous region when only electronegativity or radius descriptors are considered in the model) now become more definitive (in which the probability approaches 0 or 1 in the full-feature model). Second, the electronegativity-only or radius-only model sometimes leads to

nonsensical predictions. For example, all site distributions are predicted to be equally unlikely for MnPdSb (with low probabilities of 0.047, 0.019, and 0.007 for all possibilities) in the electronegativity-only model, and similarly for MnIrGa (with probabilities of 0.137, 0.369, and 0.063) in the radius-only model. The radius-only model also gives false negatives for experimentally validated cases of MnPtSn with Pt in 4*c* (probability of 0.182) and MnPdSn with Pd in 4*c* (probability of 0.039).

2.3.3. Total Energy Calculation and Charge Density Analysis for GdPtSb and HoPdBi

As further support for the corrected structures as suggested for GdPtSb and HoPdBi through the machine-learning model, total energy calculations were performed on these compounds with different site distributions. The site distributions with the highest probability (GdPtSb with Pt in 4*c* and HoPdBi with Pd in 4*c*) correspond to the lower total energy configurations, by 0.4–0.6 eV/atom, which is substantial.

Inspection of ELF plots and Bader charges (Figure 8), along with DOS and –COHP curves (Figure 9), shows significant differences in the electronic structures. GdPtSb in the wrong structure (Sb in 4*c*) shows negative charge found on the most electropositive atom Gd and Pt–Sb interactions that are strongly antibonding, which are chemically unreasonable characteristics. GdPtSb in the correct structure (Pt in 4*c*) shows charge distributions closer in line with expectations, the Fermi level falling near a pseudogap, and bonding interactions close to being optimized. HoPdBi in either structure appears to show reasonable charges for atoms, but the ELF plot for the wrong structure shows disturbingly high localization of electron density around the Bi atoms, which is not a realistic situation. Finally, it is interesting to point out that the oft-cited picture of the half-Heusler structure as the merging of a more ionic NaCl-type substructure with a more covalent ZnS-type substructure does not bear out on inspection of the charges.

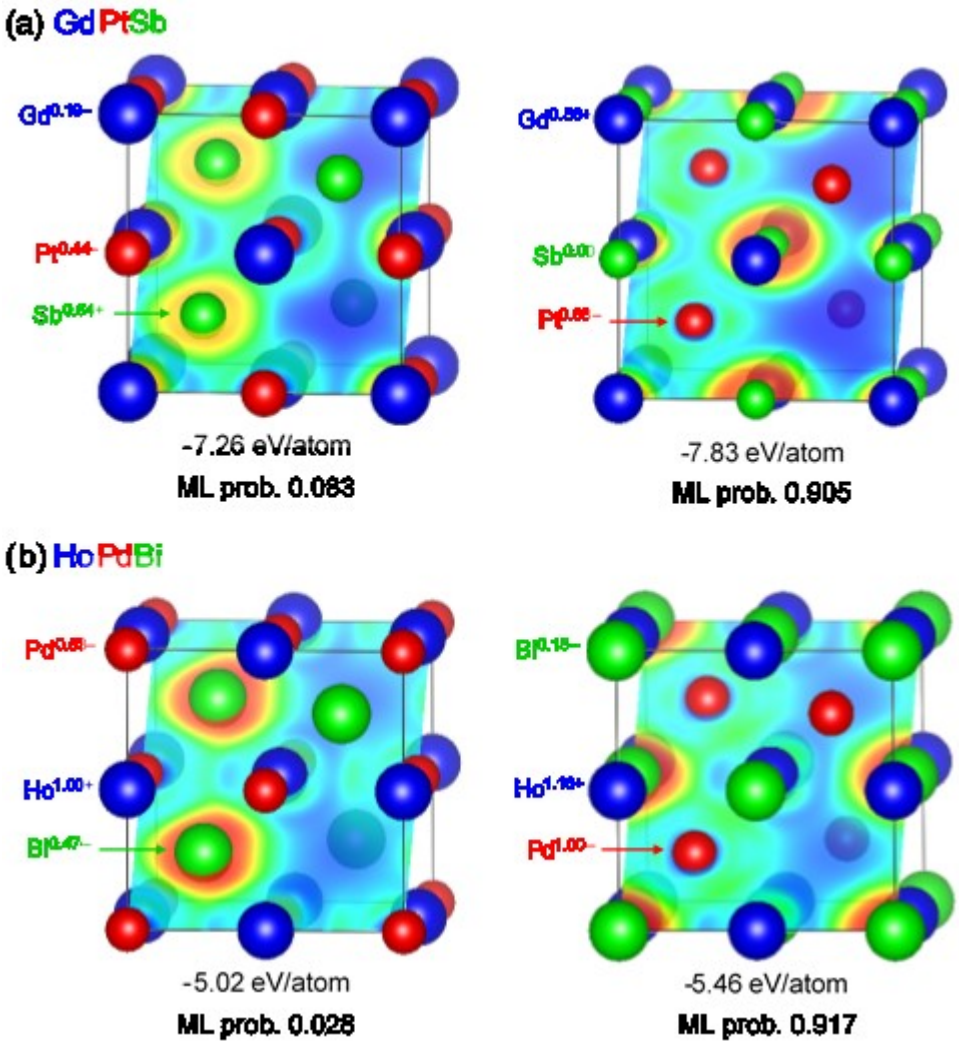


Figure 2-8. Bader charge analysis for (a) GdPtSb and (b) HoPdBi in alternative site distributions, with total energies calculated from first principles and machine-learning probabilities indicated.

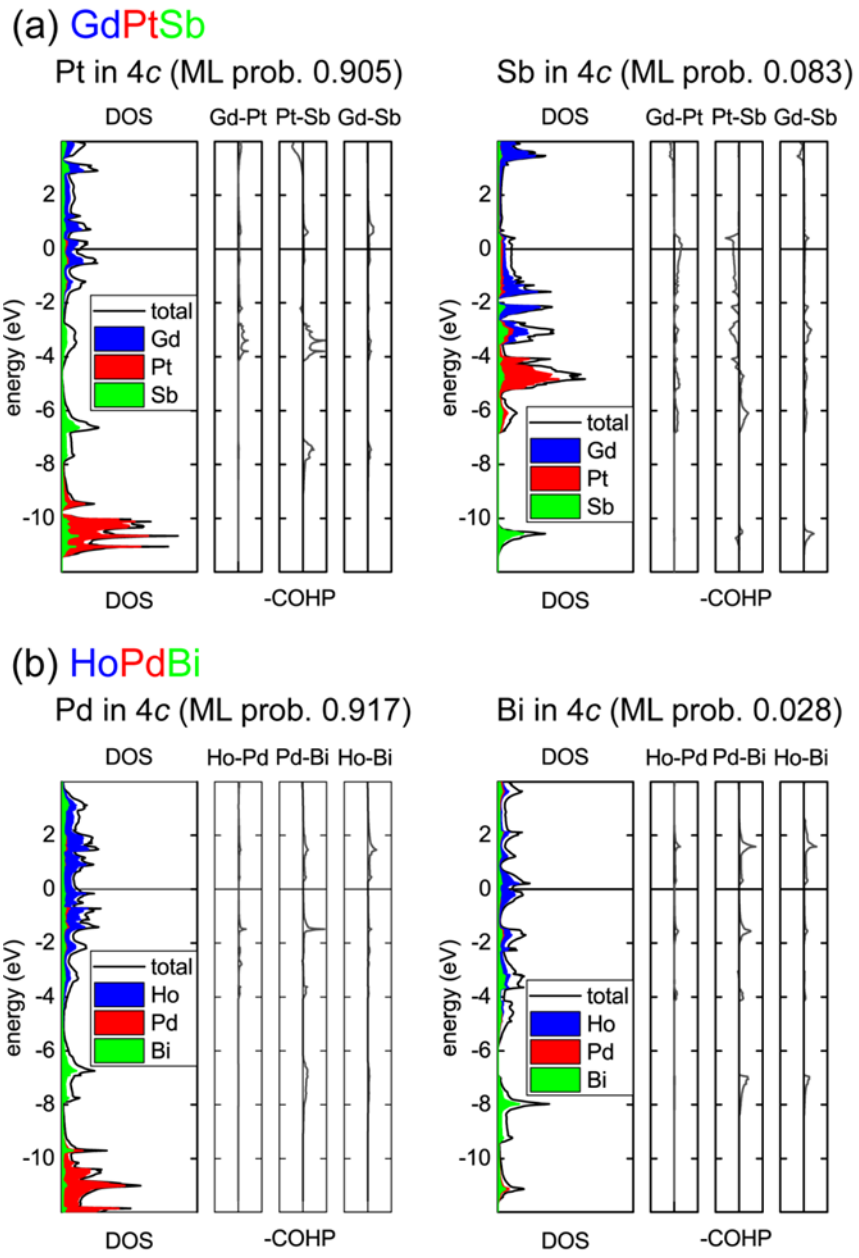


Figure 2-9. DOS and –COHP curves for (a) GdPtSb and (b) HoPdBi in alternative site distributions, with machine-learning probabilities indicated.

2.4. Conclusions

A machine-learning model has been developed to predict the correct site distribution in half-Heusler compounds based on features that depend on the composition alone. The 95%

accuracy of this model is a marked improvement over simple heuristic guidelines based on electronegativity differences that have been historically used to rationalize site distributions in these compounds. Because very few experimental structure determinations have actually been carried out on these compounds, it was not clear at the outset if the data used to build the machine-learning model can be assumed to be completely reliable. Accepting these data without thought risks compromising the quality of the model, but for a high-throughput machine-learning approach to work, it should not be necessary to inspect individual data points and painstakingly correct erroneous entries. For this reason, data cleansing is an integral part of a machine-learning workflow, especially for materials informatics where data are much scarcer (at most 10^2–10^3 entries) compared to other popular informatics applications where data are more abundant (>10^6 entries).[51] It is important to ensure that samples are not misclassified simply because the model has not yet been optimized. The CR-FS procedure is a crucial step in ensuring that the best quality model has been obtained. An interesting insight from implementing this procedure was the realization that size factors (including radii that pertain to metallic character), in addition to electronegativities, strongly influence the site distribution. Then, remaining misclassified samples are prime suspects to be re-evaluated experimentally, as we have demonstrated with GdPtSb and HoPdBi which were characterized by powder and single-crystal X-ray diffraction methods. Although combining machine-learning with first-principles approaches can be very powerful, we emphasize, as have others, that the ultimate test of any predictive model for materials discovery should be experimental validation. The model (available at https://github.com/Mar-group/Half-Heusler-site-occupancy-prediction) could serve as a useful guide to predict site occupancies of heretofore unknown half-Heusler compounds.

2.5. References

(1) Huheey, J. E.; Keiter, E. A.; Keiter, R. L. *Inorganic Chemistry: Principles of Structure and Reactivity*, 4th Ed.; Harper Collins: New York, **1993**.

(2) Burdett, J. K.; Lee, S. L.; McLarnan, T. J. The coloring problem. *J. Am. Chem. Soc.* **1985**, *107*, 3083–3089.

(3) Miller, G. J. The "coloring problem" in solids: How it affects structure, composition and properties. *Eur. J. Inorg. Chem.* **1998**, 523–536.

(4) Graf, T.; Felser, C.; Parkin, S. S. P. Simple rules for the understanding of Heusler compounds. *Prog. Solid State Chem.* **2011**, *39*, 1–50.

(5) Casper, F.; Graf, T.; Chadov, S.; Balke, B.; Felser, C. Half-Heusler compounds: novel materials for energy and spintronic applications. *Semicond. Sci. Technol.* **2012**, *27*, 063001-1–063001-8.

(6) Bos, J.-W. G.; Downie, R. A. Half-Heusler thermoelectrics: a complex class of materials. *J. Phys.: Condens. Matter* **2014**, *26*, 433201-1–433201-15.

(7) Palmstrøm, C. J. Heusler compounds and spintronics. *Prog. Cryst. Growth Charact. Mater.* **2016**, *62*, 371–397.

(8) Zeier, W. G.; Schmitt, J.; Hautier, G.; Aydemir, U.; Gibbs, Z. M.; Felser, C.; Snyder, G. J. Engineering half-Heusler thermoelectric materials using Zintl chemistry. *Nat. Rev. Mater.* **2016**, *1*, 16032-1–16032-10.

(9) Wollmann, L.; Nayak, A. K.; Parkin, S. S. P.; Felser, C. Heusler 4.0: Tunable materials. *Annu. Rev. Mater. Res.* **2017**, *47*, 247–270.

(10) Nowotny, H.; Sibert, W. Ternäre Valenzverbindungen in den Systemen Kupfer (Silber)–Arsen (Antimon, Wismut)–Magnesium. *Z. Metallkd.* **1941**, *33*, 391–394.

(11) In the literature, the 4*a* and 4*b* sites are sometimes described as being octahedral, and the 4*c* site as tetrahedral. The reader should be especially alert to understand that these geometrical descriptions are being made *in reference to the derivation from the NaCl- and ZnS-type structures*. Within the NaCl-type structure (space group $Fm\bar{3}m$), only the 4*a* and 4*b* sites are occupied, each having octahedral coordination; within the ZnS-type structure (space group $F4\bar{3}m$), only the 4*a* and 4*c* sites are occupied, each having tetrahedral coordination. When all three sites are occupied, of course, the coordination geometries are no longer preserved. The half-Heusler and related (full)-Heusler structures are also sometimes described in terms of "interpenetrating fcc lattices (or 'sublattices')," but such terminology is perhaps best avoided because it conflates the precise definition of "lattice" as established in crystallography with an extended meaning that is not intended to be conveyed. (See: Nespolo, M. Lattice versus structure, dimensionality versus periodicity: a crystallographic Babel? *J. Appl. Crystallogr.* **2019**, *52*, doi:10.1107/S1600576719000463.)

(12) Equivalently, the origin of the unit cell has merely been shifted and there is no effect on the XRD pattern.

(13) *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD)*, release 2015/16; ASM International: Materials Park, OH, 2016.

(14) Gordy, W.; Thomas, W. J. O. Electronegativities of the elements. *J. Chem. Phys.* **1956**, *24*, 439–444.

(15) Allred, A. L.; Rochow, E. G. A scale of electronegativity based on electrostatic force. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264–268.

(16) Bende, D.; Grin, Y.; Wagner, F. R. Covalence and ionicity in MgAgAs-type compounds. *Chem. Eur. J.* **2014**, *20*, 9702–9708.

(17) Bende, D.; Wagner, F. R.; Grin, Y. 8–*N* rule and chemical bonding in main-group MgAgAs-type compounds. *Inorg. Chem.* **2015**, *54*, 3970–3978.

(18) White, M. A.; Medina-Gonzalez, A. M.; Vela, J. Soft chemistry, coloring and polytypism in filled tetrahedral semiconductors: Toward enhanced thermoelectric and battery materials. *Chem. Eur. J.* **2018**, *24*, 3650–3658.

(19) Hames, F. A. Ferromagnetic-alloy phases near the compositions Ni_2MnIn, Ni_2MnGa, Co_2MnGa, Pd_2MnSb, and PdMnSb. *J. Appl. Phys.* **1960**, *31*, S370–S371.

(20) Endo, K. Magnetic studies of $C1_b$-compounds CuMnSb, PdMnSb and Cu_{1-x} (Ni or Pd)$_{xMnSb}$. *J. Phys. Soc. Jpn.* **1970**, *29*, 643–649.

(21) Webster, P. J.; Ziebeck, K. R. A. Structures of $Pd_{2-x}MnSb$ – an improved neutron polarizer? *J. Magn. Magn. Mater.* **1980**, *15–18*, 473–474.

(22) Buschow, K. H. J.; van Engen, P. G.; Jongebreur, R. Magneto-optical properties of metallic ferromagnetic materials. *J. Magn. Magn. Mater.* **1983**, *38*, 1–22.

(23) Larson, P.; Mahanti, S. D.; Kanatzidis, M. G. Structural stability of Ni-containing half-Heusler compounds. *Phys. Rev. B* **2000**, *62*, 12754–12762.

(24) Schuster, H.-U.; Dietsch, W. Eine neue ternäre Phase im System Li–Au–Sb. *Z. Naturforsch. B* **1975**, *30*, 133.

(25) Zhang, X.; Yu, L.; Zakutayev, A.; Zunger, A. Sorting stable versus unstable hypothetical compounds: The case of multi-functional ABX half-Heusler filled tetrahedral structures. *Adv. Funct. Mater.* **2012**, *22*, 1425–1435.

(26) Carrete, J.; Li, W.; Mingo, N.; Wang, S.; Curtarolo, S. Finding unprecedently low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **2014**, *4*, 011019-1–011019-9.

(27) Gauthier, R.; Zhang, X.; Hu, L.; Yu, L.; Lin, Y.; Sunde, T. O. L.; Chon, D.; Poeppelmeier, K. R.; Zunger, A. Prediction and accelerated laboratory discovery of previously known 18-electron ABX compounds. *Nat. Chem.* **2015**, *7*, 308–316.

(28) Ma, J.; Hegde, V. I.; Munira, K.; Xie, Y.; Keshavarz, S.; Mildebrath, D. T.; Wolverton, C.; Ghosh, A. W.; Butler, W. H. Computational investigation of half-Heusler compounds for spintronic applications. *Phys. Rev. B* **2017**, *95*, 024411-1–024411-25.

(29) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

(30) Oliynyk, A. O.; Mar, A. Discovery of intermetallic compounds from traditional to machine-learning approaches. *Acc. Chem. Res.* **2018**, *51*, 59–68.

(31) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G. K. H.; Mingo, N. Materials screening for the discovery of new half-Heuslers: Machine learning versus ab initio methods. *J. Phys. Chem. B* **2018**, *122*, 625–632.

(32) *PLS_Toolbox*, version 8.0.1; Eigenvector Research Inc.: Wenatchee, WA, 2018.

(33) *MATLAB Statistics and Machine Learning Toolbox*, release 2018a; The Mathworks Inc.: Natick, MA, 2018.

(34) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

(35) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

(36) Coelho, A. A. *TOPAS-Academic*, version 6; Coelho Software: Brisbane, Australia, 2007.

(37) Sheldrick, G. M. *SHELXTL*, version 6.12; Bruker AXS Inc.: Madison, WI, 2001.

(38) Kresse, G.; Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(39) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(40) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(41) Grin, Y.; Savin, A.; Silvi, B. The ELF perspective of chemical bonding. In *The Chemical Bond: Fundamental Aspects of Chemical Bonding*; Frenking, G., Shaik, S., Eds.; Wiley-VCH: Weinheim, 2014; Chap. 10, pp 345–382.

(42) Tang, W.; Sanville, E.; Henkelmann, G. A grid-based Bader analysis algorithm without lattice basis. *J. Phys. Condens. Matter* **2009**, *21*, 084204-1–084204-7.

(43) Dronskowski, R.; Blöchl, P. E. Crystal orbital Hamilton populations (COHP). Energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *J. Phys. Chem.* **1993**, *97*, 8617–8624.

(44) Tank, R.; Jepsen, O.; Burkhardt, A.; Andersen, O. K. *TB-LMTO-ASA Program*, version 4.7; Max Planck Institut für Festkörperforschung: Stuttgart, Germany, 1998.

(45) Oliynyk, A. O.; Adutwum, L. A.; Harynuk, J. A.; Mar, A. Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis. *Chem. Mater.* **2016**, *28*, 6672–6681.

(46) Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Sotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.; Brgoch, J. Disentangling structural confusion through machine

learning: Structure prediction and polymorphism of equiatomic ternary phases *ABC*. *J. Am. Chem. Soc.* **2017**, *139*, 17870–17881.

(47) The VIP score depends on the square of the normalized weight of a variable in a PLS-DA model, and is interpreted as the proportion of variances in class assignments that can be explained by that variable.

(48) *Atomic radii* are obtained from first-principles calculations and correspond to the radius of maximum charge density in the outermost shell (Suresh, C. H. A consistent approach toward atomic radii. *J. Phys. Chem. A* **2001**, *105*, 5940–5944). *Covalent radii* are obtained by halving observed homoatomic distances in crystal structures (Pauling, L.; Huggins, M. L. Covalent radii of atoms and interatomic distances in crystals containing electron-pair bonds. *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **1934**, *87*, 205–238). *Ionic and crystal radii* are obtained from empirically measured bond distances to atoms, possibly in different coordination environments, within crystal structures, mostly of oxides and halides (Pauling, L. The sizes of ions and the structure of ionic crystals. *J. Am. Chem. Soc.* **1927**, *49*, 765–790. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr., Sect. A* **1976**, *32*, 751–767). *Miracle radii* are obtained from interatomic distances in metallic glasses (Miracle, D. B. The efficient cluster packing model – An atomic structural model for metallic glasses. *Acta Mater.* **2006**, *54*, 4317–4336). *Zunger pseudopotential radii* are calculated by taking the sum of self-consistently screened nonlocal atomic pseudopotentials of the s- and p-orbitals (Zunger, A. Systematization of the stable crystal structure of all *AB*-type binary compounds: A pseudopotential orbital-radii approach. *Phys. Rev. B* **1980**, *22*, 5839–5872).

(49) Allred, A. L. Electronegativity values from thermochemical data. *J. Inorg. Nucl. Chem.* **1961**, *17*, 215–221.

(50) Little, E. J. Jr.; Jones, M. M. A complete table of electronegativities. *J. Chem. Ed.* **1960**, *37*, 231–233.

(51) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bull.* **2016**, *41*, 399–409.

Chapter 3 Predicting New Half-Heusler Compounds

3.1. Introduction

The previous chapter discussed how machine learning can be used to predict the site distributions ("colouring problem") in half-Heusler compounds. This chapter focuses on a different question, namely, will a compound having composition *ABC* adopt the half-Heusler structure? [1] In the course of this investigation, it became important to consider the close relationship between the half-Heusler and full-Heusler structures.

The family of Heusler compounds spans at least four structure types termed full-Heusler (Cu_2MnAl-type), half-Heusler (MgAgAs-type), inverse Heusler ($CuHg_2Ti$-type), and quaternary Heusler (LiMgPdSn-type). That is, they can be formed among both ternary and quaternary phases. The combination of elements is so vast (>130 000) that exploratory synthesis to discover new Heusler compounds is a difficult task. To date, there have been 1371 Heusler compounds that have been reported, 46% of which (637 compounds) have been structurally characterized by X-ray diffraction, and 5% of which (78 compounds) have had detailed atomic assignments.[2] Most of the experimental characterization has tended to be performed by powder X-ray diffraction, which can be problematic because of ambiguities as discussed previously.[3,4] For example, different structural models can lead to nearly identical intensities in the simulated powder XRD patterns. In some cases, experimental observations disagree with structures proposed by computations (e.g., NiMnCuSb), inviting skepticism about inflated claims that high-throughput calculations can guide the search for new compounds.[5-10] In fact, computational approaches have a long way to go to fulfill their promise to advance materials discovery, because they often neglect

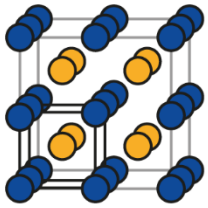
real problems faced by experimentalists, such as the occurrence of competing phases, the numerous permutations possible for site occupation, and the possibility that crystal structures may be wrong or inconsistently reported.[2,5,11]

Various rules have been formulated previously that attempt to rationalize the formation of Heusler compounds and their site occupancies.[12-14] On the basis of electronegativity differences, the structure is split up into two substructures. Vacancies are introduced if applicable so that a precise valance electron count is attained, in accordance with the 8- or 18-electron rule. Further constraints are placed on the position of the light atoms (such as Li), if they are present. Unfortunately, these rules are fallible, with frequent violations of the electron count and ambiguities in atom assignments arising from the use of different electronegativity scales. As Legrain et al. showed, even first-principles studies that assess the stability of half-Heusler phases can give inconsistent results.[15] When they compared results from first-principles and machine-learning approaches, they found the best agreement among compounds that satisfy the 8- or 18-electron rule, but suggested that additional factors (e.g., configurational entropies, quasiharmonic contributions) come into play to account for differences. Experimental validation is needed to clarify the limitations of the machine-learning approach. The challenge is that although a compound may be predicted to exist, there is no guidance on the conditions under which it forms; in fact, it might be metastable. In particular, the structures of real compounds often exhibit site disorder.

Full-Heusler compounds comprise 63% (862 compounds) of the Heusler family. They crystallize in the Cu_2MnAl-type structure (space group $Fm\bar{3}m$), which is a doubled superstructure of the CsCl-type structure (space group $Pm\bar{3}m$) (Figure 1). That is, on going from CsCl to Cu_2MnAl, the symmetry is reduced by a factor of 2 ("a *klassengleiche* transformation of index 2")

and the number of sites increases from two to three, at Wyckoff positions 8*c* (1/4, 1/4, 1/4), 4*b* (1/2, 1/2, 1/2), and 4*a* (0, 0, 0). Half-Heusler compounds comprise 27% (365 compounds) of the Heusler family. They crystallize in the MgAgAs-type structure (space group $F\bar{4}3m$), which is derived from the Cu_2MnAl-type structure by a further reduction in symmetry by a factor of 2 ("a *translationengleiche* transformation of index 2") and the number of sites increases from three to four, at Wyckoff positions 4*a* (0, 0, 0), 4*b* (1/2, 1/2, 1/2), 4*c* (1/4, 1/4, 1/4), and 4*d* (3/4, 3/4, 3/4). The local symmetry of the 4*a* and 4*b* sites is reduced from $m\bar{3}m$ to $\bar{4}3m$. The 8*c* site in Cu_2MnAl-type structure is split into the 4*c* and 4*d* sites in the MgAgAs-type structure, of which only one set is occupied while the other set is vacant. The "half-occupancy" of the formerly 8*c* site thus gives rise to the name "half-Heusler."

CsCl ($Pm\bar{3}m$)

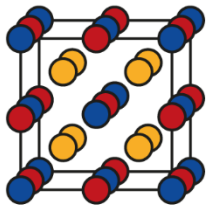


1a	1b
m-3m	*m-3m*
0	0.5
0	0.5
0	0.5

*t*1 | *k*2

full-Heusler

Cu_2MnAl ($Fm\bar{3}m$)

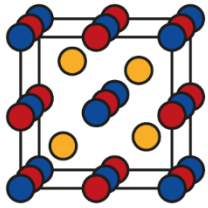


4a	4b	8c
m-3m	*m-3m*	-43*m*
0	0.5	0.25
0	0.5	0.25
0	0.5	0.25

*t*2, *k*1 | *vac.*

half-Heusler

MgAgAs ($F\bar{4}3m$)



4a	4b	4c
-43*m*	-43*m*	-43*m*
0	0.5	0.25
0	0.5	0.25
0	0.5	0.25

Figure 3-1. Relationships between CsCl, full-Heusler, and half-Heusler structures.

There exist 27 instances in which full-Heusler (AB_2C) and half-Heusler compounds (ABC) are formed from the same set of elements *A*, *B*, and *C* (Figure 2). This is an interesting phenomenon that suggests that intermediate structures may be possible, but the chemical space has not yet been extensively explored. The 8- or 18-electron rule is violated for some of these half-Heusler compounds, and more than half of the cases involve group-10 elements (Ni, Pd, Pt). In

fact, roughly 40% of all half-Heusler compounds contain Ni (Figure 3). The 18-electron rule can be satisfied by combining Ni (10 e^-) with lanthanide elements (3 e^-) and a group-15 element such as Sb (5 e^-), which is the most frequently encountered element in half-Heusler compounds. Another common combination of elements is Ni (10 e^-), group-4 elements (4 e^-), and Sn (4 e^-), which also satisfies the 18-electron rule.

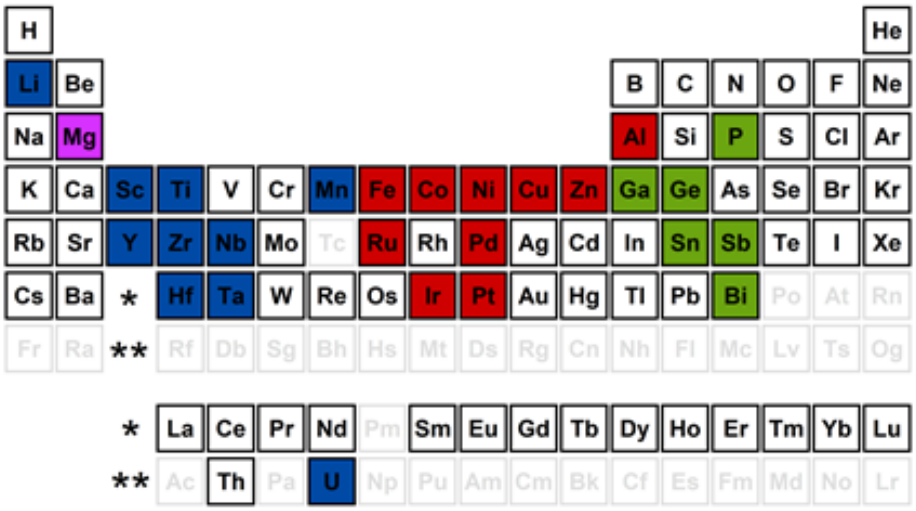


Figure 3-2. Compositional overlap between full- and half-Heusler compounds.

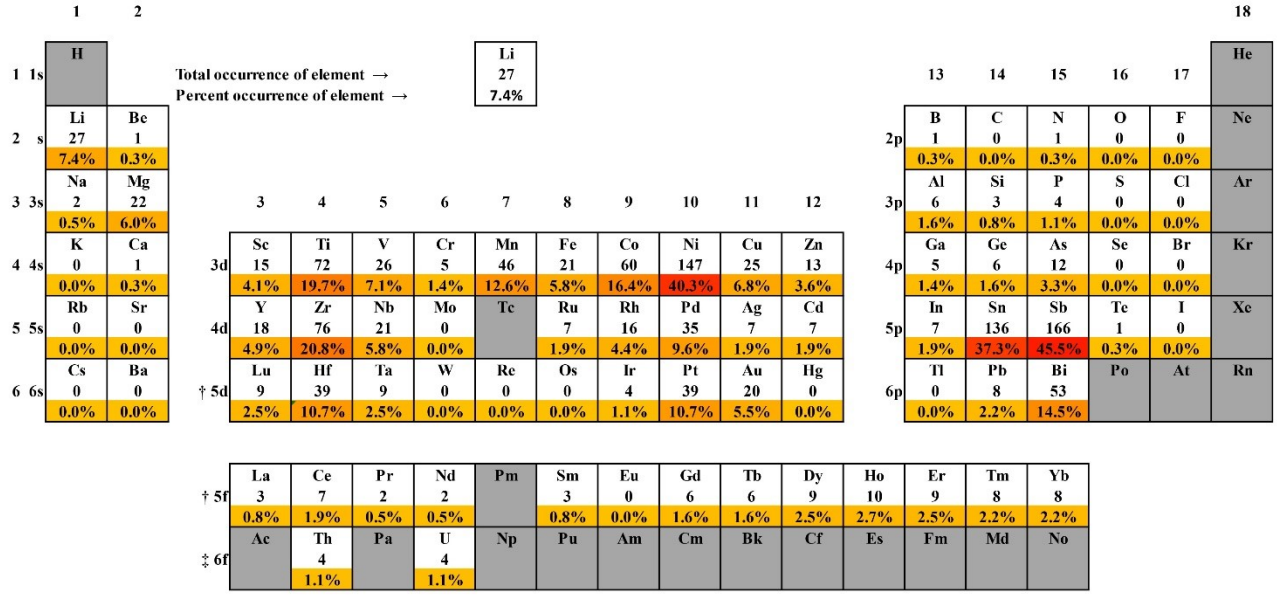


Figure 3-3. Occurrence of elements in half-Heusler structure type.

If compounds could be prepared having compositions that are intermediate between full- and half-Heusler compounds, it would be possible to correlate how structures and physical properties evolve with site occupation. Introducing site deficiencies and doping would then be a powerful tool to tune materials properties in applications such as magnetic, magnetocaloric, and thermoelectric applications. For example, varying the composition between full- and half-Heusler compounds has been shown to improve magnetocaloric response and increase saturation moments in $MnNi_{1+x}Sb$.[16,17] Complex order-disorder transitions take place in this intermediate region.[18,19] Doping in $Ti_xNb_{1-x}CoSn$ introduces transitions from a non-magnetic semiconductor to a ferromagnetic metal.[20] Self-doping with Ni in ZrNiSn improves thermoelectric properties by reducing its thermal conductivity by over 60%.[21]

In the current study, the primary goal is to discover new compounds with the half-Heusler structure through predictions made from a machine-learning approach. A secondary goal is to

identify half-Heusler compounds with compositions that have a full-Heusler counterpart, that is, the formation of a solid solution $AB_{2-x}C$ (x = 0–1).

3.2. Experimental

3.2.1. Machine-Learning Model

A machine-learning model was developed to discover new half-Heusler compounds *ABC*. The task was formulated as a classification problem. Crystallographic data were extracted for all compounds *ABC* in Pearson's Crystal Data, including those that deviated from the ideal composition by up to 20%, subject to the constraints that they do not contain hydrogen, noble gases, or elements with $Z > 83$, except that U- and Th-containing compounds were included. In total, the data set contained 2818 such compounds, of which 180 adopt the MgAgAs-type structure.[2] Descriptors were derived from 55 properties of each element; the elemental properties were then combined through 21 arithmetic operations that were weighted by composition, giving a total of 1155 descriptors at the outset. A machine-learning pipeline was developed using the PLS Toolbox software (Version 8.0.1) implemented through MATLAB (2018a release).[22,23] Two-thirds of the data were assigned to a training set and one-third to a validation set. The data were preprocessed by autoscaling and normalization. The autoscaling procedure involved mean-centering of the feature columns, followed by division by their standard deviation to obtain unit variance. The normalization procedure involved division by the sum of the absolute values of each row.

The PLS Toolbox software contains three different machine-learning algorithms (KNN, SVM, and PLS-DA) which outputs prediction probability which where combined using soft-voting script written in MATLAB. Feature selection was carried out in two ways to create more models

to consider (Figure 4). Cluster-resolution feature selection (CR-FS) and genetic algorithms (GA) were applied to generate six models.[24,25] Three iterations with 100 rounds of CR-FS were performed, giving 300 sets of features. The best performing set contained 230 features which had the highest survival rate among the 300 models. One iteration of the GA was ran using a population size of 256, which corresponds to the largest population setting to increase the number of possibilities to perform crossover and to improve accuracy. The maximum number of generations was set to 200, sufficient for the algorithm to converge on a solution. The mutation rate was set to 0.005 to address under- or overrepresentation of features in populations, with double crossover used as the breeding strategy. Partial least squares was used as the regression method to evaluate chromosomes in the population, with 25 latent variables and 10-fold-split random cross validation being used. The fittest model contained 225 features. These two sets of features were then applied with three types of algorithms. The training data set was augmented using the synthetic majority oversampling technique (SMOTE) (Figure 5).[26]

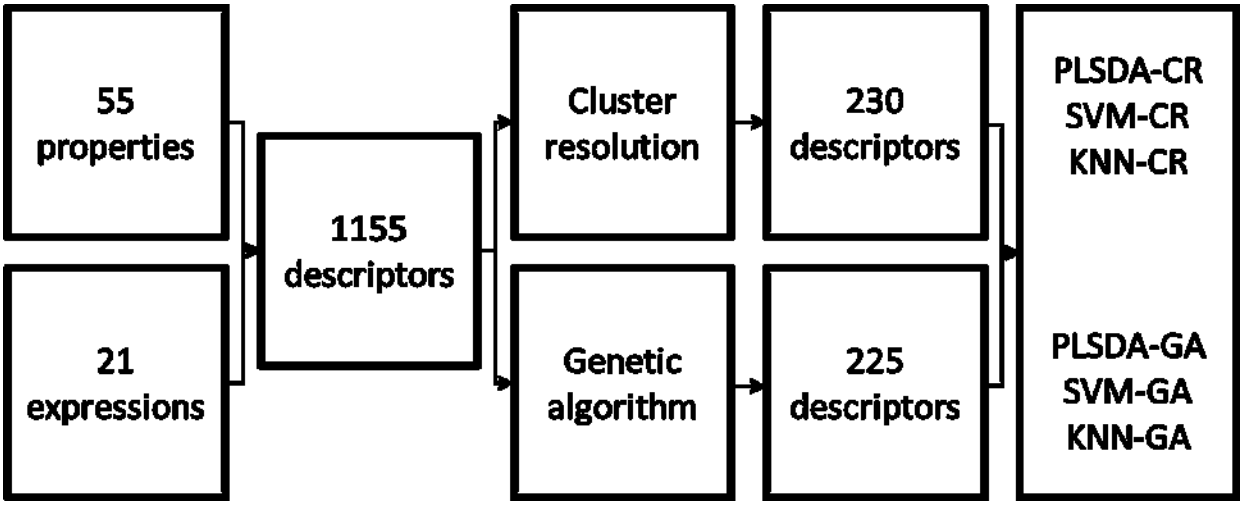


Figure 3-4. Machine-learning workflow.

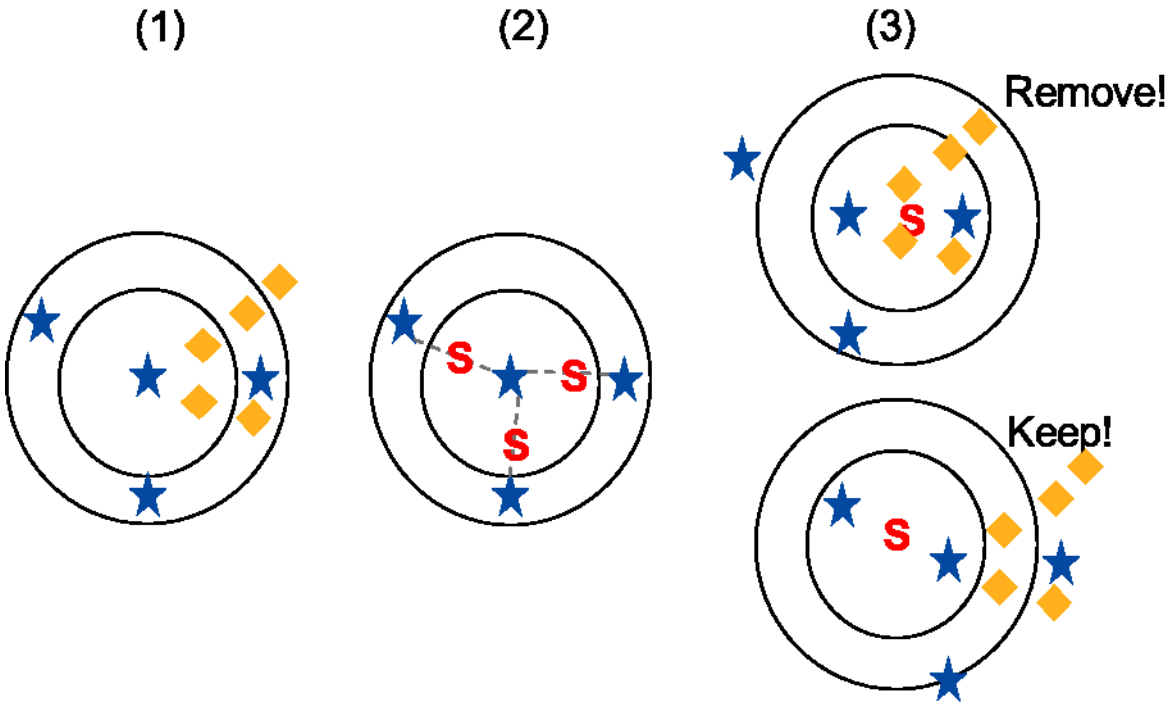


Figure 3-5. The process of SMOTE: (1) initial unbalanced dataset, (2) generation of synthetic samples using the minority class, and (3) evaluation of synthetic samples in the dataset using KNN.

The SMOTE algorithm is used to improve the individual performance of each model by increasing the number of samples in the minority class. The *k*-nearest neighbours in the minority class (half-Heusler compounds) are first calculated, and then between each neighbour, a synthetic sample is placed which is intended to be as similar as possible to the minority samples. The synthetic samples are added to the data set, and the KNN procedure is performed on the whole data set to find the nearest neighbours of the synthetic samples. If 60% of the nearest neighbours of the synthetic samples belong to the minority class, then they are kept in the data set; otherwise, they are discarded. For the training set, a total of 53 synthetic samples were generated for the model using GA, and 631 for the model using CR-FS. The models were then combined to create an ensemble (Figure 6). The votes were combined through soft voting, which is based on taking

the average of the prediction probabilities. If the average probability is greater than 50%, then the compound is assigned a label of "half-Heusler."

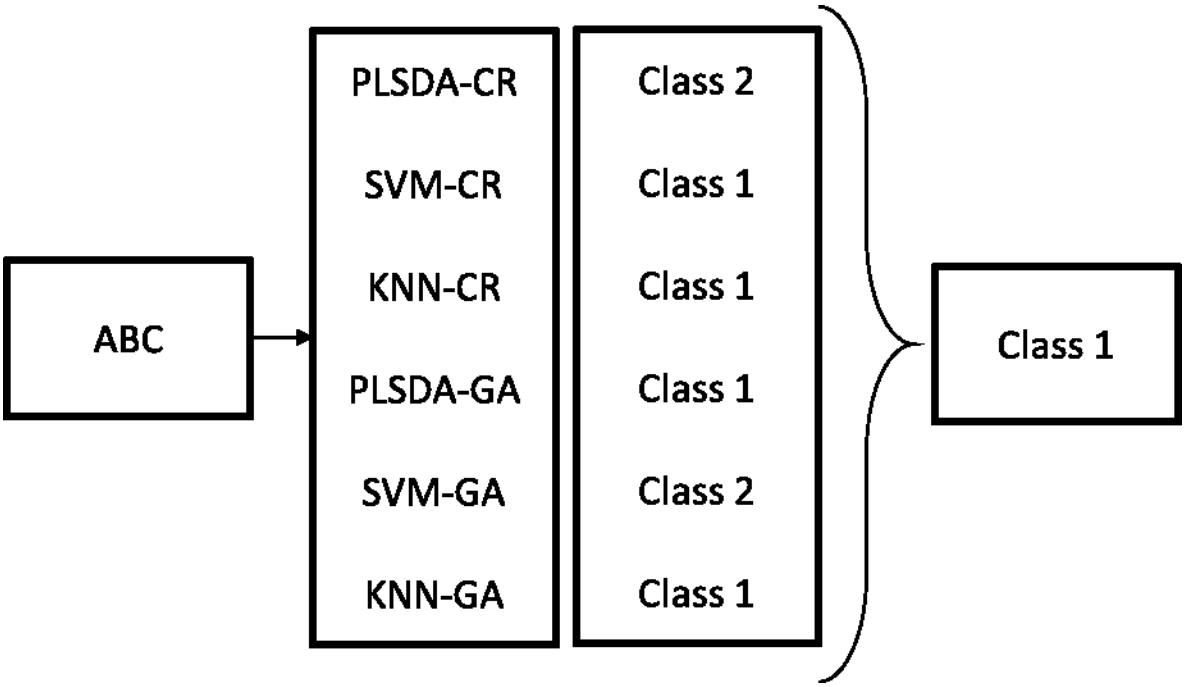


Figure 3-6. Ensemble of models used for classification.

3.3. Results and Discussion

3.3.1. Machine learning

A machine-learning model has been built to predict new compounds with half-Heusler structures, by training on 2818 compounds whose compositions were within 20% of the idealized one *ABC* found in Pearson's Crystal Data.[2] The data set consists of two classes, consisting of compounds that have half-Heusler structures (180) and those that do not (2638). The data were sorted in a similar way as described in the previous chapter for determining site occupancies in half-Heusler structures (e.g., *CBA* to *ABC*) to reflect the most probable site occupancies because

the way descriptors are defined is not order-invariant.[27] Descriptors were generated based on elemental properties (Table 1) and arithmetic operations (Table 2), with the assumption that these will reproduce previously developed rules for half-Heusler structures.

Table 3-1. Elemental properties used to generate features.

1. atomic number	29. number of f valence electrons
2. atomic weight	30. number of unfilled s orbitals
3. period number	31. number of unfilled p orbitals
4. group number	32. number of unfilled d orbitals
5. family number	33. number of unfilled f orbitals
6. metal / metalloid / nonmetal	34. number of outer-shell electrons
7. Mendeleev number	35. first ionization energy
8. quantum number *l*	36. polarizability
9. atomic radius	37. melting point
10. Miracle radius	38. boiling point
11. covalent radius	39. density
12. ionic radius	40. specific heat
13. effective ionic radius	41. heat of fusion
14. Zunger pseudopotential radii sum	42. heat of vaporization
15. CSD covalent radius	43. thermal conductivity
16. Slater radius	44. heat of atomization
17. crystal radius	45. cohesive energy
18. Pauling electronegativity	46. bulk modulus
19. Martynov-Batsanov electronegativity	47. 1st Bohr radius
20. Gordy electronegativity	48. Effective nuclear radii
21. Mulliken electronegativity	49. electron affinity
22. Allred-Rochow electronegativity	50. chemical hardness
23. metallic valence	51. NIST total energy (LDA)
24. number of valence electrons	52. NIST kinetic energy (LDA)
25. Gilman number of valence electrons	53. NIST coulomb energy (LDA)

26. number of s valence electrons	54. NIST electron-nucleus energy(LDA)
27. number of p valence electrons	55. NIST exchange energy (LDA)
28. number of d valence electrons	

Table 3-2. Arithmetic expression used to generate features.

largest number – smallest number
n_a*X_a
n_b*X_b
n_c*X_c
$2*(n_a*X_a+ n_b*X_b)$
$2*(n_a*X_a+ n_c*X_c)$
$2*(n_c*X_c+ n_b*X_b)$
$2*(n_a*X_a)+ 2*(n_b*X_b)+ 4*(n_c*X_c)$
$2*(n_c*X_c)+ 2*(n_b*X_b)+ 4*(n_a*X_a)$
$2*(n_a*X_a)+ 2*(n_c*X_c)+ 4*(n_b*X_b)$
$((n_a*X_a)+ (n_c*X_c)+ (n_b*X_b))/3$
$((4*(n_b*X_b)+4*(n_a*X_a))/sqrt(2))^3$
$((4*(n_b*X_b)+4*(n_c*X_c))/sqrt(2))^3$
$((4*(n_c*X_c)+4*(n_a*X_a))/sqrt(2))^3$
(smallest number/largest number) – 0.225
(smallest number/largest number) – 0.414
(smallest number/largest number) – 0.732
(smallest number/largest number) – 1
$(1-exp(-(((n_b*X_b- n_c*X_c)/2)^2))$
$(1-exp(-(((n_b*X_b- n_a*X_a))/2)^2))$
$(1-exp(-(((n_a*X_a- n_c*X_c))/2)^2))$

An ensemble approach was used to overcome the reliance on a single model and to cover chemical space more broadly. The idea is to avoid misclassifications and to improve prediction probabilities, with the assumption that multiple models describe different areas of chemical space and the majority decision is better than an individual one. For example, if a single model gives a prediction probability of 90% to a compound *ABC* having a half-Heusler structure, and two other

models give lower probabilities of 25%, the average probability would become 47%, thus changing the classification.

Table 3-3. Comparison of model performance.

Model	Training set	Validation set
Before feature selection (ensemble)		
sensitivity (Class 1 / Class 2)	0.858 / 0.976	0.733 / 0.971
specificity (Class 1 / Class 2)	0.976 / 0.858	0.971 / 0.733
accuracy	0.969	0.964
After feature selection (ensemble)		
sensitivity (Class 1 / Class 2)	0.900 / 0.989	0.867 / 0.983
specificity (Class 1 / Class 2)	0.989 / 0.900	0.983 / 0.867
accuracy	0.978	0.975
After SMOTE (ensemble)		
sensitivity (Class 1 / Class 2)	0.950 / 0.986	0.900 / 0.982
specificity (Class 1 / Class 2)	0.986 / 0.950	0.982 / 0.900
accuracy	0.981	0.977
After SMOTE (best individual)		
sensitivity (Class 1 / Class 2)	0.954 / 0.997	0.816 / 0.985
specificity (Class 1 / Class 2)	0.997 / 0.954	0.985 / 0.816
accuracy	0.993	0.974

The results indicate that the ensemble approach is more generalizable than with a single model (Table 3), with an improved performance in the validation set. Specifically, the sensitivity (rate of true positives) is an important metric because we are interested in predicting new half-Heusler compounds. The ensemble method outperforms the single-model method by 9% in this regard. The sensitivity and accuracy increase after feature selection (to avoid overfitting data) and SMOTE are applied (to give the model more class-1 samples from which to learn), demonstrating

that these methods were indeed helpful in improving the model. There is a large discrepancy between the two classes, which decreases the sensitivity relative to class 1, but this discrepancy was softened by adding 54 (using GA) or 631 synthetic samples (using CR-FS) to the training sets. The huge difference in the number of synthetic samples for each feature set can be attributed to how well they cluster classes into homogeneous neighbourhoods. It can be concluded the CR-FS was better, but the comparison should be tempered by the fact they contain different numbers of features. It should be pointed out that the synthetic samples were only generated from the training set, to avoid bias in which samples generated from the validation set appear in the training set. In a previous comparison of first-principles vs machine-learning methods to the discovery of half-Heusler compounds, Legrain et al. developed an ensemble model using the random forest algorithm (with 1000 trees). Legrain's model had a precision of 0.90, a recall of 0.52, and a Matthew's correlation coefficient of 0.68; in comparison, our model had a precision of 0.77, a recall of 0.90, and a Matthew's correlation coefficient of 0.82.[15] Our model performs better balanced (MCC) statistical measures and has a significantly lower amount of false negatives but a higher number of false positives. So, our model is a bit greedier in terms of our half Heusler predictions but overall exhibits better performance. Our model also diverges a bit when it comes to the prediction for the formation of half-Heusler compounds. We also focus in on the predictions for the existence of half-Heusler compounds that have counterparts among full-Heusler compounds (Table 4), in addition to identifying the most probable candidates to form half-Heusler structures (Table 5).

Table 3-4. Full-Heusler counterpart probability

sample	probability	Sample	probability
MnRhPb	0.940318	LiMgTl	0.699684
SbUPd	0.937233	LiPdSn	0.699185
MnSnPd	0.923915	LiCdSn	0.69472
MnRhSn	0.904742	MnSnCo	0.691028
LiCdGe	0.886437	MnRuSn	0.683711
MnPdIn	0.885059	LiPdGe	0.677328
MnSnNi	0.883066	LiGeCo	0.671179
MgSnNi	0.870451	LiGaRu	0.667919
LiMgPb	0.860629	MnRhAl	0.650874
MnSnCu	0.852327	TiPbLi	0.64874
NbSnNi	0.850677	VSnCo	0.646496
ZrSbNi	0.847263	LiMgIn	0.644947
LiMgGe	0.83755	LiNiSn	0.643281
VSnNi	0.816952	LiMgCd	0.641821
LiZnGe	0.806029	MnPtAl	0.639986
LiHgGe	0.801488	ScSnNi	0.635793
PdPbLi	0.799741	VSnRh	0.634538
TiSnPd	0.792943	MnRhGe	0.626066
LiSbCo	0.771848	CrSnRu	0.621542
LiMgGa	0.766507	LiPdGa	0.615888
LiMgSn	0.765717	VSnRu	0.613479
MgInPd	0.760959	MnPdGe	0.601354
ZrSnCo	0.753862	VPbLi	0.58175
LiMgSi	0.751294	LiNiSi	0.576855
LiSbSn	0.748943	TiSnIr	0.57591
TiSbCu	0.742239	SnIrLi	0.575458
YBiPd	0.73048	LiGaRh	0.571219
LiZnSn	0.730323	LiSnPt	0.568707
TiBiLi	0.729338	MgPdGa	0.548588
LiAlPd	0.717202	LiAlPt	0.544319
MnRuSb	0.713088	ScPdPb	0.540573
TiSnRu	0.712848	LiGeRh	0.536203
MgInNi	0.712028	TiSnCu	0.534805
SnHfCo	0.707933	LiPdIn	0.527771
VSnFe	0.76901	CrGeRu	0.512912
LiGeNi	0.700193	LiCuGe	0.508319

Table 3-5. Highest probability for half-Heusler structures.

samples	probabilities	samples	probabilities	samples	probabilities
PdCdBi	0.9925665	NbBiPt	0.941015	MnSnPd	0.923915
SbTaNi	0.9892582	ZrBiAu	0.940976	NbBiCo	0.923572
BiLaNi	0.985461	MnRhPb	0.940318	CrSbHg	0.923149
NiCdBi	0.9847144	AgHgBi	0.940167	TiBiAu	0.923033
BiPrPt	0.9835108	MgSbPb	0.939578	MnRhIn	0.922864
BiLaAu	0.9833609	SbTaPt	0.939273	MgSbIr	0.922363
BiLaHg	0.982716	ZnRhPb	0.939022	MnSbHg	0.921615
SbUPt	0.9821082	MnRhTl	0.938971	TiSbOs	0.92123
BiLuHg	0.9798159	ZnPdSb	0.938014	BiTbHg	0.921187
BiSmPt	0.9727946	CuZnBi	0.937885	LiMgI	0.920599
HfBiAu	0.9722375	NbBiNi	0.937402	BiGdZn	0.920305
BiLuAu	0.9719914	SbUPd	0.937233	NbSbNi	0.920065
BiCeHg	0.9659705	MgSbCo	0.936808	NbBiAu	0.920052
BiLaZn	0.9647496	BiTbZn	0.936188	ZnPdIn	0.91858
BiDyPd	0.9619523	LiGaBi	0.935529	CrSbAu	0.918371
PdCdPb	0.9609973	BiTmHg	0.935256	RhCdBi	0.91725
BiGdHg	0.960758	MgBiSn	0.934833	TiBiRu	0.915882
BiYbNi	0.9581364	MgSbAu	0.934196	TiBiRh	0.915459
NbSbPt	0.957862	MgAgBi	0.933027	BiPrZn	0.915128
MgBiPt	0.9488026	MgSbFe	0.932152	ZnPdTe	0.914501
ZrBiPt	0.9479557	CrBiAu	0.931845	MgSbZn	0.914197
ZnPdBi	0.9479335	BiUPt	0.930309	ZnRhTl	0.913619
MnAuTl	0.9476165	MgSbRh	0.929953	BiTmPb	0.913425
CuCdBi	0.9475496	CrCdBi	0.92977	CuCdSn	0.913366
ZnAgBi	0.9475029	BiCeAu	0.929528	NbBiFe	0.911177
MgBiCo	0.9471268	SbTaPd	0.929069	AsPtTh	0.911061
ZnRhBi	0.9470903	MgSbOs	0.928982	ZnRhSb	0.911022
TiBiPt	0.9463257	BiYbHg	0.928741	NbCdBi	0.910216
MgBiFe	0.9455812	ZrBiRu	0.928405	MgPbCo	0.907448
ZnPdPb	0.9454404	ZnPdTl	0.928359	BiDyZn	0.906976
BiLaPb	0.9448877	MgSbIn	0.927812	BiGdAu	0.906375
MgBiRh	0.9446175	LiAlBi	0.927354	MnSbTl	0.905418
MgBiAu	0.9446037	MgSbSn	0.926572	MnRhSn	0.904742
CdPtBi	0.9443456	MgBiCr	0.92642	BiLuZn	0.904638
MgBiRu	0.9439655	MnAuBi	0.925895	MgBiNb	0.902902
MgBiIr	0.9428587	BiLaTl	0.925168	NbBiRu	0.902436
MgBiPd	0.9425756	ZnRhSn	0.924664	BiTmZn	0.90069
CoCdBi	0.941657	MgPbNi	0.924595		
TiSbPt	0.941534	CuCdPb	0.924557		

CdHgBi	0.9414863	MgBiMo	0.923983

Future experimental work includes the synthesis and characterization of a set of some of the predicted compounds. Several candidates that follow the 18-electron rule (e.g., VSnCo, CrSnRu, and TiSnPd) would be a good starting point to determine if semiconducting compounds with a full-Heusler counterpart can be prepared. CrSnRu is particularly interesting because there is only one other Cr-containing half-Heusler compound known to date. Synthesis of the compounds with the top ten probabilities would help evaluate if the model is overoptimistic or not. An inspection of the other candidates suggests that Li- and Mg-containing candidates are worthwhile exploring.

3.4. Conclusion

A machine learning model was developed to predict new members of the half-Heusler family based on features that depend on composition alone. The final model showed an improvement over previous work with a validation accuracy of 98%. In developing the model, two machine-learning techniques not commonly used in the materials informatics were compared. The first technique, minority oversampling, resulted in a marked improvement in the sensitivity of the model. The second technique, an ensemble of several models whose outputs have been combined via a soft voting algorithm, resulted in a larger improvement in the sensitivity of model. Combination of both techniques resulted in a model that performs just as well or better in the performance metrics reported in this work but also shows a smaller performance drop when comparing the training and validation performance metric. This indicates that these techniques have made the model more reliable by combatting the phenomenon of overfitting. Experimental work is underway to validate the model with a focus on half-Heusler compounds that have a full-Heusler counterpart and have the highest prediction probabilities.

3.5. References

1) Heusler, F.; Stark, W.; Haupt, E. Über die synthese ferromagnetischer Manganlegierungen. *Verh. Deut. Phys. Ges* **1903**, *144*, 340–223.

2) *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD)*, release 2015/16; ASM International: Materials Park, OH, **2016**.

3) Oliynyk, A.O.; Antono, E.; Sparks,T.D.; Ghadbeigi, L.; Gaultois, M.W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

4) Takamura, Y.; Nakane, R.; Sugahara, S. Quantitative analysis of atomic disorders in full-Heusler Co_2FeSi alloy thin films using x-ray diffraction with Co Kα and Cu Kα sources. *J. Appl. Phys.* **2010**, *107*, 09B111.

5) Bose, S.K.; Kudrnovsky, J.; Liu, Y. Structure and physical properties of quaternary Heusler alloy NiMnCuSb. *J. Magn. Magn. Mater.* **2017**, *444*, 338–343.

6) Yu, G.H.; Xu, Y.L.; Qiu, H.M.; Zhu, Z.Y.; Huang, X.P. Pan, L.Q. Recent progress in Heusler-type magnetic shape memory alloys. *Rare Met.* **2015**, *34*, 527–539.

7) Carrete, J.; Li,W.; Mingo, N.; Wang, S.; Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **2014**, *4*, 011019.

8) Bhattacharya, S.; Madsen, G.K.H. A novel p-type half Heusler from high-throughput transport and defect calculations. *J. Mater. Chem. C.* **2016**, *4*, 11261.

9) Sanvito, S.; Oses, C.; Xue, A.; Tiwari, A.; Zic, M.; Arcger, T.; Tozman, P.; Venkatesan,M.; Coey, M.; Curtarolo, S. Accelarated discovery of new magnets in the Heusler alloy family. *Sci. Adv.* **2017**, *3*, e1602241.

10) Gao, Q.; Opahle, I.; Zhang, H. High-throughput screening for spin-gapless semiconductors in quaternary Heusler compounds. *Phys. Rev. Mater.* **2019**, *3*, 024410.

11) See http://heusleralloys.mint.ua.edu for University of Alabama MINT Heusler Database, **2015**.

12) Anand, S.; Xia, K.; Hegde, V.I.; Ayedemir, U.; Kocevski, V.; Zhu, T.; Wolverton, C.; Snyder, J. A valence balanced rule for discovery of 18-electron half-Heuslers with defects. *Energy Environ. Sci.* **2018**, *11*, 1480–1488.

13) Graf, T.; Felser, C.; Parkin, S. S. P. Simple rules for the understanding of Heusler compounds. *Prog. Solid State Chem.* **2011**, *39*, 1–50.

14) Wang, X.; Cheng,Z.; Yuan, H.;Khenata, R.; $L2_1$ and Xa ordering completion in titanium-based full-Heusler alloys. *J. Mater. Chem. C.* **2017**, *5*, 11559–11564.

15) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G.K.H.; Mingo, N. Materials screening for the discovery of new half-Heuslers: machine learning versus ab initio methods. *J. Phys. Chem. B* **2018**, *122*, 625–632.

16) Levin, E.E.; Bocarsly, J.D.; Wyckoff, K.E.; Pollock, T.M.; Seshadri, R.; Tuning the magnetocaloric response in half-Heusler/Heusler $MnNi_{1+x}Sb$ solid solutions. *Phys. Rev. mater.* **2017**, *1*, 075003.

17) Webster, P.J.; Mankikar, R.M. Chemical order and magnetic properties of the $Ni_{2-x}MnSb$ system. *J. Magn. Magn. Mater.* **1984**, *42*, 300–308.

18) Nagasako, M.; Taguchi, Y.; Miyamoto, T.; Kanomata, T.; Ziebeck, K.R.A.; Kainuma, R. Order-disorder transition of vacancies from the full- to the half-Heusler structure in $Ni_{2-x}MnSb$ alloys. *Intermetallics* **2014**, *61*, 38–41.

19) Brown, P.J.; Gandy, A.P.; Kainuma, R.; Kanomata, T.; Miyamoto, T.; Nagasako, M.; Neumann, K.U.; Sheikh, A.: Ziebeck, K.R.A. Atomic order and magnetization distribution in the half metallic and nearly half metallic C1b compounds NiMnSb and PdMnSb. *J. Phys. Condens. Matter* **2010**, *22*, 206004.

20) Kouacou, M.A.; Koua, A.A.; Zoueu, J.T.; Konan, K.; Pierre, J. Onset of itinerant ferromagnetism associated with semiconductor-metal transition in $Ti_xNb_{1-x}CoSn$ half Heusler solid solution compounds. *Pramana* **2008**, *71*, 157–166.

21) Chauhan, N.S.; Gahtpri, B.; Sivaiah, B.; Mahanti, D.S.; Dhar, A.; Bhattacharya, A. Modulating the lattice dynamics of n-type Heusler compounds via tuning Ni concentration. *Appl. Phys. Let.* **2018**, *113*,013902.

22) *PLS_Toolbox*, version 8.0.1; Eigenvector Research Inc.: Wenatchee, WA, **2018**.

23) *MATLAB Statistics and Machine Learning Toolbox*, release 2018a; The Mathworks Inc.: Natick, MA, **2018**.

24) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

25) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

26) Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE:synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

27) A.S. Gzyl, A.O. Oliynyk, L. Adutwum, A. Mar, Solving the coloring problem in half-Heusler structures: Machine-learning predictions and experimental validation, *Inorg. Chem*. **2019**, *58*, 9280–9289.

Chapter 4 Conclusions

4.1. Machine learning approach to data sanitization

A machine learning approach is a powerful recent addition to the chemists' toolbox but has specific requirements in order for it to be useful. Its effectiveness depends on data quality and representation of the samples. Data quality relies on good science being done in a standardized way and the accuracy of the database. It is testament to the high and rigorous standards of solid state chemists and crystallographers who have long had the tradition of storing useful crystallographic information in databases so that future generations can take advantage of it. Nevertheless, there are two problems that pertain and are particular to the half-Heusler compounds. The first problem is the ambiguity in site distributions, which are not typically analyzed because of a lack of incentive to do so when there are no refinable position parameters. The second problem is inconsistency between what is reported in databases and what was actually written in the paper. In Chapter 2, we address this problem by applying machine learning to sanitize the site occupancy for half-Heusler compounds in a high-throughput manner. This achieves several goals. First, transcription errors can be identified in the database. Second, only with correct site assignments can reliable chemical rules and accurate structure-property relationships be formulated. Third, DFT calculations can be initiated properly by indicating the correct structures to input. Fourth, clean data lead to more reliable representations to be used in building machine-learning models. For example, if a set of features is generated that is not invariant to the order in which elements are listed in a formula, several representations of the same compound are possible. Elements are then sorted in columns, each representing the occupation of a crystallographic site. If the site occupation is reported incorrectly, the structure of the compound is also represented incorrectly. Additionally, the machine-learning model was experimentally verified by synthesizing the flagged

compounds GdPtSb and HoPdBi, which adds to the credibility of using a machine learning approach. The machine-learning model outperforms the traditional approach to assigning site occupancy with a validation accuracy of 0.929 vs 0.659 (Allred-Rochow).

4.2. Machine learning approach to materials discovery

Chapter 3 builds upon the work of Chapter 2 by utilizing the corrected site occupancies to develop a machine-learning model for the high-throughput discovery of new Heusler compounds. In a traditional approach to synthesis of new compounds, obvious chemical substitutions of existing compounds lead researchers to investigate only a narrow portion of chemical space, introducing a distribution bias. Applying rules based on this biased distribution only exacerbates this problem of tunnel vision. Serendipitous discovery can add new members that extends the distribution in new directions, but this is not generally an efficient way to search the space. Utilizing a machine-learning approach, we can quickly search through chemical space and discover new compounds by learning from relations in higher dimensions. We employed a novel machine-learning framework and sought to experimentally validate its output. This serves two purposes: (1) to discover new members of an important family of compounds, and (2) to convince the scientific community of the validity of a machine-learning approach. An ensemble method was used to combat overfitting and SMOTE was applied to tackle the class imbalance problem which is ever present in this field. The performance of the model showed incremental improvement after application of the techniques (Table 4.1).

Table 4-1. Comparison of model performance

Model	Training set	Validation set
Before feature selection (ensemble)		
(sensitivity / specificity)	0.858 / 0.976	0.733 / 0.971
accuracy	0.969	0.964
After feature selection (ensemble)		
(sensitivity / specificity)	0.900 / 0.989	0.867 / 0.983
accuracy	0.978	0.975
After SMOTE (ensemble)		
(sensitivity / specificity)	0.950 / 0.986	0.900 / 0.982
accuracy	0.981	0.977
After SMOTE (best individual)		
(sensitivity / specificity)	0.954 / 0.997	0.816 / 0.985
accuracy	0.993	0.974

The ability to correctly identify half-Heusler compounds (sensitivity) is highest for the ensemble method combined with SMOTE. The same model also has the highest validation accuracy and the lowest degree of overfitting. Synthetic will be focused on predicted half-Heusler compounds that have a full Heusler counterpart. This will narrow down synthetic efforts even further while simultaneously increasing the number of systems where the transition between the two can be studied.

4.3. Future work

In this work, machine learning has been applied to solve the coloring problem for half-Heusler structures and to predicted new half-Heusler compounds. Although a first draft of a manuscript has been written for Chapter 3, more experimental exploration can be done in this system. The family of Heusler compounds is very large, spanning a wide range of compositions whose formation is still not completely understood. For example, why are there are so few half-Heusler counterparts to the other family members? The transition from half-Heusler to full-Heusler structures (solid solutions with composition ABC_{2-x}) is of particular interest because of the structures transform from noncentrosymmetric to centrosymmetric. Using machine learning to populate the families and help pick out erroneously characterized samples from the past can aid in understanding the fundamental chemistry of this important family of compounds. While features were being developed for the machine-learning algorithms, it became apparent that there is no simple way to represent nonstoichiometric compounds, or structures exhibiting disorder. Although developing machine-learning representations of molecular organic and inorganic compounds is now a popular area of study, the same cannot be said of non-molecular inorganic solids. A long term goal is to address the balance of machine representation and chemical interpretability. It would be interesting to apply generative models instead of classification models to tackle materials prediction problems.

Bibliography

1) Heusler, F.; Stark, W.; Haupt, E. Über die synthese ferromagnetischer Manganlegierungen. *Verh. Deut. Phys. Ges* **1903**, *144*, 340–223.

2) Kübler, J.; Williams, A.R.; Sommers, C.B. Formation and coupling of magnetic moments in Heusler alloys. *Phys. Rev. B.* **1983**, *28*, 1745–1755.

3) Bradley, A. J.; Rodgers, J. W. The Crystal Structure of the Heusler Alloys. *Proc. R. Soc. London, Ser. A* **1934**, *144*, 340–359.

4) Khawam, A.; Flanagan, D.R. Solid-state kinetic models: basics and mathematical fundamentals. *J. Phys. Chem. B.* ***2006****, 110*, 17315–17328.

5) Vlack, V. *Materials science for engineers 6th ed*., Addison-Wesley: Reading, Massachusetts, **1975**.

6) Borg, R.J.; Dienes, G.J. *An introduction to solid state diffusion,* Academic press, **1988.**

7) Merkle, R.; Maier, J. On the Tammann rule. *Z. Anorg. Allg. Chem.* **2005**, *631*, 1163–1166.

8) West, A. R. Basic Solid State Chemistry, 2nd ed., Wiley: New York, **1999**.

9) Massa, W. Crystal Structure Determination, 2nd ed., Springer–Verlag: Berlin, **2004.**

10) Dronkowski, R. *Computational chemistry of solid-state materials*, Wiley-VCH, Weinheim, **2005**.

11) Dronkowski, R.; Blöchl, P.E. Crystal orbital Hamiltonian populations (COHP) Energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *J. Phys. Chem.* **1993**, *97*, 8617–8624.

12) Deringer, V.L; Tchougreeff, A.L.; Dronskowski, R. Crystal orbital Hamiltonian population analysis as projected from plane-wave basis sets. *J. Phys. Chem. A* **2011**, *115*, 5461–5466.

13) Skriver, H.L. *The LMTO method,* Springer-Verlag Berlin Heidelberg, **1984.**

14) Tank, R.; Jepsen, O.; Burkhardt, A.; Andersen, O.K.; *TB-LMTO-ASA program*, version 4.7; Max Planck Institut für Festkörperforschung: Stuttgart, Germany, **1998.**

15) Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229.

16) Mitchell, T.M. *Machine learning*, Mcgraw-Hill, **1997**.

17) Tehrani, A.M.; Oliynyk, A.O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T.D.; Brgoch, J. Machine learning directed search for ultraincompressable, superhard materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844–9853.

18) Zhang, H.; Moon, S.K.; Ngo, T.H. Hybrid machine learning method to determine the optimal operating process window in aerosol jet 3d printing. *ACS Appl. Mater. Interfaces* **2019**, *11*, 17994–18003.

19) Rashidi, M.; Wolkow, R.A. Autonomous scanning probe microscopy in situ tip conditioning through machine learning. *ACS Nano* **2018**, *12*, 5185–5189.

20) Hou, Z.; Takagiwa, Y.; Shinohara, Y.; Xu, Y.; Tsuda, K. Machine-learning-assisted development and theoretical consideration for the $Al_2Fe_3Si_3$ thermoelectric material. *ACS Appl. Mater. Interfaces* **2019**, *11*, 11545–11554.

21) Liu, R.; Kumar, A.; Chen, Z.; Agrawal, A.; Sundararaghavan, V.; Choudhary, A. A predictive machine learning approach for microstructure optimization and materials design. *Sci. Rep.* **2015**, 11551.

22) Hong, S.; Nomura, K.; Krishnamoorthy, A.; Rajak, P.; Sheng, C.; Kalia, R.K.; Nakano, A.; Vashishta, P. Defect healing in layered materials: a machine learning-assisted characterization of MoS2 crystal phases. *J. Phys. Chem. Lett.* **2019**, *10*, 2739–2744.

23) Batra, R.; Tran, H.D.; Kim, C.; Chapman, J.; Chen, L.; Chandrasekaran, A.; Ramprasad, R. A general atomic neighbourhood fingerprint for machine learning based methods *J. Phys. Chem. C.* **2019**, *123*, 15859–15866.

24) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564–3572.

25) Xie, T.; Grossman, J.C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys.Rev. Lett.* **2018**, 145301.

26) Lengeling-Sanchez, B.; Outeiral, C.; Guimaraes, G.L.; Aspuru-Guzik, A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *Chemrxiv.* **2017**, Preprint.

27) Hill J., Mannodi-Kanakkithodi A., Ramprasad R., Meredig B. *Materials Data Infrastructure and Materials Informatics*. In: Shin D., Saal J. (eds) *Computational Materials System Design*. Springer: Cham **2018**.

28) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

29) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

30) Abdi, H.; Williams, L. J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010,** *2*, 433–459.

31) Lee, L.C.; Liong, C.Y.; Jemain, A.A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. *Analyst,* **2018**, *143*, 3526–3539.

32) Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

33) Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Transactions on intelligent systems and technology,* **2011**, *2*, 1–27.

34) Huheey, J. E.; Keiter, E. A.; Keiter, R. L. *Inorganic Chemistry: Principles of Structure and Reactivity*, 4th Ed.; Harper Collins: New York, **1993**.

35) Burdett, J. K.; Lee, S. L.; McLarnan, T. J. The coloring problem. *J. Am. Chem. Soc.* **1985**, *107*, 3083–3089.

36) Miller, G. J. The "coloring problem" in solids: How it affects structure, composition and properties. *Eur. J. Inorg. Chem.* **1998**, 523–536.

37) Graf, T.; Felser, C.; Parkin, S. S. P. Simple rules for the understanding of Heusler compounds. *Prog. Solid State Chem.* **2011**, *39*, 1–50.

38) Casper, F.; Graf, T.; Chadov, S.; Balke, B.; Felser, C. Half-Heusler compounds: novel materials for energy and spintronic applications. *Semicond. Sci. Technol.* **2012**, *27*, 063001-1–063001-8.

39) Bos, J.-W. G.; Downie, R. A. Half-Heusler thermoelectrics: a complex class of materials. *J. Phys.: Condens. Matter* **2014**, *26*, 433201-1–433201-15.

40) Palmstrøm, C. J. Heusler compounds and spintronics. *Prog. Cryst. Growth Charact. Mater.* **2016**, *62*, 371–397.

41) Zeier, W. G.; Schmitt, J.; Hautier, G.; Aydemir, U.; Gibbs, Z. M.; Felser, C.; Snyder, G. J. Engineering half-Heusler thermoelectric materials using Zintl chemistry. *Nat. Rev. Mater.* **2016**, *1*, 16032-1–16032-10.

42) Wollmann, L.; Nayak, A. K.; Parkin, S. S. P.; Felser, C. Heusler 4.0: Tunable materials. *Annu. Rev. Mater. Res.* **2017**, *47*, 247–270.

43) Nowotny, H.; Sibert, W. Ternäre Valenzverbindungen in den Systemen Kupfer (Silber)–Arsen (Antimon, Wismut)–Magnesium. *Z. Metallkd.* **1941**, *33*, 391–394.

44) Nespolo, M. Lattice versus structure, dimensionality versus periodicity: a crystallographic Babel? *J. Appl. Crystallogr.* **2019**, *52*, doi: 10.1107/S1600576719000463.

45) *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD)*, release 2015/16; ASM International: Materials Park, OH, 2016.

46) Gordy, W.; Thomas, W. J. O. Electronegativities of the elements. *J. Chem. Phys.* **1956**, *24*, 439–444.

47) Allred, A. L.; Rochow, E. G. A scale of electronegativity based on electrostatic force. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264–268.

48) Bende, D.; Grin, Y.; Wagner, F. R. Covalence and ionicity in MgAgAs-type compounds. *Chem. Eur. J.* **2014**, *20*, 9702–9708.

49) Bende, D.; Wagner, F. R.; Grin, Y. 8–*N* rule and chemical bonding in main-group MgAgAs-type compounds. *Inorg. Chem.* **2015**, *54*, 3970–3978.

50) White, M. A.; Medina-Gonzalez, A. M.; Vela, J. Soft chemistry, coloring and polytypism in filled tetrahedral semiconductors: Toward enhanced thermoelectric and battery materials. *Chem. Eur. J.* **2018**, *24*, 3650–3658.

51) Hames, F. A. Ferromagnetic-alloy phases near the compositions Ni_2MnIn, Ni_2MnGa, Co_2MnGa, Pd_2MnSb, and PdMnSb. *J. Appl. Phys.* **1960**, *31*, S370–S371.

52) Endo, K. Magnetic studies of $C1_b$-compounds CuMnSb, PdMnSb and Cu_{1-x} (Ni or Pd) $_{xMnSb}$. *J. Phys. Soc. Jpn.* **1970**, *29*, 643–649.

53) Webster, P. J.; Ziebeck, K. R. A. Structures of $Pd_{2-x}MnSb$ – an improved neutron polarizer? *J. Magn. Magn. Mater.* **1980**, *15–18*, 473–474.

54) Buschow, K. H. J.; van Engen, P. G.; Jongebreur, R. Magneto-optical properties of metallic ferromagnetic materials. *J. Magn. Magn. Mater.* **1983**, *38*, 1–22.

55) Larson, P.; Mahanti, S. D.; Kanatzidis, M. G. Structural stability of Ni-containing half-Heusler compounds. *Phys. Rev. B* **2000**, *62*, 12754–12762.

56) Schuster, H.-U.; Dietsch, W. Eine neue ternäre Phase im System Li–Au–Sb. *Z. Naturforsch. B* **1975**, *30*, 133.

57) Zhang, X.; Yu, L.; Zakutayev, A.; Zunger, A. Sorting stable versus unstable hypothetical compounds: The case of multi-functional ABX half-Heusler filled tetrahedral structures. *Adv. Funct. Mater.* **2012**, *22*, 1425–1435.

58) Carrete, J.; Li, W.; Mingo, N.; Wang, S.; Curtarolo, S. Finding unprecedently low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **2014**, *4*, 011019-1–011019-9.

59) Gauthier, R.; Zhang, X.; Hu, L.; Yu, L.; Lin, Y.; Sunde, T. O. L.; Chon, D.; Poeppelmeier, K. R.; Zunger, A. Prediction and accelerated laboratory discovery of previously known 18-electron ABX compounds. *Nat. Chem.* **2015**, *7*, 308–316.

60) Ma, J.; Hegde, V. I.; Munira, K.; Xie, Y.; Keshavarz, S.; Mildebrath, D. T.; Wolverton, C.; Ghosh, A. W.; Butler, W. H. Computational investigation of half-Heusler compounds for spintronic applications. *Phys. Rev. B* **2017**, *95*, 024411-1–024411-25.

61) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

62) Oliynyk, A. O.; Mar, A. Discovery of intermetallic compounds from traditional to machine-learning approaches. *Acc. Chem. Res.* **2018**, *51*, 59–68.

63) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G. K. H.; Mingo, N. Materials screening for the discovery of new half-Heuslers: Machine learning versus ab initio methods. *J. Phys. Chem. B* **2018**, *122*, 625–632.

64) *PLS_Toolbox*, version 8.0.1; Eigenvector Research Inc.: Wenatchee, WA, 2018.

65) *MATLAB Statistics and Machine Learning Toolbox*, release 2018a; The Mathworks Inc.: Natick, MA, 2018.

66) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

67) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

68) Coelho, A. A. *TOPAS-Academic*, version 6; Coelho Software: Brisbane, Australia, 2007.

69) Sheldrick, G. M. *SHELXTL*, version 6.12; Bruker AXS Inc.: Madison, WI, 2001.

70) Kresse, G.; Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

71) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

72) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

73) Grin, Y.; Savin, A.; Silvi, B. The ELF perspective of chemical bonding. In *The Chemical Bond: Fundamental Aspects of Chemical Bonding*; Frenking, G., Shaik, S., Eds.; Wiley-VCH: Weinheim, 2014; Chap. 10, pp 345–382.

74) Tang, W.; Sanville, E.; Henkelmann, G. A grid-based Bader analysis algorithm without lattice basis. *J. Phys. Condens. Matter* **2009**, *21*, 084204-1–084204-7.

75) Dronskowski, R.; Blöchl, P. E. Crystal orbital Hamilton populations (COHP). Energy-resolved visualization of chemical bonding in solids based on density-functional calculations. *J. Phys. Chem.* **1993**, *97*, 8617–8624.

76) Tank, R.; Jepsen, O.; Burkhardt, A.; Andersen, O. K. *TB-LMTO-ASA Program*, version 4.7; Max Planck Institut für Festkörperforschung: Stuttgart, Germany, 1998.

77) Oliynyk, A. O.; Adutwum, L. A.; Harynuk, J. A.; Mar, A. Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis. *Chem. Mater.* **2016**, *28*, 6672–6681.

78) Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Sotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.; Brgoch, J. Disentangling structural confusion through machine

learning: Structure prediction and polymorphism of equiatomic ternary phases *ABC*. *J. Am. Chem. Soc.* **2017**, *139*, 17870–17881.

79) The VIP score depends on the square of the normalized weight of a variable in a PLS-DA model, and is interpreted as the proportion of variances in class assignments that can be explained by that variable.

80) Suresh, C. H. A consistent approach toward atomic radii. *J. Phys. Chem. A* **2001**, *105*, 5940–5944).

81) Pauling, L.; Huggins, M. L. Covalent radii of atoms and interatomic distances in crystals containing electron-pair bonds. *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **1934**, *87*, 205–238.

82) Pauling, L. The sizes of ions and the structure of ionic crystals. *J. Am. Chem. Soc.* **1927**, *49*, 765–790.

83) Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr., Sect. A* **1976**, *32*, 751–767.

84) Miracle, D. B. The efficient cluster packing model – An atomic structural model for metallic glasses. *Acta Mater.* **2006**, *54*, 4317–4336.

85) Zunger, A. Systematization of the stable crystal structure of all *AB*-type binary compounds: A pseudopotential orbital-radii approach. *Phys. Rev. B* **1980**, *22*, 5839–5872.

86) Allred, A. L. Electronegativity values from thermochemical data. *J. Inorg. Nucl. Chem.* **1961**, *17*, 215–221.

87) Little, E. J. Jr.; Jones, M. M. A complete table of electronegativities. *J. Chem. Ed.* **1960**, *37*, 231–233.

88) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bull.* **2016**, *41*, 399–409.

89) Heusler, F.; Stark, W.; Haupt, E. Über die synthese ferromagnetischer Manganlegierungen. *Verh. Deut. Phys. Ges* **1903**, *144*, 340–223.

90) *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD)*, release 2015/16; ASM International: Materials Park, OH, **2016**.

91) Oliynyk, A.O.; Antono, E.; Sparks,T.D.; Ghadbeigi, L.; Gaultois, M.W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.

92) Takamura, Y.; Nakane, R.; Sugahara, S. Quantitative analysis of atomic disorders in full-Heusler Co_2FeSi alloy thin films using x-ray diffraction with Co Kα and Cu Kα sources. *J. Appl. Phys.* **2010**, *107*, 09B111.

93) Bose, S.K.; Kudrnovsky, J.; Liu, Y. Structure and physical properties of quaternary Heusler alloy NiMnCuSb. *J. Magn. Magn. Mater.* **2017**, *444*, 338–343.

94) Yu, G.H.; Xu, Y.L.; Qiu, H.M.; Zhu, Z.Y.; Huang, X.P. Pan, L.Q. Recent progress in Heusler-type magnetic shape memory alloys. *Rare Met.* **2015**, *34*, 527–539.

95) Carrete, J.; Li, W.; Mingo, N.; Wang, S.; Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **2014**, *4*, 011019.

96) Bhattacharya, S.; Madsen, G.K.H. A novel p-type half Heusler from high-throughput transport and defect calculations. *J. Mater. Chem. C.* **2016**, *4*, 11261.

97) Sanvito, S.; Oses, C.; Xue, A.; Tiwari, A.; Zic, M.; Arcger, T.; Tozman, P.; Venkatesan,M.; Coey, M.; Curtarolo, S. Accelarated discovery of new magnets in the Heusler alloy family. *Sci. Adv.* **2017**, *3*, e1602241.

98) Gao, Q.; Opahle, I.; Zhang, H. High-throughput screening for spin-gapless semiconductors in quaternary Heusler compounds. *Phys. Rev. Mater.* **2019**, *3*, 024410.

99) See http://heusleralloys.mint.ua.edu for University of Alabama MINT Heusler Database, **2015**.

100) Anand, S.; Xia, K.; Hegde, V.I.; Ayedemir, U.; Kocevski, V.; Zhu, T.; Wolverton, C.; Snyder, J. A valence balanced rule for discovery of 18-electron half-Heuslers with defects. *Energy Environ. Sci.* **2018**, *11*, 1480–1488.

101) Graf, T.; Felser, C.; Parkin, S. S. P. Simple rules for the understanding of Heusler compounds. *Prog. Solid State Chem.* **2011**, *39*, 1–50.

102) Wang, X.; Cheng,Z.; Yuan, H.;Khenata, R.; $L2_1$ and Xa ordering completion in titanium-based full-Heusler alloys. *J. Mater. Chem. C.* **2017**, *5*, 11559–11564.

103) Legrain, F.; Carrete, J.; van Roekeghem, A.; Madsen, G.K.H.; Mingo, N. Materials screening for the discovery of new half-Heuslers: machine learning versus ab initio methods. *J. Phys. Chem. B* **2018**, *122*, 625–632.

104) Levin, E.E.; Bocarsly, J.D.; Wyckoff, K.E.; Pollock, T.M.; Seshadri, R.; Tuning the magnetocaloric response in half-Heusler/Heusler $MnNi_{1+x}Sb$ solid solutions. *Phys. Rev. Mater.* **2017**, *1*, 075003.

105) Webster, P.J.; Mankikar, R.M. Chemical order and magnetic properties of the $Ni_{2-x}MnSb$ system. *J. Magn. Magn. Mater.* **1984**, *42*, 300–308.

106) Nagasako, M.; Taguchi, Y.; Miyamoto, T.; Kanomata, T.; Ziebeck, K.R.A.; Kainuma, R. Order-disorder transition of vacancies from the full- to the half-Heusler structure in $Ni_{2-x}MnSb$ alloys. *Intermetallics* **2014**, *61*, 38–41.

107) Brown, P.J.; Gandy, A.P.; Kainuma, R.; Kanomata, T.; Miyamoto, T.; Nagasako, M.; Neumann, K.U.; Sheikh, A.: Ziebeck, K.R.A. Atomic order and magnetization distribution in the half metallic and nearly half metallic C1b compounds NiMnSb and PdMnSb. *J. Phys. Condens. Matter* **2010**, *22*, 206004.

108) Kouacou, M.A.; Koua, A.A.; Zoueu, J.T.; Konan, K.; Pierre, J. Onset of itinerant ferromagnetism associated with semiconductor-metal transition in $Ti_xNb_{1-x}CoSn$ half Heusler solid solution compounds. *Pramana* **2008**, *71*, 157–166.

109) Chauhan, N.S.; Gahtpri, B.; Sivaiah, B.; Mahanti, D.S.; Dhar, A.; Bhattacharya, A. Modulating the lattice dynamics of n-type Heusler compounds via tuning Ni concentration. *Appl. Phys. Let.* **2018**, *113*,013902.

110) *PLS_Toolbox*, version 8.0.1; Eigenvector Research Inc.: Wenatchee, WA, **2018**.

111) *MATLAB Statistics and Machine Learning Toolbox*, release 2018a; The Mathworks Inc.: Natick, MA, **2018**.

112) Sinkov, N. A.; Harynuk, J. J. Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling. *Talanta* **2011**, *83*, 1079–1087.

113) Adutwum, L. A.; de la Mata, A. P.; Bean, H. D.; Hill, J. E.; Harynuk, J. J. Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios. *Anal. Bioanal. Chem.* **2017**, *409*, 6699–6708.

114) Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE:synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.

115) A.S. Gzyl, A.O. Oliynyk, L. Adutwum, A. Mar, Solving the coloring problem in half-Heusler structures: Machine-learning predictions and experimental validation, *Inorg. Chem*. **2019**, *58*, 9280–9289.

Appendix 1 Supplementary Data for Chapter 2

Table S1. Elemental Properties

1. atomic number	23. cohesive energy
2. atomic mass	24. number of s valence electrons
3. period number	25. number of p valence electrons
4. group number	26. number of d valence electrons
5. family number	27. number of f valence electrons
6. metal / metalloid / nonmetal	28. number of unfilled s orbitals
7. Mendeleev number	29. number of unfilled p orbitals
8. quantum number *l*	30. number of unfilled d orbitals
9. atomic radius	31. number of unfilled f orbitals
10. Miracle radius	32. number of outer-shell electrons
11. covalent radius	33. first ionization energy
12. Zunger pseudopotential radii sum	34. polarizability
13. ionic radius	35. melting point
14. crystal radius	36. boiling point
15. Pauling electronegativity	37. density
16. Martynov-Batsanov electronegativity	38. specific heat
17. Gordy electronegativity	39. heat of fusion
18. Mulliken electronegativity	40. heat of vaporization
19. Allred-Rochow electronegativity	41. thermal conductivity
20. metallic valence	42. heat of atomization
21. number of valence electrons	43. bulk modulus
22. Gilman number of valence electrons	

Table S2. Arithmetic Operations Applied to Elemental Properties

operation	description
$½(X_a + X_b) - X_c$	difference between (average of atoms in 4*a* and 4*b*) and atoms in 4*c*
$½(X_a + X_b) / X_c$	ratio of (average of atoms in 4*a* and 4*b*) and atoms in 4*c*
$\min\{X_a, X_b\} - X_c$	difference between (minimum of atoms in 4*a* and 4*b*) and atoms in 4*c*
$\max\{X_a, X_b\} - X_c$	difference between (maximum of atoms in 4*a* and 4*b*) and atoms in 4*c*
$\min\{X_a, X_b\} / X_c$	ratio of (minimum of atoms in 4*a* and 4*b*) and atoms in 4*c*
$\max\{X_a, X_b\} / X_c$	ratio of (maximum of atoms in 4*a* and 4*b*) and atoms in 4*c*

Table S3. Probabilities for Correctness of Site Distributions in Half-Heusler Compounds

- Each sample *ABC* is labeled by specifying the occupation of the 4*c* site. For example, MgAgAs-Mg, MgAsAs-Ag, and MgAgAs-As refer to model structures in which Mg, Ag, and As atoms occupy the 4*c* site respectively.
- The samples were split into a training set (2/3) and a validation set (1/3).
- Class 1 refers to site distributions as experimentally reported in the literature; Class 2 refers to alternative site distributions with the 4*c* site being occupied by other atoms.
- Probabilities were evaluated on an SVM model after features were chosen through a CR-FS procedure.

no.	sample	correctness probability
1	ZnAgAs-As	0.881
2	AgCdSb-Sb	0.881
3	AlBeB-B	0.881
4	CrAlCo-Co	0.881
5	AlLiSi-Si	0.906
6	CdLiAs-As	0.881
7	MnLiAs-As	0.921
8	ZnLiAs-As	0.881
9	CaBiAu-Au	0.937
10	BiYbAu-Au	0.924
11	DyPbAu-Au	0.962
12	ErPbAu-Au	0.939
13	GdPbAu-Au	0.959
14	HoPbAu-Au	0.960
15	LiAuSb-Sb	0.881
16	SnLuAu-Au	0.900
17	AuMnSb-Sb	0.881
18	SnMnAu-Au	0.638
19	YPbAu-Au	0.955
20	ScSnAu-Au	0.881
21	CeBiPd-Pd	0.904
22	BiCePt-Pt	0.881
23	MgBiCu-Cu	0.881
24	BiDyNi-Ni	0.922
25	BiErNi-Ni	0.935
26	BiErPd-Pd	0.938
27	ZnFeBi-Bi	0.160
28	BiGdNi-Ni	0.915
29	BiGdPt-Pt	0.900
30	BiHoNi-Ni	0.924
31	BiHoPt-Pt	0.913
32	BiLaPd-Pd	0.893
33	MgLiBi-Bi	0.600
34	BiVLi-Li	0.625
35	BiLuPd-Pd	0.940
36	BiLuPt-Pt	0.928
37	BiNdNi-Ni	0.874
38	PtNdBi-Bi	0.226
39	BiScNi-Ni	0.951
40	BiSmNi-Ni	0.906
41	BiTmNi-Ni	0.934
42	BiYNi-Ni	0.918
43	ZrBiNi-Ni	0.955
44	BiPrPd-Pd	0.900
45	BiSmPd-Pd	0.914
46	BiTbPd-Pd	0.904
47	BiYbPd-Pd	0.902
48	BiTbPt-Pt	0.881
49	BiYPt-Pt	0.905
50	BiYbPt-Pt	0.882
51	ZrBiRh-Rh	0.931
52	CuCdSb-Sb	0.881
53	HfSbCo-Co	0.888
54	MnCoSb-Sb	0.373
55	SnNbCo-Co	0.950
56	TaSbCo-Co	0.900
57	SbVCo-Co	0.881
58	ZrSbCo-Co	0.881
59	TiSnCo-Co	0.951
60	MgPbCu-Cu	0.888
61	CuMgSn-Sn	0.881
62	SbMnCu-Cu	0.881
63	SbDyPd-Pd	0.946
64	SbDyPt-Pt	0.926
65	SbErPd-Pd	0.949
66	SbErPt-Pt	0.930
67	TiSbFe-Fe	0.930
68	VSbFe-Fe	0.881
69	SnTiFe-Fe	0.976
70	IrMnGa-Ga	0.881

Training Set, Class 1

no.	sample	correctness probability
71	RuMnGa-Ga	0.881
72	GaTiRh-Rh	0.881
73	GdPtSb-Sb	0.081
74	LiInGe-Ge	0.881
75	HfSbNi-Ni	0.966
76	HfSnNi-Ni	0.989
77	SnHfPt-Pt	0.981
78	HfSbRh-Rh	0.948
79	HoSbNi-Ni	0.935
80	SbHoPd-Pd	0.947
81	LiInSn-Sn	0.881
82	MgLiP-P	0.954
83	LiZnN-N	0.968
84	ZnLiP-P	0.881
85	LuSbNi-Ni	0.937
86	LuSbPd-Pd	0.946
87	NiMgSb-Sb	0.140
88	SbMgPd-Pd	0.900
89	SbMnNi-Ni	0.881
90	PdMnSb-Sb	0.680
91	MnPtSb-Sb	0.663
92	PtMnSn-Sn	0.881
93	NbSbRh-Rh	0.881
94	SnNbRh-Rh	0.881
95	SbScNi-Ni	0.955
96	TbSbNi-Ni	0.923
97	TmSbNi-Ni	0.935
98	VSbNi-Ni	0.971
99	SbYbNi-Ni	0.900
100	SbZnNi-Ni	0.637
101	SnTiNi-Ni	0.979
102	USnNi-Ni	0.986
103	SbScPd-Pd	0.954
104	SbTmPd-Pd	0.945
105	YbSbPd-Pd	0.919
106	SnZrPd-Pd	0.959
107	SbSmPt-Pt	0.903
108	SbTbPt-Pt	0.917
109	SbYPt-Pt	0.916
110	SbYbPt-Pt	0.897
111	SnThPt-Pt	0.960
112	SnTiPt-Pt	0.881

no.	sample	correctness probability
113	SnZrPt-Pt	0.971
114	ThSbRh-Rh	0.894
115	USbRh-Rh	0.881
116	SbZrRh-Rh	0.913
117	TaSbRu-Ru	0.926
118	TiSbRu-Ru	0.881
119	ZrSbRu-Ru	0.907

Training Set, Class 2

no.	sample	correctness probability
120	MgAgAs-Ag	0.022
121	ZnAgAs-Ag	0.022
122	ZnAgAs-Zn	0.006
123	AgCdSb-Ag	0.036
124	SbMgAg-Mg	0.030
125	AlBeB-Be	0.077
126	AlBeB-Al	0.021
127	CrAlCo-Cr	0.051
128	LiAlGe-Al	0.077
129	AlLiSi-Li	0.012
130	AlLiSi-Al	0.030
131	CdLiAs-Cd	0.017
132	MgLiAs-Li	0.029
133	MnLiAs-Li	0.076
134	MnLiAs-Mn	0.021
135	ZnLiAs-Zn	0.038
136	ZnNaAs-Na	0.032
137	CaBiAu-Bi	0.057
138	CaBiAu-Ca	0.067
139	BiYbAu-Bi	0.077
140	AuCdSb-Cd	0.007
141	DyPbAu-Pb	0.077
142	DyPbAu-Dy	0.022
143	ErPbAu-Er	0.020
144	SnErAu-Er	0.058
145	GdPbAu-Pb	0.044
146	GdPbAu-Gd	0.026
147	HoPbAu-Ho	0.024
148	SnHoAu-Ho	0.055
149	LiAuSb-Au	0.486
150	LiAuSb-Li	0.077
151	SnLuAu-Sn	0.029
152	SnMgAu-Mg	0.021
153	AuMnSb-Mn	0.007
154	AuMnSb-Au	0.183
155	SnMnAu-Sn	0.077
156	TbPbAu-Pb	0.033
157	YPbAu-Pb	0.046
158	YPbAu-Y	0.031
159	ScSnAu-Sc	0.057
160	SnTmAu-Tm	0.043

no.	sample	correctness probability
161	CeBiPd-Bi	0.045
162	CeBiPd-Ce	0.038
163	BiCePt-Bi	0.037
164	ZrBiCo-Bi	0.022
165	MgBiCu-Bi	0.107
166	MgBiCu-Mg	0.072
167	BiDyNi-Bi	0.021
168	BiDyPt-Dy	0.041
169	BiErNi-Er	0.046
170	BiErNi-Bi	0.014
171	BiErPd-Bi	0.020
172	BiErPt-Er	0.048
173	ZnFeBi-Fe	0.106
174	ZnFeBi-Zn	0.077
175	BiGdNi-Bi	0.015
176	BiGdPd-Gd	0.043
177	BiGdPt-Gd	0.041
178	BiGdPt-Bi	0.027
179	BiHoNi-Bi	0.018
180	PdHoBi-Ho	0.050
181	BiHoPt-Ho	0.047
182	BiHoPt-Bi	0.026
183	BiLaPd-Bi	0.072
184	BiLaPt-La	0.034
185	MgLiBi-Li	0.064
186	MgLiBi-Mg	0.077
187	BiVLi-Bi	0.057
188	BiLuNi-Lu	0.051
189	BiLuPd-Lu	0.052
190	BiLuPd-Bi	0.025
191	BiLuPt-Bi	0.023
192	BiMgNi-Mg	0.077
193	BiNdNi-Nd	0.027
194	BiNdNi-Bi	0.077
195	PtNdBi-Pt	0.852
196	BiPrNi-Pr	0.034
197	BiScNi-Sc	0.069
198	BiScNi-Bi	0.012
199	BiSmNi-Bi	0.026
200	BiTbNi-Tb	0.040
201	BiTmNi-Tm	0.041

no.	sample	correctness probability
202	BiTmNi-Bi	0.024
203	BiYNi-Bi	0.015
204	BiZnNi-Zn	0.037
205	ZrBiNi-Bi	0.012
206	ZrBiNi-Zr	0.049
207	BiPrPd-Bi	0.051
208	BiScPd-Sc	0.077
209	BiSmPd-Sm	0.039
210	BiSmPd-Bi	0.035
211	BiTbPd-Bi	0.025
212	BiTmPd-Tm	0.046
213	BiYbPd-Yb	0.054
214	BiYbPd-Bi	0.017
215	BiTbPt-Bi	0.023
216	BiTmPt-Tm	0.044
217	BiYPt-Y	0.047
218	BiYPt-Bi	0.027
219	BiYbPt-Bi	0.020
220	TmBiRh-Bi	0.034
221	ZrBiRh-Bi	0.006
222	ZrBiRh-Zr	0.067
223	CuCdSb-Cu	0.422
224	LiCdP-Cd	0.045
225	HfSbCo-Sb	0.030
226	HfSbCo-Hf	0.048
227	MnCoSb-Mn	0.004
228	NbSbCo-Sb	0.026
229	SnNbCo-Nb	0.043
230	SnNbCo-Sn	0.023
231	TaSbCo-Ta	0.050
232	TiSbCo-Sb	0.077
233	SbVCo-V	0.030
234	SbVCo-Sb	0.051
235	ZrSbCo-Zr	0.048
236	TaSnCo-Sn	0.020
237	TiSnCo-Sn	0.039
238	TiSnCo-Ti	0.030
239	MgPbCu-Mg	0.038
240	SbMgCu-Mg	0.035
241	CuMgSn-Mg	0.024
242	CuMgSn-Cu	0.865
243	SbMnCu-Sb	0.657

no.	sample	correctness probability
244	DySbNi-Sb	0.051
245	SbDyPd-Dy	0.035
246	SbDyPd-Sb	0.077
247	SbDyPt-Sb	0.077
248	ErSbNi-Sb	0.029
249	SbErPd-Er	0.042
250	SbErPd-Sb	0.044
251	SbErPt-Sb	0.054
252	NbSbFe-Sb	0.017
253	TiSbFe-Sb	0.051
254	TiSbFe-Ti	0.039
255	VSbFe-V	0.019
256	ZnFeSb-Fe	0.077
257	SnTiFe-Ti	0.013
258	SnTiFe-Sn	0.075
259	IrMnGa-Ir	0.076
260	PtMnGa-Mn	0.005
261	RuMnGa-Mn	0.028
262	RuMnGa-Ru	0.040
263	GaTiRh-Ga	0.077
264	GdSbNi-Sb	0.037
265	GdPtSb-Pt	0.920
266	GdPtSb-Gd	0.037
267	LiInGe-Li	0.077
268	TiGePt-Ge	0.304
269	HfSbNi-Sb	0.013
270	HfSbNi-Hf	0.043
271	HfSnNi-Hf	0.036
272	HfSnPd-Sn	0.077
273	SnHfPt-Hf	0.028
274	SnHfPt-Sn	0.006
275	HfSbRh-Hf	0.061
276	HfSbRu-Sb	0.008
277	HoSbNi-Sb	0.042
278	HoSbNi-Ho	0.037
279	SbHoPd-Sb	0.059
280	SbHoPt-Ho	0.039
281	LiInSn-In	0.077
282	LiInSn-Li	0.077
283	MgLiP-Mg	0.070
284	MgLiSb-Li	0.077
285	LiZnN-Zn	0.027

no.	sample	correctness probability
286	LiZnN-Li	0.030
287	ZnLiP-Zn	0.077
288	LiSbV-Sb	0.077
289	LuSbNi-Sb	0.026
290	LuSbNi-Lu	0.045
291	LuSbPd-Lu	0.046
292	LuSbPt-Sb	0.065
293	NiMgSb-Mg	0.046
294	NiMgSb-Ni	0.926
295	SbMgPd-Sb	0.077
296	SbMgPt-Mg	0.057
297	SbMnNi-Mn	0.003
298	SbMnNi-Sb	0.325
299	PdMnSb-Pd	0.023
300	TeMnPd-Mn	0.006
301	MnPtSb-Pt	0.089
302	MnPtSb-Mn	0.000
303	PtMnSn-Pt	0.078
304	MnRhSb-Rh	0.077
305	NbSbRh-Sb	0.010
306	NbSbRh-Nb	0.007
307	SnNbRh-Sn	0.032
308	NbSbRu-Sb	0.006
309	SbScNi-Sc	0.060
310	SbScNi-Sb	0.023
311	TbSbNi-Tb	0.035
312	TiSbNi-Sb	0.043
313	TmSbNi-Sb	0.061
314	TmSbNi-Tm	0.033
315	VSbNi-V	0.004
316	YSbNi-Sb	0.033
317	SbYbNi-Yb	0.028
318	SbYbNi-Sb	0.067
319	SbZnNi-Sb	0.413
320	ThSnNi-Sn	0.020
321	SnTiNi-Ti	0.014
322	SnTiNi-Sn	0.077
323	USnNi-U	0.016
324	ZrSnNi-Sn	0.023
325	SbScPd-Sc	0.070
326	SbScPd-Sb	0.026
327	SbTmPd-Sb	0.072

no.	sample	correctness probability
328	SbYPd-Y	0.043
329	YbSbPd-Sb	0.074
330	YbSbPd-Yb	0.034
331	SnZrPd-Sn	0.058
332	SbScPt-Sc	0.066
333	SbSmPt-Sm	0.028
334	SbSmPt-Sb	0.077
335	SbTbPt-Sb	0.077
336	SbTmPt-Tm	0.035
337	SbYPt-Y	0.042
338	SbYPt-Sb	0.077
339	SbYbPt-Sb	0.055
340	SnScPt-Sc	0.053
341	SnThPt-Th	0.028
342	SnThPt-Sn	0.010
343	SnTiPt-Sn	0.076
344	SnUPt-U	0.013
345	SnZrPt-Zr	0.030
346	SnZrPt-Sn	0.005
347	ThSbRh-Th	0.035
348	SbTiRh-Ti	0.023
349	USbRh-Sb	0.010
350	USbRh-U	0.017
351	SbZrRh-Sb	0.009
352	SnTiRh-Ti	0.045
353	TaSbRu-Sb	0.004
354	TaSbRu-Ta	0.011
355	TiSbRu-Ti	0.024
356	VSbRu-Sb	0.023
357	ZrSbRu-Sb	0.007
358	ZrSbRu-Zr	0.052

Validation Set, Class 1

no.	sample	correctness probability
359	MgAgAs-As	0.899
360	SbMgAg-Ag	0.106
361	LiAlGe-Ge	0.884
362	MgLiAs-As	0.950
363	ZnNaAs-As	0.855
364	AuCdSb-Sb	0.811
365	SnErAu-Au	0.877
366	SnHoAu-Au	0.944
367	SnMgAu-Au	0.726
368	TbPbAu-Au	0.964
369	SnTmAu-Au	0.896
370	ZrBiCo-Co	0.886
371	BiDyPt-Pt	0.909
372	BiErPt-Pt	0.926
373	BiGdPd-Pd	0.920
374	PdHoBi-Bi	0.026
375	BiLaPt-Pt	0.866
376	BiLuNi-Ni	0.937
377	BiMgNi-Ni	0.914
378	BiPrNi-Ni	0.886
379	BiTbNi-Ni	0.898
380	BiZnNi-Ni	0.550
381	BiScPd-Pd	0.951
382	BiTmPd-Pd	0.940
383	BiTmPt-Pt	0.925
384	TmBiRh-Rh	0.909
385	LiCdP-P	0.890
386	NbSbCo-Co	0.906
387	TiSbCo-Co	0.899
388	TaSnCo-Co	0.945

no.	sample	correctness probability
389	SbMgCu-Cu	0.778
390	DySbNi-Ni	0.935
391	ErSbNi-Ni	0.937
392	NbSbFe-Fe	0.932
393	ZnFeSb-Sb	0.505
394	PtMnGa-Ga	0.684
395	GdSbNi-Ni	0.930
396	TiGePt-Pt	0.051
397	HfSnPd-Pd	0.966
398	HfSbRu-Ru	0.944
399	SbHoPt-Pt	0.928
400	MgLiSb-Sb	0.787
401	LiSbV-V	0.530
402	LuSbPt-Pt	0.928
403	SbMgPt-Pt	0.906
404	TeMnPd-Pd	0.061
405	MnRhSb-Sb	0.717
406	NbSbRu-Ru	0.893
407	TiSbNi-Ni	0.974
408	YSbNi-Ni	0.928
409	ThSnNi-Ni	0.967
410	ZrSnNi-Ni	0.987
411	SbYPd-Pd	0.939
412	SbScPt-Pt	0.942
413	SbTmPt-Pt	0.927
414	SnScPt-Pt	0.977
415	SnUPt-Pt	0.960
416	SbTiRh-Rh	0.912
417	SnTiRh-Rh	0.924
418	VSbRu-Ru	0.563

Validation Set, Class 2

no.	sample	correctness probability
419	MgAgAs-Mg	0.023
420	AgCdSb-Cd	0.004
421	SbMgAg-Sb	0.860
422	CrAlCo-Al	0.430
423	LiAlGe-Li	0.016
424	CdLiAs-Li	0.096
425	MgLiAs-Mg	0.032
426	ZnLiAs-Li	0.073
427	ZnNaAs-Zn	0.061
428	BiYbAu-Yb	0.053
429	AuCdSb-Au	0.411
430	ErPbAu-Pb	0.040
431	SnErAu-Sn	0.035
432	HoPbAu-Pb	0.064
433	SnHoAu-Sn	0.049
434	SnLuAu-Lu	0.061
435	SnMgAu-Sn	0.080
436	SnMnAu-Mn	0.011
437	TbPbAu-Tb	0.026
438	ScSnAu-Sn	0.028
439	SnTmAu-Sn	0.049
440	BiCePt-Ce	0.036
441	ZrBiCo-Zr	0.061
442	BiDyNi-Dy	0.039
443	BiDyPt-Bi	0.031
444	BiErPd-Er	0.049
445	BiErPt-Bi	0.023
446	BiGdNi-Gd	0.041
447	BiGdPd-Bi	0.031
448	BiHoNi-Ho	0.044
449	PdHoBi-Pd	0.931
450	BiLaPd-La	0.034
451	BiLaPt-Bi	0.061
452	BiVLi-V	0.591
453	BiLuNi-Bi	0.013
454	BiLuPt-Lu	0.051
455	BiMgNi-Bi	0.041
456	PtNdBi-Nd	0.026
457	BiPrNi-Bi	0.020
458	BiSmNi-Sm	0.034
459	BiTbNi-Bi	0.014

no.	sample	correctness probability
460	BiYNi-Y	0.047
461	BiZnNi-Bi	0.114
462	BiPrPd-Pr	0.035
463	BiScPd-Bi	0.012
464	BiTbPd-Tb	0.042
465	BiTmPd-Bi	0.029
466	BiTbPt-Tb	0.041
467	BiTmPt-Bi	0.030
468	BiYbPt-Yb	0.048
469	TmBiRh-Tm	0.060
470	CuCdSb-Cd	0.000
471	LiCdP-Li	0.051
472	MnCoSb-Co	0.745
473	NbSbCo-Nb	0.029
474	TaSbCo-Sb	0.016
475	TiSbCo-Ti	0.044
476	ZrSbCo-Sb	0.061
477	TaSnCo-Ta	0.070
478	MgPbCu-Pb	0.096
479	SbMgCu-Sb	0.529
480	SbMnCu-Mn	0.004
481	DySbNi-Dy	0.033
482	SbDyPt-Dy	0.034
483	ErSbNi-Er	0.041
484	SbErPt-Er	0.042
485	NbSbFe-Nb	0.013
486	VSbFe-Sb	0.041
487	ZnFeSb-Zn	0.010
488	IrMnGa-Mn	0.019
489	PtMnGa-Pt	0.033
490	GaTiRh-Ti	0.262
491	GdSbNi-Gd	0.036
492	LiInGe-In	0.012
493	TiGePt-Ti	0.044
494	HfSnNi-Sn	0.022
495	HfSnPd-Hf	0.024
496	HfSbRh-Sb	0.009
497	HfSbRu-Hf	0.047
498	SbHoPd-Ho	0.040
499	SbHoPt-Sb	0.058
500	MgLiP-Li	0.023

no.	sample	correctness probability
501	MgLiSb-Mg	0.136
502	ZnLiP-Li	0.045
503	LiSbV-Li	0.173
504	LuSbPd-Sb	0.057
505	LuSbPt-Lu	0.046
506	SbMgPd-Mg	0.049
507	SbMgPt-Sb	0.119
508	PdMnSb-Mn	0.000
509	TeMnPd-Te	0.098
510	PtMnSn-Mn	0.000
511	MnRhSb-Mn	0.008
512	SnNbRh-Nb	0.014
513	NbSbRu-Nb	0.006
514	TbSbNi-Sb	0.039
515	TiSbNi-Ti	0.030
516	VSbNi-Sb	0.043
517	YSbNi-Y	0.041
518	SbZnNi-Zn	0.006
519	ThSnNi-Th	0.030

no.	sample	correctness probability
520	USnNi-Sn	0.023
521	ZrSnNi-Zr	0.038
522	SbTmPd-Tm	0.035
523	SbYPd-Sb	0.071
524	SnZrPd-Zr	0.025
525	SbScPt-Sb	0.026
526	SbTbPt-Tb	0.035
527	SbTmPt-Sb	0.062
528	SbYbPt-Yb	0.033
529	SnScPt-Sn	0.052
530	SnTiPt-Ti	0.007
531	SnUPt-Sn	0.011
532	ThSbRh-Sb	0.050
533	SbTiRh-Sb	0.041
534	SbZrRh-Zr	0.073
535	SnTiRh-Sn	0.111
536	TiSbRu-Sb	0.023
537	VSbRu-V	0.016

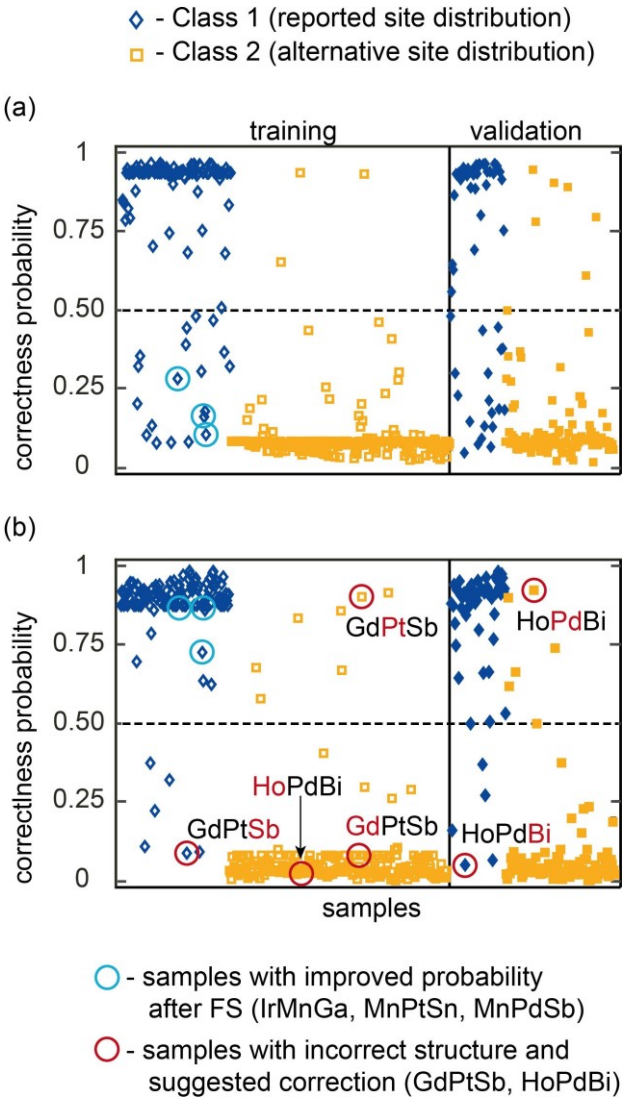


Figure S1. Highlighted samples on prediction probability figure.