JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Information-Theoretic Approach for the Discovery of Design Rules for Crystal Chemistry

Chang Sun Kong,[†] Wei Luo,[†] Sergiu Arapan,[‡,⊥] Pierre Villars,[§] Shuichi Iwata,[∥] Rajeev Ahuja,[‡,#] and Krishna Rajan*[,†]

[†]Department of Materials Science and Engineering, Iowa State University, Ames, Iowa 50011, United States

[‡]Condensed Matter Theory Group, Department of Physics and Astronomy, Uppsala University, Box 530, S-751 21 Uppsala, Sweden

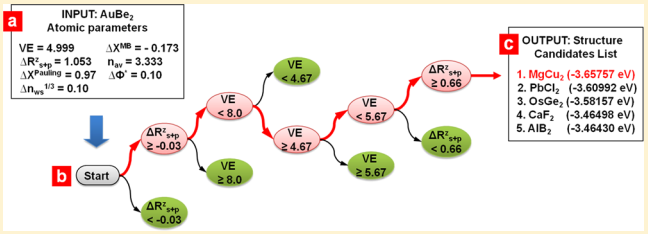
[⊥]Institute of Electronic Engineering and Industrial Technologies, Academy of Sciences of Moldova, Academiei 3/3, MD-2028 Chişinău, Moldova

[§]Materials Phases Data System, Schwanden 400, Vitznau CH-6354, Switzerland

[∥]The Graduate School of Project Design, 3-13-16 Minami-aoyama, Minato-ku, Tokyo 107-8411, Japan

[#]Applied Materials Physics, Department of Materials Science and Engineering, Royal Institute of Technology, SE-100 44 Stockholm, Sweden

ABSTRACT: In this work, it is shown that for the first time that, using information-entropy-based methods, one can quantitatively explore the relative impact of a wide multidimensional array of electronic and chemical bonding parameters on the structural stability of intermetallic compounds. Using an inorganic AB_2 compound database as a template data platform, the evolution of design rules for crystal chemistry based on an information-theoretic partitioning classifier for a high-dimensional manifold of crystal chemistry descriptors is monitored. An application of this data-mining approach to establish chemical and structural design rules for crystal chemistry is demonstrated by showing that, when coupled with first-principles calculations, statistical inference methods can serve as a tool for significantly accelerating the prediction of unknown crystal structures.



1. INTRODUCTION

One of the great questions in materials crystal chemistry is, why do atoms arrange themselves in the way they do? To address this challenge, the concept of phase homologies has been widely used. Homologous series of compounds are those that seem chemically diverse but can be expressed in terms of a mathematical formula that is capable of producing each chemical member in a particular crystal structure "type". A well-established strategy to help discover new compounds—or at least to try to develop chemical design strategies for discovery—is to search, organize, and classify homologous compounds from known data. These classification schemes, often called structure maps, are developed with the hope that they can provide sufficient insight to help forecast, with some certainty, specific new phases or compounds. Yet, although the classification schemes (over a dozen have been reported in the past 50 years[1−22]) have proved to be instructive, mostly in hindsight, they have had limited impact, if at all, on the a priori design of materials chemistry.

The objective of this article is to show how one can transform the concept of homologous compounds from one of mapping phenomenological observations to one that can serve as a tool for learning about structure−chemistry relationships and identify, using statistical inference methods, the rules that appear to govern the structural stability of known compounds. We show that, from this information, one can identify and lay the foundations for structure prediction. For the purposes of this study, we focused on AB_2 compounds as a template to demonstrate the value of this approach.

2. METHODS

2.1. Data Compilation. In general, structure−property relationships are guided by precisely defined functional relationships (e.g., electronic structure calculations to define energy landscapes associated with crystal chemistry). Our data-driven methodology involves applying statistical learning tools to analyze correlations between numerous scalar descriptors of electronic and crystal structure parameters of known intermetallic compounds and using that information, in turn, to develop predictive models that can suggest new structures/chemistries and properties based purely on the formalism of statistical learning methods.

This methodology is quite different from the approach that is widely reported by many groups in which large numbers of high-throughput electronic structure computations are conducted to seek compound chemistries with energy minima (where data-mining-related techniques are embedded in the

Received: December 29, 2011
Published: July 1, 2012

   dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812−1820

computation to help the efficiency of the calculations), and then potentially new stable compounds are identified by identifying those that have energy minima but are not reported in known experimental databases.[23−29] In this study, we instead propose an approach to establish such a structure−property relationship where we do not assume any specific formulation linking structure with property. We take a data-driven approach in which we seek to establish structure−property relationships by identifying patterns of behavior between known discrete scalar descriptors associated with crystal and electronic structures and observed properties of the material. From this information, we extract design rules that allow the systematic identification of critical structure−property relationships resulting in the quantitative identification of the exact role of specific combinations of materials descriptors. In our group, we have applied this approach to explore a variety of questions associated with crystal chemistry.[30−34] In this article, we expand on our prior work to demonstrate that, by using tools of statistical inference, we can build design rules that correlate atomic attributes of constituents with the structural stability of compounds.

Our approach requires the careful establishment of a data set of descriptors on which statistical learning tools are directly applied. The number of parameters needed to predict even relatively simple structures can be large if one has to capture both geometrical and bonding characteristics of that crystal chemistry. Although the potential number of variables can, in fact, be large, data dimensionality reduction and information-theoretic techniques can help reduce it to a manageable number. This article describes a data-mining strategy from which effective classification models can be developed using high-dimensional information. The classification models, in turn, permit one to screen and rapidly target the few compounds of interest for subsequent computations to aid in structure prediction.

To build a classification model, the crystal structure data of intermetallic compounds were collected from the Linus Pauling File[35] (LPF). Of 103 elements (atomic numbers 1−103) in the periodic table, 68 elements were taken into account after excluding nonmetallic elements, namely, hydrogen, chalcogen, halogen, and noble-gas group elements and some of the lanthanide/actinide elements (i.e., Pm, Eu, Tb, Yb, Pa, Np, Am, Cm, Bk, Cf, Es, Fm, Md, No, and Lr) for which experimental data do not exist in the LPF. In addition, the data for compounds prepared at high temperature and pressure or at low temperature, for compounds stabilized by impurities, and for metastable and/or polymorphic states were not used for this study. In version 1.0 of the LPF, AB_2 binary compounds formed from the selected 68 elements have 973 entries that can be categorized into 109 structure-type classes. Among them, 840 compounds (86.3%) are represented by 34 main structure types, each of which has more than six entries. The remaining 133 compounds are classified into an additional 75 structure types. These 75 structure types (i.e., minor structure types), each of which has less than five compounds as members, were excluded from the data set.

2.2. Determination of Parameter Space. The selection of parameter sets is a critical step in the construction of a good predictive model. In this study, we chose seven parameters suggested by three separate classical studies from Mooser and Pearson, Miedema, and Villars[1−7] in their structure-map models. Detailed descriptions of the parameters are provided in Table 1. Based on these well-established criteria, we integrate the collective impact of these parameters and show how a predictive model can be established by applying the partition-based classifier and a high-dimensional coordinate system to a structure data manifold.

Table 1. Atomic and Physical Parameters for Crystal Structure Prediction[1−7]

parameter	description	model
VE	average number of valence electrons per atom	Villars
ΔX^{MB}	weighted difference of Martynov−Batsanov electronegativities	
ΔR^{Z}_{s+p}	weighted difference of the sum of Zunger pseudopotential radii	
n_{av}	average principal quantum number	Mooser and Pearson
$\Delta X^{Pauling}$	Pauling electronegativity difference	
$\Delta\Phi^*$	chemical potential difference for electronic charges	Miedema
$\Delta n_{ws}^{1/3}$	electron density difference in the Wigner−Seitz atomic cell	

2.3. Classification and Partitioning of High-Dimensional Data. In this study, we use the Shannon[36] information entropy function as a splitting criterion for the classification of crystal structure data. The information entropy, H, of a data set is a quantitative measure of the uncertainty and is mathematically defined as[36,37]

$$H = -\sum_{i=i}^{n} p(x_i) \log p(x_i) \tag{1}$$

where $p(x_i) \geq 0$, $\sum_{i=1}^{n} p(x_i) = 1$ $(i = 1, 2, ..., n)$. In eq 1, $p(x_i)$ is the likelihood of the occurrence of a particular structure type, x_i, in the given data set, in this case, the LPF database.

The aim of the classification is to reduce the uncertainty of the data set by grouping the observations (i.e., data) according to their similarities and thus to extract useful information from the data. In practice, the classification in this study, the partitioning of data manifolds, involves consecutively finding a discrete parameter value that minimizes the information entropy of the entire data set by dividing the parameter space into smaller subregions with a set of hyperplane boundaries. This partitioning consists of a series of recursive processes to seek the position of a hyperplane that bisects a portion of the parameter space to minimize the entropy. The classification tree built as the result of the partitioning connects the respective compounds to the crystal structures with the parameters and extracts useful knowledge from the causal linkage between the parameters and crystal structures. Initially, the respective compounds $C\{c_1, c_2, ..., c_m\}$ in a data set (here, m = 840) in which each compound is characterized by the parameters $V = V(v_1, v_2, ..., v_l)$ (here, l = 7) are connected to one of the classes (i.e., structure types) $X\{x_1, x_2, ..., x_n\}$ (here, the number of structure types is n = 34). Then, the entropy criterion applied to the partitioning is

$$x_i(m) = \underset{xi}{\text{argmax}}[p_m(x_i)] \tag{2}$$

where $x_i(m)$ is a major class (i.e., structure type) in node m, $p_m(x_i)$ is the probability of class x_i in node m, and argmax denotes the argument of the maximum. That is, the class homogeneity in each node is increased as the classification tree grows and finally is reached at the maximum state. When this classification as a mapping (M) process, M: $C(V) \rightarrow X$, is correctly performed, the uncovered regularities in the data are developed as sets of if−then rules.

dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812−1820

Starting from the root, or single-group node, the compounds are divided into two offspring subgroups according to the "cutting" value of the parameters (Figure 1). These numeric

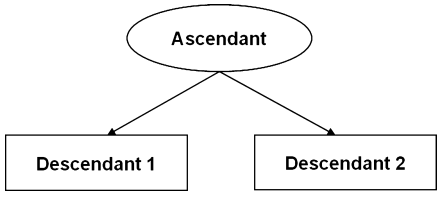


Figure 1. Classification tree description of recursive partitioning.

constraint values that make the best bisection of the parameter space are achieved by maximizing the reduction of entropy (H) before and after a partition step. Thus, the information gain (IG), ΔH (i.e., the goodness-of-split that can be achieved by a partition step), is defined as

$$\Delta H = H_{\text{ascendant}} - (p_1 H_{\text{descendant},1} + p_2 H_{\text{descendant},2}) \quad (3)$$

where H is the information entropy of the data at a given level of the tree defined by eq 1; p_1 and p_2 are the fractions of descendant nodes 1 and 2, respectively, and thus, $p_1 + p_2 = 1$.

Although the grouping turns into more homogeneous subsets as the partition proceeds, the splitting of the tree should be stopped at a certain level to prevent overfitting. That is, when no improvement of the model performance is observed, the bifurcation step is terminated.[38] This pruning procedure is highly significant as the predictability of a tree model is not only evaluated during the pruning but also directly related to the number of the structure types suggested by the classification model and thus to the computation time for the ab initio calculations. The pruning optimization is implemented until the prediction reliability is not improved during the validation test. We assessed our model by a random-sampling-based cross-validation method; specifically, one-tenth of the compound data was randomly selected and set aside for the test from the entire 840 compounds list. Then, a classification tree was formed with the remainder, called the training data set. Finally, we checked, using the test data, whether the predictive model (i.e., classification tree) correctly suggests the possible structure type. This procedure was repeated 10 times with random sampling. The validation process was carried out at various tree levels, and the partitioning was stopped at the level at which the prediction error rate became the lowest. Once the prediction model of if–then rules is constructed, the crystal structure of a new compound can be estimated by simply applying the model.

2.4. Energy Calculation by Density Functional Theory.

Ab initio calculations of the electronic structure were performed within the framework of the density functional theory (DFT) using the projector-augmented-wave (PAW) method as implemented in the Vienna Ab Initio Simulation Package (VASP).[39] We used PAW potentials[40,41] derived

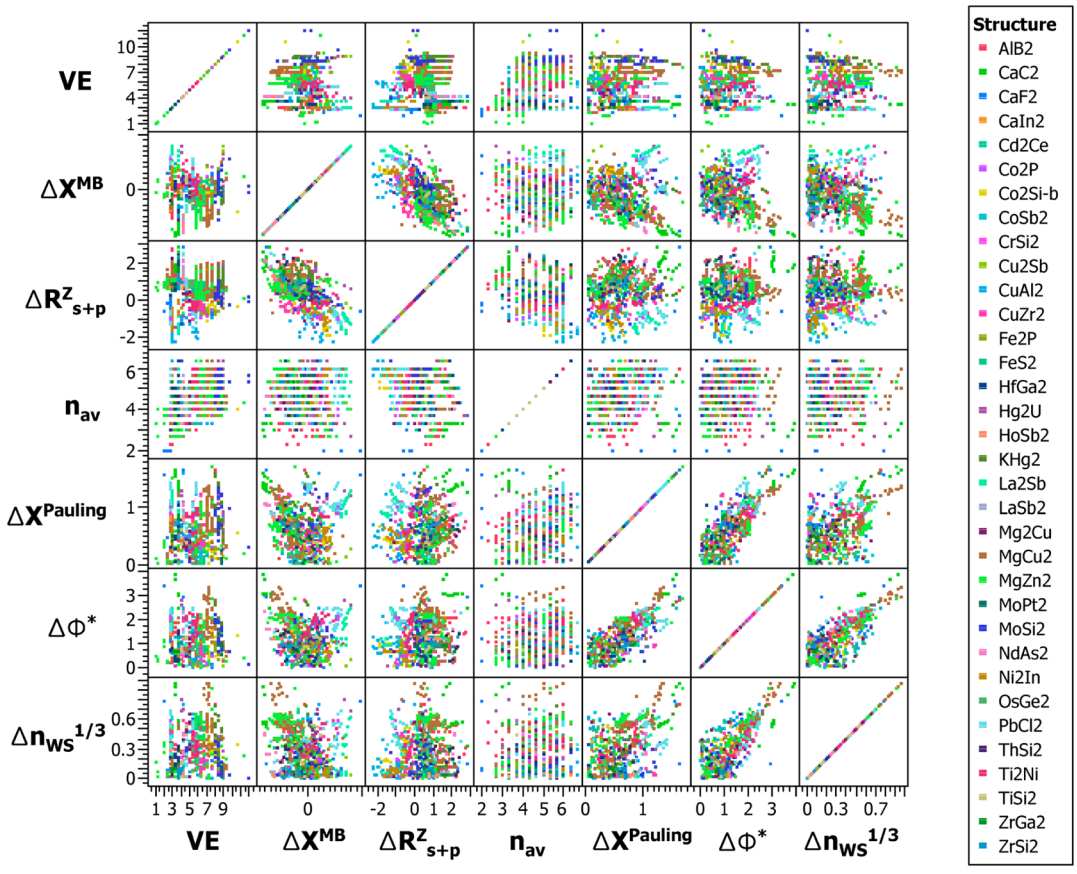


Figure 2. Two-dimensional feature spaces of AB_2-type inorganic compounds. The coordinates of each plot are atomic, physical parameters for the compounds originating from their constituent elements. The color-coded points on the plot represent 840 AB_2 compounds that are classified into 34 different crystal-structure types. Each subplot indicates two-dimensional structure-map feature space constructed by two of seven parameters. For the definition of each parameter, see Table 1.

dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812–1820

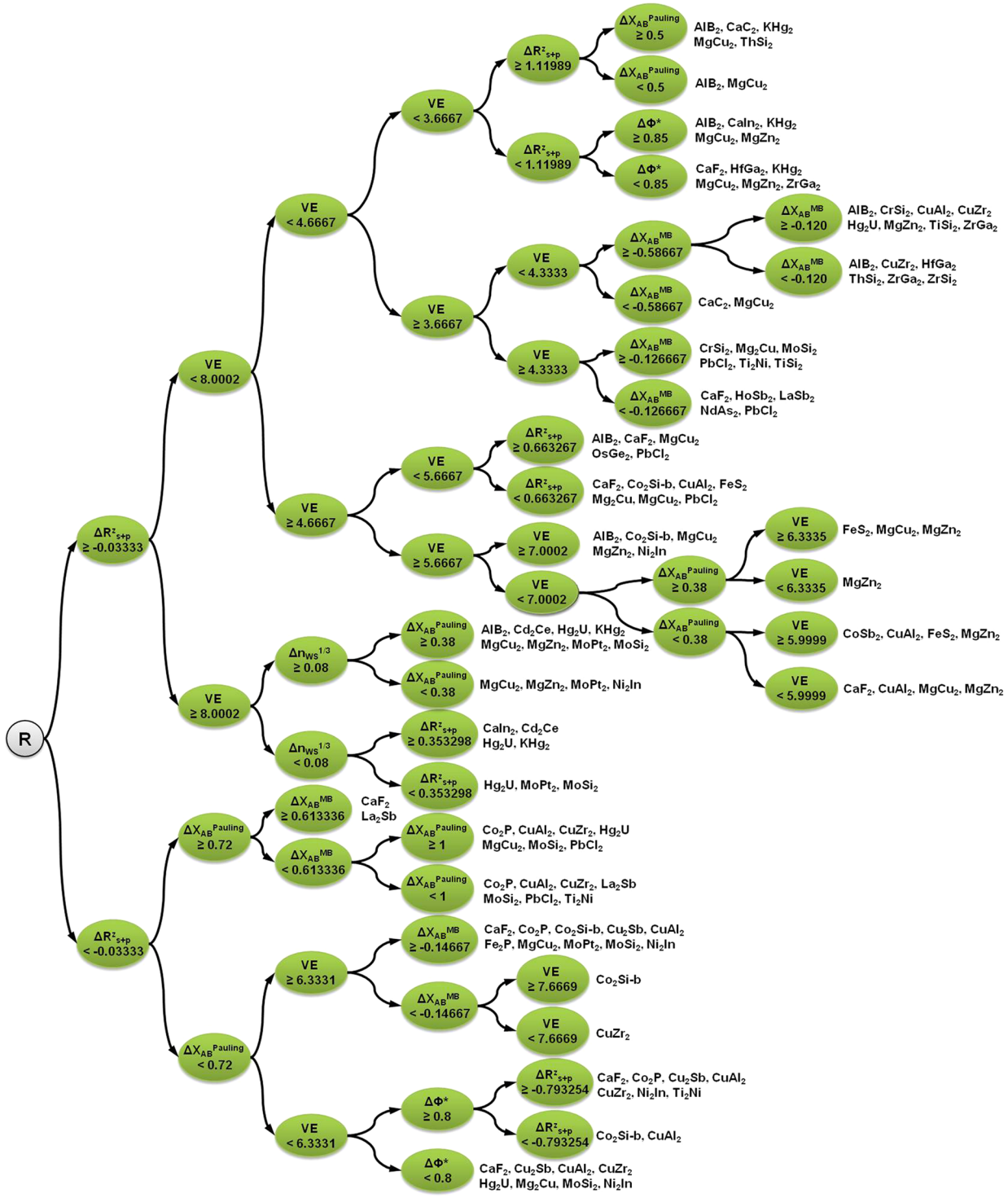


Figure 3. Classification tree for the crystal structures of AB_2 compounds. By a series of recursive partitioning, this tree structure with corresponding if−then classification rules was achieved. At the terminal leaves, a few crystal structures (prototypes) are suggested as the possible stable structures. Note that only two to four parameters are variably used in the series of if−then rules for each branch; in most cases, the combination of atomic size, electrochemical, and valence-electron factors (after ref 4) is used for the partitioning.

within the generalized gradient approximation (GGA) description of the electronic exchange-correlation energy.[42] All structures were fully relaxed until the Hellman−Feynman forces acting on each ion became less than 10^{-3} eV/Å. To ensure accurate results during the structure optimization procedure, Kohn−Sham orbitals were expanded in a plane-wave basis to an energy cutoff 150% larger than the default energy cutoff provided by PAW potentials. We used the Monkhorst−Pack[43] scheme to generate an automatic *k*-mesh sampling of the Brillouin zone and carried out the integration in reciprocal space using Methfessel−Paxton[44,45] smearing during the relaxation and the linear tetrahedron method with Blöchl corrections[46] for the relaxed structures. For all structures,

dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812−1820

convergence within 10^{-3}eV/ion of the total energy with respect to the number of k points was achieved.

3. RESULTS AND DISCUSSION

3.1. Extracting Design Rules for Crystal Chemistry. Our objective was to extract information based on the apparent cause and effect of the characteristics of the atomic species in a given compound from the data present in existing structure maps. We employed an informatics-based structure mapping for the extraction of materials design rules. A model using a multidimensional partitioning-based classifier was constructed from the crystal structure data of 840 AB_2 binary compounds and seven physical parameters. The AB_2 intermetallics include some important crystal structures for inorganic compounds such as AlB_2, CaF_2, and Laves phases.

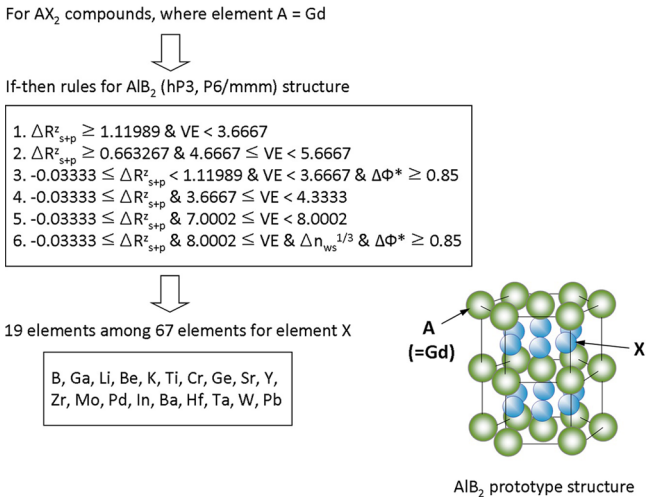


Figure 4. Selection of constituent elements from classification rules for crystal structure design of GdX_2. To obtain a specific crystal structure type, the classification rules (i.e., if−then rules) can be utilized for the selection of appropriate materials system. As an element, A, is predetermined, the possible partners for X element can be selected using if−then classification rules achieved from known structure data. Here, boron and gallium are already known in the data as forming AlB_2 structure with gadolinium, namely, GdB_2 and $GdGa_2$, respectively. However, the crystal structures of the other 17 elements are not experimentally known.

Generally, structure classification is based on the a priori selection of a given set of rules. However, we made no such a priori assumption and simultaneously selected all seven, or more, if necessary, parameters proposed for different types of compound systems. It should be noted that different types of electronegativity scales were used. The physical meanings of the respective definitions of electronegativity scales can be understood from the units of the various parameters. Pauling electronegativity, $\Delta X^{\text{Pauling}}$, has units of $(\text{energy})^{1/2}$, whereas the Martynov−Batsanov scale, ΔX^{MB}, has units of (energy/valence electron), and Miedema's parameter $\Delta\Phi^*$ has units of (volt), similarly to a work function. According to Villars,[4] these seven parameters can be categorized into four different subgroups, namely, the atomic-size factor (for ΔR^{Z}_{s+p}), valence-electron factor (for VE), atomic-number factor (for n_{av}), and electrochemical factor (for ΔX^{MB}, $\Delta X^{\text{Pauling}}$, $\Delta\Phi^*$, and $\Delta n_{ws}^{1/3}$) groups. From the classification results achieved by the principle of minimum information entropy, it will be shown that different combinations of the parameters appear to serve as crystal chemistry rules influencing the stability of different crystal structure types.

3.2. Dimensionality of the Data Manifold. Removing the constraints of variable selection, the incorporation of all of the parameters used in classical structure maps introduces a level of complexity in identifying correlations. This can be appreciated by studying the scatter plots of all of the possible combinations of parameters mapping the positions of the different crystal structures for AB_2 compounds (Figure 2). Hence the challenge is to partition the resulting high-dimensional data geometry in a way that can capture a classification scheme in AB_2 compounds that a simple two-dimensional analysis does not. This was addressed by using a recursive partitioning strategy where a measure of information entropy was used as a criterion for the selection of key parameters.[36,47] The outcome of classification, that is, the relationships between different crystal structures and atomic parameters, can be represented in the form of a classification tree that shows distinct classes (structure types) and the corresponding parameter(s) governing the partitioning of chemical space.

As pointed out previously, when a suitable set of parameters is selected for a structure map, the compounds of the same structure are grouped together closely in the parameter space, and thus, each structure type occupies its distinct region, a structure domain, in the map. As the selected parameter space is subdivided into smaller subspaces by the adaptation of hyperplane boundaries, the uncertainty measured by the information entropy function of data is decreased. Because the possible position of hyperplanes is determined to minimize the information entropy, this classification corresponds to the recursive partitioning process to determine the boundaries between different (structure-type) classes maximizing the reduction of the entropy of entire data space.[38] The partitioning steps can be represented by the corresponding tree structure. This tree structure diagram provides the information on the classified groups at each level of the tree along with the splitting criteria for the boundaries between them. In the multidimensional space partitioning, a set of splitting conditions can be considered as the pathways to reach each compartment. Once the classification of the given data is completed, the crystal structure of any new material not explored elsewhere can be identified by investigating where the material is located in the subdivided parameter space.

In this study, AB_2 intermetallic compounds available in the LPF database were classified according to their structure types by correlating the parameters originating from three different structure-map models. The classification tree constructed is shown in Figure 3. Each node denotes the splitting condition of different structure types used to achieve the minimum information entropy at the given depth of the tree. Before the classification, all compounds are contained in the root node (R). At each step of the partition, a specific parameter (or attribute) is selected to subdivide the compounds into two subgroups according to their structure types while reducing the information entropy of the system. After a series of partition steps, the number of possible structure types for each branch is optimally minimized. In general, the tree is overgrown and then pruned to obtain the best subtree structure that provides the highest performance. The performance and validity of the tree model were evaluated by a cross-validation procedure (internal evaluation of a model within the given data set), as described in detail at the previous section. According to the cross-validation

Table 2. If−Then Rules for the Estimation of the Structural Stability of AB_2 Compounds

group	if (decision criteria)	then (possible structure-type candidates)
1	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.72 \rightarrow VE < 6.33 \rightarrow \Delta\Phi^* < 0.8$	CaF_2, Cu_2Sb, $CuAl_2$, $CuZr_2$, Hg_2U, Mg_2Cu, $MoSi_2$, Ni_2In
2	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.72 \rightarrow VE < 6.33 \rightarrow \Delta\Phi^* \geq 0.8 \rightarrow \Delta R^{Z}_{s+p} < -0.79$	Co_2Si-b, $CuAl_2$
3	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.72 \rightarrow VE < 6.33 \rightarrow \Delta\Phi^* \geq 0.8 \rightarrow \Delta R^{Z}_{s+p} \geq -0.79$	CaF_2, Co_2P, Cu_2Sb, $CuAl_2$, $CuZr_2$, Ni_2In, Ti_2Ni
4	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow VE \geq 6.33 \rightarrow \Delta X_{AB}{}^{MB} < -0.15 \rightarrow VE < 7.67$	$CuZr_2$
5	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow VE \geq 6.33 \rightarrow \Delta X_{AB}{}^{MB} < -0.15 \rightarrow VE \geq 7.67$	Co_2Si-b
6	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow VE \geq 6.33 \rightarrow \Delta X_{AB}{}^{MB} \geq -0.15$	CaF_2, Co_2P, Co_2Si-b, Cu_2Sb, $CuAl_2$, Fe_2P, $MgCu_2$, $MoPt_2$, $MoSi_2$, Ni_2In
7	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow \Delta X_{AB}{}^{MB} < 0.61 \rightarrow \Delta X_{AB}{}^{Pauling} < 1$	Co_2P, $CuAl_2$, $CuZr_2$, La_2Sb, $MoSi_2$, $PbCl_2$, Ti_2Ni
8	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow \Delta X_{AB}{}^{MB} < 0.61 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 1$	Co_2P, $CuAl_2$, $CuZr_2$, Hg_2U, $MgCu_2$, $MoSi_2$, $PbCl_2$
9	$\Delta R^{Z}_{s+p} < -1/30 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.72 \rightarrow \Delta X_{AB}{}^{MB} \geq 0.61$	CaF_2, La_2Sb
10	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE \geq 8 \rightarrow \Delta n_{ws}{}^{1/3} < 0.08 \rightarrow \Delta R^{Z}_{s+p} < 0.35$	Hg_2U, $MoPt_2$, $MoSi_2$
11	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE \geq 8 \rightarrow \Delta n_{ws}{}^{1/3} < 0.08 \rightarrow \Delta R^{Z}_{s+p} \geq 0.35$	$CaIn_2$, Cd_2Cu, Hg_2U, KHg_2
12	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE \geq 8 \rightarrow \Delta n_{ws}{}^{1/3} \geq 0.08 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.38$	$MgCu_2$, $MgZn_2$, $MoPt_2$, Ni_2In
13	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE \geq 8 \rightarrow \Delta n_{ws}{}^{1/3} \geq 0.08 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.38$	AlB_2, Cd_2Ce, Hg_2U, KHg_2, $MgCu_2$, $MgZn_2$, $MoPt_2$, $MoSi_2$
14	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE \geq 5 + 2/3 \rightarrow VE < 7 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.38 \rightarrow VE < 6$	CaF_2, $CuAl_2$, $MgCu_2$, $MgZn_2$
15	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE \geq 5 + 2/3 \rightarrow VE < 7 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.38 \rightarrow VE \geq 6$	$CoSb_2$, $CuAl_2$, FeS_2, $MgZn_2$
16	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE \geq 5 + 2/3 \rightarrow VE < 7 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.38 \rightarrow VE < 6.33$	$MgZn_2$
17	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE \geq 5 + 2/3 \rightarrow VE < 7 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.38 \rightarrow VE \geq 6.33$	FeS_2, $MgCu_2$, $MgZn_2$
18	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE \geq 5 + 2/3 \rightarrow VE \geq 7$	AlB_2, Co_2Si-b, $MgCu_2$, $MgZn_2$, Ni_2In
19	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE < 5 + 2/3 \rightarrow \Delta R^{Z}_{s+p} < 0.66$	CaF_2, Co_2Si-b, $CuAl_2$, FeS_2, Mg_2Cu, $MgCu_2$, $PbCl_2$
20	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE \geq 4 + 2/3 \rightarrow VE < 5 + 2/3 \rightarrow \Delta R^{Z}_{s+p} \geq 0.66$	AlB_2, CaF_2, $MgCu_2$, $OsGe_2$, $PbCl_2$
21	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE \geq 3 + 2/3 \rightarrow VE \geq 4 + 1/3 \rightarrow \Delta X_{AB}{}^{MB} < -0.13$	CaF_2, $HoSb_2$, $LaSb_2$, $NdAs_2$, $PbCl_2$
22	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE \geq 3 + 2/3 \rightarrow VE \geq 4 + 1/3 \rightarrow \Delta X_{AB}{}^{MB} \geq -0.13$	$CrSi_2$, Mg_2Cu, $MoSi_2$, $PbCl_2$, Ti_2Ni, $TiSi_2$
23	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE \geq 3 + 2/3 \rightarrow VE < 4 + 1/3 \rightarrow \Delta X_{AB}{}^{MB} < -0.59$	CaC_2, $MgCu_2$
24	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE \geq 3 + 2/3 \rightarrow VE < 4 + 1/3 \rightarrow \Delta X_{AB}{}^{MB} \geq -0.59 \rightarrow \Delta X_{AB}{}^{MB} < -0.12$	AlB_2, $CuZr_2$, $HfGa_2$, $ThSi_2$, $ZrGa_2$, $ZrSi_2$
25	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE \geq 3 + 2/3 \rightarrow VE < 4 + 1/3 \rightarrow \Delta X_{AB}{}^{MB} \geq -0.59 \rightarrow \Delta X_{AB}{}^{MB} \geq -0.12$	AlB_2, $CrSi_2$, $CuAl_2$, $CuZr_2$, Hg_2U, $MgZn_2$, $TiSi_2$, $ZrGa_2$
26	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE < 3 + 2/3 \rightarrow \Delta R^{Z}_{s+p} < 1.12 \rightarrow \Delta\Phi^* < 0.85$	CaF_2, $HfGa_2$, KHg_2, $MgCu_2$, $MgZn_2$, $ZrGa_2$
27	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE < 3 + 2/3 \rightarrow \Delta R^{Z}_{s+p} < 1.12 \rightarrow \Delta\Phi^* \geq 0.85$	AlB_2, $CaIn_2$, KHg_2, $MgCu_2$, $MgZn_2$
28	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE < 3 + 2/3 \rightarrow \Delta R^{Z}_{s+p} \geq 1.12 \rightarrow \Delta X_{AB}{}^{Pauling} < 0.5$	AlB_2, $MgCu_2$
29	$\Delta R^{Z}_{s+p} \geq -1/30 \rightarrow VE < 8 \rightarrow VE < 4 + 2/3 \rightarrow VE < 3 + 2/3 \rightarrow \Delta R^{Z}_{s+p} \geq 1.12 \rightarrow \Delta X_{AB}{}^{Pauling} \geq 0.5$	AlB_2, CaC_2, KHg_2, $MgCu_2$, $ThSi_2$

of the model that we constructed, 91.7% of the candidate lists suggested the correct structure type for the test compounds. Most of the compounds in this example could be classified with only two to four parameters, rather than requiring all seven parameters. Note that principal quantum number (n_{av}) is not even included in the classification tree in Figure 3. This indicates that the relatively significant crystal chemistry attributes are automatically detected during the classification of different types of compounds.

Once the physical constraints on the stable atomic configuration of the compounds are formulated with a set of parameters, a series of rules can be used for the selection of appropriate chemical components to design the crystal structure. Figure 4 shows one example of the application of crystal chemistry design rules described by if−then statements. For a given composition of AX_2, when an element is determined, for example, gadolinium (Gd, the A component), then the possible pair of X components for the formation of AlB_2 prototype structure can be selected according to the if−then rules for AlB_2 structure. Only 19 of 67 elements are expected to form AlB_2 structure with Gd. As mentioned previously, structure maps are a practical and effective approach that capture the relationships between the crystal structure of compounds and the relevant properties from their constituent elements. The regularities achieved by mapping inorganic compounds to their structure classes with the parameters become the rules of crystal chemistry. Given a stoichiometry, one can estimate the stable structure of a particular combination of elements and achieve a guideline for the development of new materials. The tree representation of if {the range of parameters}−then {structure types} rules as a multivariate structure map shows the pathways to achieve specific crystal structures that are restricted by a discrete range of parameters. In Table 2, the if−then rules to obtain stable structures of AB_2 compounds are summarized.

3.3. Structure Prediction. The prediction of the crystal structures of unknown compounds can be carried out by tracking the routes in the tree according to the "window" of criteria given at each node. This screening process to find the stable crystal structure of a compound extensively reduces the

dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812−1820

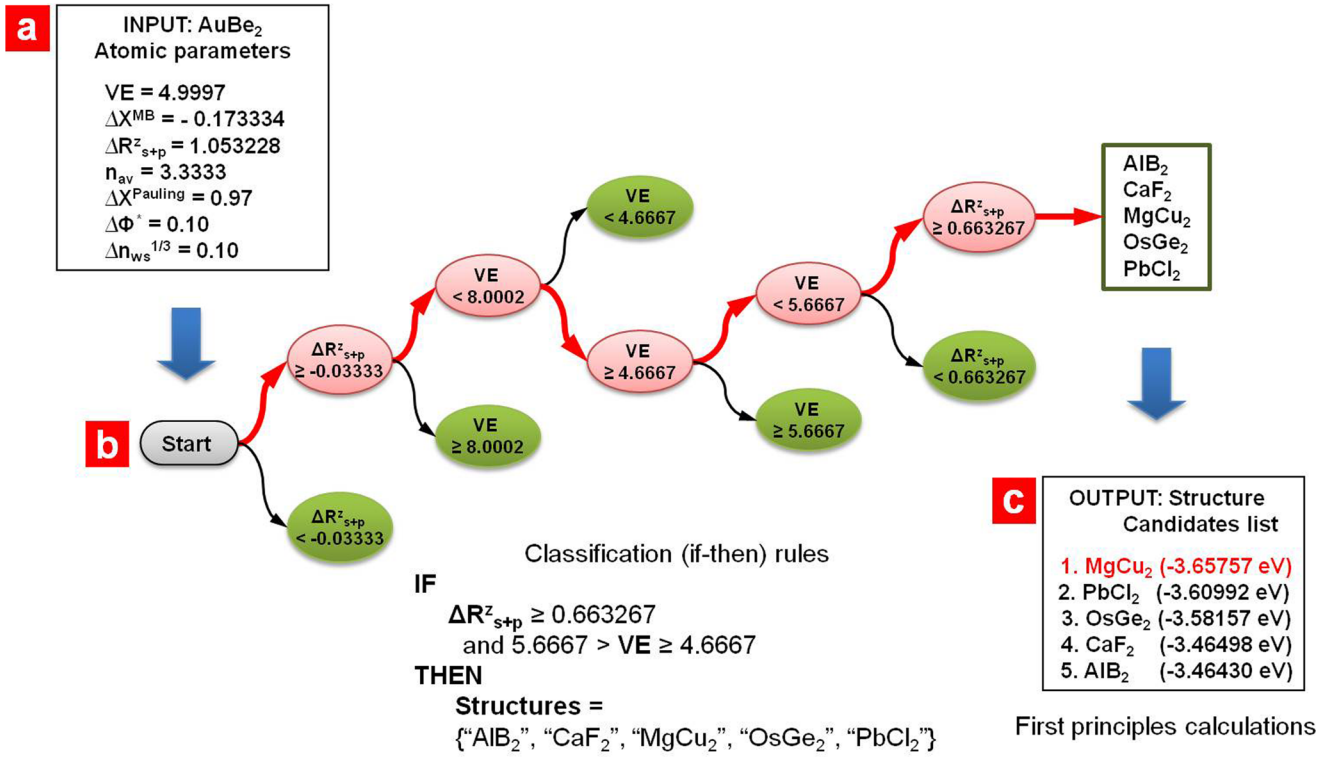


Figure 5. Prediction of the crystal structure of a hypothetical compound $AuBe_2$. By searching through a route according to the if−then rule, the possible structure types are nominated as the candidates. (1) The parameters calculated from atomic and physical properties of the constituents of $AuBe_2$ are used as the input variables for the structure prediction. (2) The classification tree suggests the possible structure-type candidates (the pathway indicated by red arrows). (3) A structure type with the lowest total energy, namely, $MgCu_2$, is confirmed as the most stable crystal structure by first-principles calculations.

Table 3. Prediction of the Crystal Structure (Prototype) of AB_2-Type Compounds

compound	structures estimated by classification tree	structure predicted by DFT
$AlRu_2$	**$CuZr_2$**, Cu_2Sb	$CuZr_2$
$FeMn_2$	**$CuZr_2$**, $CuAl_2$	$CuZr_2$
FeB_2	**$PbCl_2$**, CaF_2, $HoSb_2$, $LaSb_2$, $NdAs_2$	$PbCl_2$
Co_2Na	**$MgZn_2$**, FeS_2, $MgCu_2$	$MgZn_2$
$AuCa_2$	**$PbCl_2$**, Co_2P, $CuAl_2$, $CuZr_2$, Hg_2U, $MgCu_2$, $MoSi_2$	$PbCl_2$
Au_2Cr	**$PbCl_2$**, Co_2P, $CuAl_2$, $CuZr_2$, La_2Sb, $MoSi_2$, Ti_2Ni	$PbCl_2$
$CsGa_2$	**KHg_2**, AlB_2, CaC_2, CaF_2, $MgCu_2$, $ThSi_2$	KHg_2
$GdPd_2$	**Ni_2In**, AlB_2, Co_2Si-b, $MgCu_2$, $MgZn_2$	Ni_2In
$RuGe_2$	**$OsGe_2$**, AlB_2, CaF_2, $MgCu_2$, $PbCl_2$	$OsGe_2$

search space of ab initio calculations. In a classification tree model built using 840 compounds, five candidates, on average, are suggested as possible stable structure types for each test compound. Considering that we took 34 structure types into account for AB_2 compounds, the computational effort is reduced to approximately 15%. The most important aspect is that the underlying physical and chemical rules behind the atomic ordering of multielement systems can be explicitly expressed only with the parametrization of atomic-level governing factors, which facilitates the design of new materials.

Figure 5 illustrates how to find the most stable structure of a hypothetical compound, $AuBe_2$, by combining our classification-tree model and first-principles calculations. First, the parameters affecting the crystal structure of $AuBe_2$ are calculated from the atomic properties of the constituent elements. Then, the parameters of the test compound are applied to the criteria shown in the classification tree, which is tracked from the root node ("start" node in Figure 5) to the leaves. At the end of a branch, a list of the crystal structure types, AlB_2, CaF_2, $MgCu_2$, $OsGe_2$, and $PbCl_2$, is suggested as the most probable crystal structure of $AuBe_2$. That is, one of these five candidates is the most stable structure being sought. The results of the calculations suggest that the structure type $MgCu_2$, which has the lowest-energy state among the candidates, is the most stable structure for $AuBe_2$. In the same way, the most probable crystal structures of some compounds—experimentally known although the structure types have not yet been ascertained—have been identified, and the results are summarized in Table 3. (The identified most probable structures are highlighted in bold.)

4. CONCLUSIONS

This study serves to highlight the value of statistical inference methods in integrating diverse chemical and structural information to establish design rules for crystal chemistry. The use of Shannon entropy in developing classifiers among a multidimensional parameter space of descriptors provides a new approach to the characterization of homologous inorganic compounds. In the present study, the validation of our statistical-learning-based predictions is accomplished by first-principles calculations, but it clearly opens the door for guiding experimental structure determination studies and can be extended to other stoichiometries of multicomponent crystal chemistries.

dx.doi.org/10.1021/ci200628z | J. Chem. Inf. Model. 2012, 52, 1812−1820

AUTHOR INFORMATION

Corresponding Author

*E-mail: krajan@iastate.edu. Tel.: 515-294-2670.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge support from the National Science Foundation (DMS-1125909) and the Army Research Office (Grant W911NF-10-0397). K.R. also acknowledges support from Iowa State University through the Wilkinson Professorship in Interdisciplinary Engineering.

REFERENCES

(1) Villars, P. A Three-Dimensional Structural Stability Diagram for 998 Binary AB Intermetallic Compounds. *J. Less-Common Met.* **1983**, *92*, 215−238.

(2) Villars, P. A Three-Dimensional Structural Stability Diagram for 1011 Binary AB2 Intermetallic Compounds: II. *J. Less-Common Met.* **1984**, *99*, 33−43.

(3) Villars, P. Three-Dimensional Structural Stability Diagrams for 648 Binary AB_3 and 389 Binary A_3B_5 Intermetallic Compounds: III. *J. Less-Common Met.* **1984**, *102*, 199−211.

(4) Villars, P. Factors governing crystal structures. In *Intermetallic Compounds: Crystal Structures of Intermetallic Compounds*; Westbrook, J. H., Fleischer, R. L., Eds.; John Wiley & Sons: New York, 2000; Vol. *1*, pp 1−49.

(5) Mooser, E.; Pearson, W. B. On the Crystal Chemistry of Normal Valence Compounds. *Acta Crystallogr.* **1959**, *12*, 1015−1022.

(6) Miedema, A. R. The electronegativity parameter for transition metals: Heat of formation and charge transfer in alloys. *J. Less-Common Met.* **1973**, *32*, 117−136.

(7) de Boer, F. R.; Pettifor, D. G. *Cohesion in Metals: Transition Metal Alloys*; North-Holland: Amsterdam, The Netherlands, 1989.

(8) 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.

(9) Savitskii, E. M.; Gribulya, V. B.; Kiselyova, N. N. On the Application of Cybernetic Prediction Systems in the Search for New Magnetic Materials. *J. Less-Common Met.* **1980**, *72*, 307−315.

(10) Burdett, J. K.; Price, G. D.; Price, S. L. Factors influencing solid-state structure—An analysis using pseudopotential radii structure maps. *Phys. Rev. B* **1981**, *24*, 2903−2912.

(11) Pettifor, D. G. A Chemical Scale for Crystal-Structure Maps. *Solid State Commun.* **1984**, *51*, 31−34.

(12) Pettifor, D. G. The Structures of Binary Compounds I. Phenomenological Structure Maps. *J. Phys. C: Solid State Phys.* **1986**, *19*, 285−313.

(13) Pettifor, D. G. Structure Maps for Pseudobinary and Ternary Phases. *Mater. Sci. Technol.* **1988**, *4*, 675−691.

(14) Villars, P.; Mathis, K.; Hulliger, F. In *The Structures of Binary Compounds*; de Boer, F. R., Pettifor, D. G., Eds.; North-Holland: Amsterdam, The Netherlands, 1989; Vol. *2*, pp 1−103.

(15) Kiselyova, N. N. Information-Predicting Systems for the Design of New Materials. *J. Alloys Compd.* **1993**, *197*, 159−165.

(16) Makino, Y. Interpretation of Bandgap, Heat of Formation and Structural Mapping for sp-Bonded Binary Compounds on the Basis of Bond Orbital Model and Orbital Electronegativity. *Intermetallics* **1994**, *2*, 55−66.

(17) Makino, Y. Structural Mapping of Binary Compounds between Transitional Metals on the Basis of Bond Orbital Model and Orbital Electronegativity. *Intermetallics* **1994**, *2*, 67−72.

(18) Pettifor, D. G. Structure mapping. In *Intermetallic Compounds: Crystal Structures of Intermetallic Compounds*; Westbrook, J. H., Fleischer, R. L., Eds.; John Wiley & Sons: New York, 2000; Vol. *1*, pp 195−214.

(19) Chen, Y.; Iwata, S.; Liu, J.; Villars, P.; Rodgers, J. Structural stability of atomic environment types in AB intermetallic compounds. *Modell. Simul. Mater. Sci. Eng.* **1996**, *4*, 335−348.

(20) Pettifor, D. G. Structure maps revisited. *J. Phys.: Condens. Matter.* **2003**, *15*, V13−V16.

(21) Villars, P.; Cenzual, K.; Daams, J.; Chen, Y.; Iwata, S. Data-driven atomic environment prediction for binaries using Mendeleev number: Part 1. Composition AB. *J. Alloys Compd.* **2004**, *367*, 167−175.

(22) Villars, P.; Daams, J.; Shikata, Y.; Rajan, K.; Iwata, S. A new approach to describe elemental-property parameters. *Chem. Met. Alloys* **2008**, *1*, 1−23.

(23) Jóhannesson, G. H.; Bligaard, T.; Ruban, A. V.; Skriver, H. L.; Jacobsen, K. W.; Nørskov, J. K. Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.* **2002**, *88*, 255506.

(24) Curtarolo, S.; Morgan, D.; Persson, K.; Rodgers, J.; Ceder, G. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **2003**, *91*, 135503.

(25) Dudiy, S. V.; Zunger, A. Searching for alloy configurations with target physical properties: Impurity design via a genetic algorithm inverse band structure approach. *Phys. Rev. Lett.* **2006**, *97*, 046401.

(26) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **2006**, *5*, 641−646.

(27) Sluiter, M. H. F. Lattice stability prediction of elemental tetrahedrally closed packed structure. *Acta Mater.* **2007**, *55*, 3707−3718.

(28) Mohn, C. E.; Kob, W. A genetic algorithm for the atomistic design and global optimization of substitutionally disordered materials. *Comput. Mater. Sci.* **2009**, *45*, 111−117.

(29) Oganov, A. R.; Valle, M. How to quantify energy landscapes of solids. *J. Chem. Phys.* **2009**, *130*, 104−504.

(30) Gadzuric, S.; Suh, C.; Gaune-Escard, M.; Rajan, K. Extracting information from molten salt database. *Met. Trans. A* **2006**, *37*, 3411−3414.

(31) Rajagopalan, A.; Rajan, K. Informatics based optimization of crystallographic descriptors for framework structures. In *Combinatorial and High-Throughput Discovery and Optimization of Catalysts and Materials*; Maier, W., Potyrailo, R. A., Eds.; CRC Press: Boca Raton, FL, 2006; pp 47−59.

(32) George, L.; Hrubiak, R.; Rajan, K.; Saxena, S. Principal component analysis on properties of binary and ternary hydrides and a comparison of metal versus metal hydride properties. *J. Alloys Compd.* **2009**, *478*, 731−735.

(33) Rajan, K. Data mining and inorganic crystallography. In *Data Mining in Crystallography*; Kuleshova, D. W. M., Liudmila, N., Eds.; Structure and Bonding Series; Springer-Verlag: Berlin, Germany, 2010; Vol. *134*, pp 59−87.

(34) Broderick, S. R.; Rajan, K. Eigenvalue Decomposition of Spectral Features in Density of States Curves. *Europhys. Lett.* **2011**, *95*, 57005.

(35) Villars, P.; Berndt, M.; Brandenburg, K.; Cenzual, K.; Daams, J.; Hulliger, F.; Massalski, T.; Okamoto, H.; Osaki, K.; Prince, A.; Putz, H.; Iwata, S. The Pauling File, Binaries Edition. *J. Alloys Compd.* **2004**, *367*, 293−297.

(36) Shannon, C. E. A mathematical theory of communication. *Bell Sys. Tech. J.* **1948**, *27*, 379−423, 623−656.

(37) Khinchin, A. I. *Mathematical Foundations of Information Theory*; Dover: New York, 1957.

(38) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall: New York, 1984.

(39) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953−17979.

(40) 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.

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

(42) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46*, 6671−6687.
(43) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188−5192.
(44) Methfessel, M.; Paxton, A. T. High-precision sampling for Brillouin-zone integration in metals. *Phys. Rev. B* **1989**, *40*, 3616−3621.
(45) Paxton, A. T.; Methfessel, M.; Polatoglou, H. M. Structural energy−volume relations in first-row transition metals. *Phys. Rev. B* **1990**, *41*, 8127−8138.
(46) Blöchl, P. E.; Jepsen, O.; Andersen, O. K. Improved tetrahedron method for Brillouin-zone integrations. *Phys. Rev. B* **1994**, *49*, 16223−16233.
(47) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.