CHEMISTRY OF MATERIALS

Article

pubs.acs.org/cm

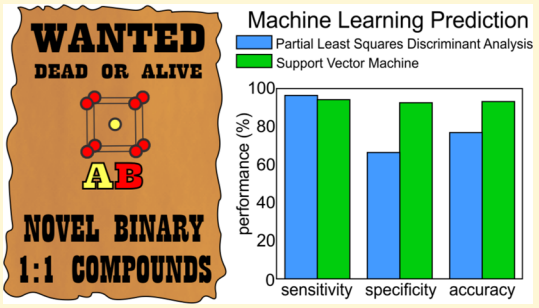
Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis

Anton O. Oliynyk, Lawrence A. Adutwum, James J. Harynuk,* and Arthur Mar*

Department of Chemistry, University of Alberta, Edmonton, Alberta T6G 2G2, Canada

Supporting Information

ABSTRACT: Partial least-squares discriminant analysis (PLS-DA) and support vector machine (SVM) techniques were applied to develop a crystal structure predictor for binary AB compounds. Models were trained and validated on the basis of the classification of 706 AB compounds adopting the seven most common structure types (CsCl, NaCl, ZnS, CuAu, TlI, β-FeB, and NiAs), through data extracted from Pearson's Crystal Data and ASM Alloy Phase Diagram Database. Out of 56 initial variables (descriptors based on elemental properties only), 31 were selected in as unbiased manner as possible through a procedure of forward selection and backward elimination, with the quality of the model evaluated by measuring the cluster resolution at each step. PLS-DA gave sensitivity of 96.5%, specificity of 66.0%, and accuracy of 77.1% for the validation set data, whereas SVM gave sensitivity of 94.2%, specificity of 92.7%, and accuracy of 93.2%, a significant improvement. Radii, electronegativity, and valence electrons, previously chosen intuitively in structure maps, were confirmed as important variables. PLS-DA and SVM could also make quantitative predictions of hypothetical compounds, unlike semiclassical approaches. The new compound RhCd was predicted to have the CsCl-type structure by PLS-DA (0.669 probability) and, at an even stronger confidence level, by SVM (0.918 probability). RhCd was synthesized by reaction of the elements at 800 °C and confirmed by X-ray diffraction to adopt the CsCl-type structure. SVM is thus a superior classification method in crystallography that is fast and makes correct, quantitative predictions; it may be more broadly applicable to help identify the structure of unknown compounds with any arbitrary composition.



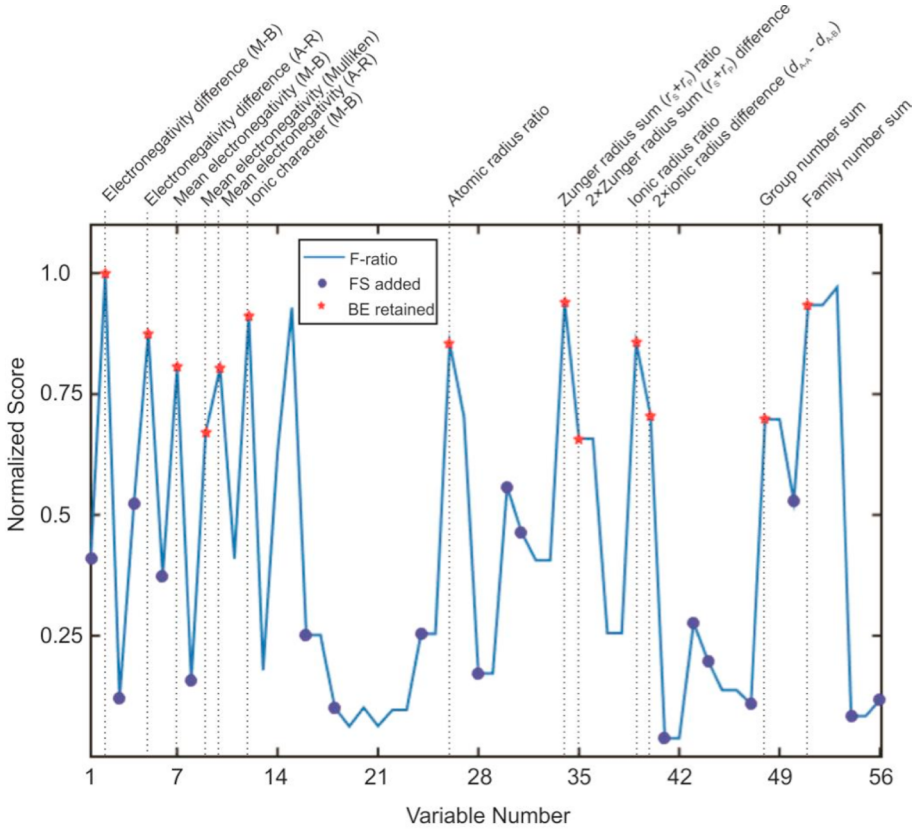
1. INTRODUCTION

A fundamental goal in chemistry is identifying what compounds form given an arbitrary combination of elements and what structures they adopt. Even for the simplest case, that of equiatomic binary compounds AB, where A and B are any elements in the periodic table, the problem is complex because there are many factors that influence structure formation. In the early days of crystallography, when structure determination was still difficult, it was hoped that, by correlating atomic properties and systematizating empirical structural information, "*perhaps we had come to a time when we could predict what the structures are without X-ray diffraction patterns.*"[1] Size factors were first invoked through radius ratio rules to rationalize the structures of ionic solids AB and the preferred coordination geometries of ions, but they failed to account for the observation that NaCl-type structures are far more prevalent than predicted.[2] Later, other atomic and physical properties were included, such as electronegativities and valence electron numbers, giving a more nuanced picture and generating structure maps (e.g., Mooser and Pearson,[3,4] Phillips and van Vechten,[5] Pettifor,[6] Zunger,[7] Villars[8,9]) that succeeded in segregating structure types. For example, focusing on intermetallic compounds AB, Villars considered 182 variables and tested mathematical combinations of these variables to identify three expressions, difference in Zunger pseudopotential radii sums, difference in Martynov-Batsanov electronegativity, and sum of valence electrons, that separated 988 compounds into 20 structure types with <3% violations, an impressive achievement.[8] The elucidation of such structure maps could be described as a semiclassical or semiempirical approach toward structure prediction. At the other extreme, first-principles electronic structure calculations can be performed to evaluate the stability of compounds; this approach is feasible if powerful computational facilities are available and can provide guidance to discovering new compounds.[10−13] However, regardless of the methods, predictions are worthless unless they can be validated experimentally.

Chemometric techniques have been applied to understand a variety of chemical systems such as predicting optimal experimental conditions,[14,15] identifying patterns in jet fuels,[16] classifying gasoline samples according to type or origin,[17−21] and discovering biomarkers.[22,23] In materials science, the wealth of information in databases[24,25] offers opportunities for data mining to address problems such as engineering band gaps of semiconductors,[26] enhancing hardness of nitrides,[27] and designing topologies of zeolites.[28] Cluster analysis and principal

Received: July 15, 2016
Revised: August 19, 2016
Published: August 22, 2016

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681



1. •Electronegativity difference (Pauling scale)	13. Ionic character (Gordy scale)	37. Ionic radius sum (d_{A-B})
2. ★Electronegativity difference (Martynov-Batsanov scale)	14. Ionic character (Mulliken scale)	38. Mean ionic radius
3. •Electronegativity difference (Gordy scale)	15. Ionic character (Allred-Rochow scale)	39. ★Ionic radius ratio
4. •Electronegativity difference (Mulliken scale)	16. •Sum of valence electrons	40. ★2×ionic radius difference (d_{A-A} - d_{A-B})
5. ★Electronegativity difference (Allred-Rochow scale)	17. Mean number of electrons	41. •Crystal radius sum (d_{A-B})
6. •Mean electronegativity (Pauling scale)	18. •Atomic number sum	42. Mean crystal radius
7. ★Mean electronegativity (Martynov-Batsanov scale)	19. Atomic number difference	43. •Crystal radius ratio
8. •Mean electronegativity (Gordy scale)	20. Mean atomic number	44. •2×crystal radius difference (d_{A-A} - d_{A-B})
9. ★Mean electronegativity (Mulliken scale)	21. Atomic weight difference	45. Period number sum
10. ★Mean electronegativity (Allred-Rochow scale)	22. Mean atomic weight	46. Mean period number
11. Ionic character (Pauling scale)	23. Atomic weight sum	47. •Period number difference
12. ★Ionic character (Martynov-Batsanov scale)	24. •Atomic radius sum (d_{A-B})	48. ★Group number sum
	25. Mean atomic radius	49. Mean group number
	26. ★Atomic radius ratio	50. •Group number difference
	27. 2×atomic radius difference (d_{A-A} - d_{A-B})	51. ★Family number sum
	28. •Covalent radius sum (d_{A-B})	52. Mean Family number
	29. Mean covalent radius	53. Family number difference
	30. •Covalent radius ratio	54. •Quantum number (l) sum
	31. •2×covalent radius difference (d_{A-A} - d_{A-B})	55. Mean quantum number (l) mean
	32. Zunger radius sum (r_S+r_P) sum	56. •Quantum number (l) difference
	33. Mean Zunger radius sum (r_S+r_P)	
	34. ★Zunger radius sum (r_S+r_P) ratio	
	35. ★2×Zunger radius sum (r_S+r_P) difference	
	36. Zunger radius sum (r_S+r_P) difference	

Figure 1. Fisher ratio scores for all variables (identified in the legend) selected during backward elimination (red stars) and forward selection (blue circles).

component analysis (PCA) have been widely used to identify inherent patterns in chemical data.[29,30] Arguably the most popular technique for data exploration, PCA is effective in reducing the dimensionality of a large data set, to help visualize similarities among samples and correlations between variables in the data. Supervised classification or pattern recognition tools such as linear discriminant analysis and partial least-squares discriminant analysis (PLS-DA) are also applied routinely to chemical data.[31]

Another supervised approach is the group of support vector machine (SVM) methods. SVMs are boundary-based methods that aim to maximize the gap (a hyperplane in a higher dimensional space) separating samples belonging to two classes; they do not model the entire ensemble of samples in each class but, instead, are constructed on the basis of those samples close to the boundary between the two classes.[32] SVMs have been popular in biochemistry and medicine[33] and have been used extensively to model and study ligand binding,[34] protein folding,[35] and other biological processes. In contrast, they have been less commonly applied in other branches of chemistry and, to our knowledge, never before in inorganic crystallography. Compared to linear discriminant analysis methods, SVMs are more flexible because the kernel function (used to build the model through a radial or linear basis) can be changed to optimize the performance. Automatic tuning of the kernel function to maximize the separation boundaries between classes improves classification.

When a chemometric model is constructed, feature selection is an important step. The goal is to remove (or minimize) the influence of noisy or irrelevant variables that degrade model

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

quality in terms of its performance for the classification problem, so that the final model is thereby built primarily upon variables having high information content.[36] Model quality can be evaluated through statistical measures such as sensitivity (proportion of true positives), specificity (proportion of true negatives), and accuracy (proportion of true results). Feature selection is commonly achieved through inspection of Fisher (*F*-ratio) scores, the ratio of between-group variability (explained variance) to within-group variability (unexplained variance);[16] selectivity (*S*-ratio) scores, which evaluate the usefulness of each variable in a regression model;[37] or loadings and variable importance in projection (VIP) scores[38] (or similar statistics) during model optimization. Generally, variables with higher scores contribute more information to the model. Another parameter that can be used to guide feature selection is the model quality metric termed cluster resolution (CR), which is the product of the maximal size of noncolliding confidence (hyper)ellipses that can be generated around samples clustered by class assignment in any reduced dimensionality score space (e.g., PCA space). An initial subset of variables that are likely to have high information content is selected and tested sequentially, from the lowest-ranked (least likely to be useful) to the highest-ranked (most likely to be useful). Through a process called cluster resolution-guided feature selection (CR-FS), implemented by means of a hybrid backward elimination/forward selection of variables, the contribution of all variables to model quality can be evaluated in the context of other variables.[39]

Here, we revisit the longstanding problem of predicting the structures of binary AB compounds, with several goals. First, we use the CR-FS algorithm applied in PCA space to determine what combinations of variables (atomic and physical properties) best optimize the discrimination of structure types and thereby evaluate the reliability of previous structure maps and gain insight on factors influencing structural preference. Second, variables retained after feature selection were used to build PLS-DA and SVM models, with the superior one chosen to predict the structure of a new compound. Third, we confirm the existence of a heretofore unknown AB compound through experiment. Although more than half of the possible AB compounds (out of all combinations of elements) remain uninvestigated, the latest report of a newly synthesized CsCl-type binary compound (at ambient conditions and containing nonradioactive elements) was that of RhZn, over 15 years ago.[40] The overarching motivation is to develop tools to guide and accelerate the search for other AB compounds and, more broadly, new materials.

2. EXPERIMENTAL SECTION

2.1. Chemometric Analysis of AB Phases. Crystallographic data of AB compounds were extracted from Pearson's Crystal Data[24] and ASM Alloy Phase Diagram Database;[25] additional data (up to September 2015) were obtained from searches on SciFinder.[41] All possible AB combinations were considered provided that (i) they did not contain hydrogen, noble gases, or elements with $Z > 83$ (radioactive elements and actinides) and (ii) they exhibit exact 1:1 stoichiometry. Out of 2926 possible combinations satisfying these conditions, 974 compounds are experimentally confirmed to exist under ambient temperatures and pressures, crystallizing in 107 unique structure types (Table S1 and XLSX file).

Variables used to describe atomic properties were chosen from those which have well-defined values for all or most elements (or which can be interpolated, such as for the lanthanide series). They generally fall into a small number of categories: (i) electronegativities in different scales,[42−46] (ii) various types of radii,[2] and (iii) properties derived from position in the periodic table (e.g., number of valence electrons, group number, and others).[47] Mathematical expressions (such as sums or differences for two elements A and B) derived from these properties were also treated as variables. In total, 56 variables were considered (see legend of Figure 1 later).

The data for these AB combinations and variables were represented in a 974 × 56 matrix. To ensure good statistical reliability, only those compounds (706) crystallizing in structure types containing more than 30 representatives were retained in this analysis: 257 in CsCl, 205 in NaCl, 102 in TlI, 42 in β-FeB, 36 in NiAs, 33 in ZnS, and 31 in CuAu structure types. The data were normalized, mean-centered, and scaled to unit variances along the columns (variables) of the data matrix. The preprocessed data were split into two parts, two-thirds (470) for training (i.e., variable selection and model building) and one-third (236) for external validation, such that each set had approximately the same proportions of compounds belonging to the different structure types. (In general, at least half of the data should be assigned to the training set.) The training data were split in half, with 235 samples being used for feature ranking and 235 samples for optimization. Although the data splitting was performed only once, based on past experience with the CR-FS algorithm, we consider the size of the data set to be sufficiently large that the feature selection procedure will not depend on the assignment of particular compounds to the training and validation sets.

The CR-FS algorithm was implemented in a three-dimensional PCA score space (PC1 vs PC2 vs PC3) with variables ranked by *F*-ratio. The 20 most highly ranked variables were used for initial model construction and subjected to backward elimination. The remaining variables were tested during forward selection. In this procedure, the initial subset of variables that are likely to have a high information content was selected and tested sequentially, from the lowest-ranked (least likely to be useful) to the highest-ranked (most likely to be useful). If removing a variable improved CR, it was eliminated from the model; otherwise, the algorithm proceeded to the next variable. The forward selection step began with those variables that survived backward elimination and then tested progressively lower-ranked variables which had not been initially considered, in turn. Those variables whose inclusion improved the model quality were retained, while those that did not were discarded. For large data sets (with millions of variables), the forward selection step is stopped after progressing for sufficiently long that the likelihood of finding a new useful variable is essentially zero. For small data sets (with <1000 variables), all variables are tested, as was the case here.

Subsequently, PLS-DA and SVM models were generated with all samples from the training set (feature ranking and optimization sets combined) using only those variables retained by CR-FS. The SVM classification was performed with a radial basis function and a venetian blind cross-validation with 10-fold data split to optimize the model. The ability of SVM vs PLS-DA models to correctly predict the crystal structure of new compounds was evaluated using the validation set data. The model was then used to predict the crystal structure of a completely unknown compound, RhCd.

Data handling and feature selection were performed with in-house written algorithms in Matlab 2015a (The Mathworks, Natick, MA). PLS-DA and SVM models were generated using PLS Toolbox Version 8.0.1 (Eigenvector Research Inc., Wenatchee, WA). Results for objective comparison were class predicted probabilities of the PLS-DA and SVM models.[48] All computations were performed on a Windows PC, running on an Intel Core i7-4790 CPU with 32 GB RAM.

2.2. Synthesis and Characterization of RhCd. From the chemometric analysis above, RhCd was predicted to adopt a CsCl-type structure. A pressed pellet of Rh powder (99.95%, Alfa-Aesar) and filed Cd pieces (99.95%, Alfa-Aesar) in a 1:1 molar ratio with a total mass of 0.2 g was placed in a fused-silica tube, which was evacuated and sealed. The tube was heated to 800 °C, kept at that temperature for 1 week, and quenched in cold water. The product was examined by powder X-ray diffraction (XRD) performed on an Inel diffractometer equipped with a curved position-sensitive detector and

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

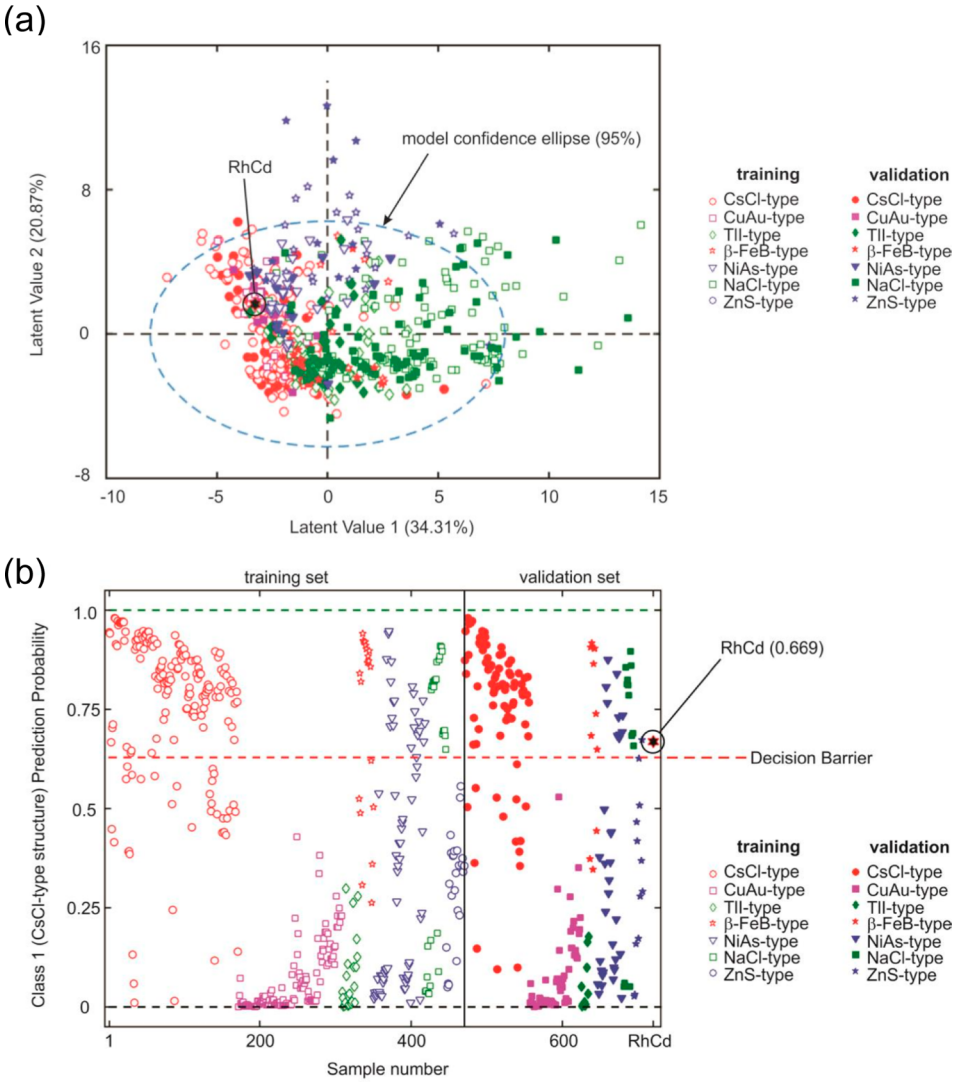


Figure 2. (a) Latent value scores for AB compounds and (b) predicted probability for CsCl-type structures for PLS-DA models using 31 selected features.

by energy-dispersive X-ray (EDX) analysis on a JEOL JSM-6010LA scanning electron microscope.

Small single crystals, confirmed by EDX analysis to have the composition RhCd, were selected for structure determination. Intensity data were collected at room temperature on a Bruker PLATFORM diffractometer equipped with a SMART APEX II CCD area detector and a graphite-monochromated Mo Kα radiation source, using ω scans at 8 different ϕ angles with a frame width of 0.3° and an exposure time of 15 s per frame. The structure was solved and refined with use of the SHELXTL (version 6.12) program package.[49] Face-indexed absorption corrections were applied. The cubic space group $Pm\bar{3}m$ was chosen on the basis of Laue symmetry, intensity statistics, and systematic absences. Full crystallographic data, in CIF format, have been sent to Fachinformationszentrum Karlsruhe, Abt. PROKA, 76344 Eggenstein-Leopoldshafen, Germany, as supplementary material No. CSD-431550 and can be obtained by contacting FIZ (quoting the article details and the corresponding CSD numbers).

3. RESULTS AND DISCUSSION

3.1. Cluster Resolution Feature Selection. The CR-FS algorithm is well-suited to the simultaneous optimization of multiple-class (i.e., $n \geq 3$) problems, by using the product of all pairwise cluster resolution values as the objective function.[39,50−52] In this case, the optimization was for a seven-class problem, with each class representing one of the seven common structure types adopted by binary compounds AB: CsCl, NaCl, ZnS, CuAu, TlI, β-FeB, and NiAs.

For feature selection, the 56 variables used to describe atomic properties were first ranked according to Fisher (F-ratio) or selectivity ratio (S-ratio) scores (Figure S1). The choice of which ratio to use was not found to be critical because both tended to arrive at similar results. The variables included in the final model according to F-ratio scores consist of those retained in the backward elimination step (red stars) and those added in the forward selection step (blue circles) (Figure 1). Some high-ranked variables were eliminated while some low-ranked ones were added, indicating that high F-ratios only suggest potential importance but do not guarantee actual importance of variables to the intended classification model. After backward elimination and forward selection, 31 out of 56 variables were retained. The initially high-ranked variables that were removed through backward elimination were average Martynov-Batsanov or Mulliken electronegativities, Pauling electronegativities (and expressions derived from them), interatomic distances, and differences of Zunger radii sums ($r_s + r_p$). Conversely, the initially low-ranked variables that were included through

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

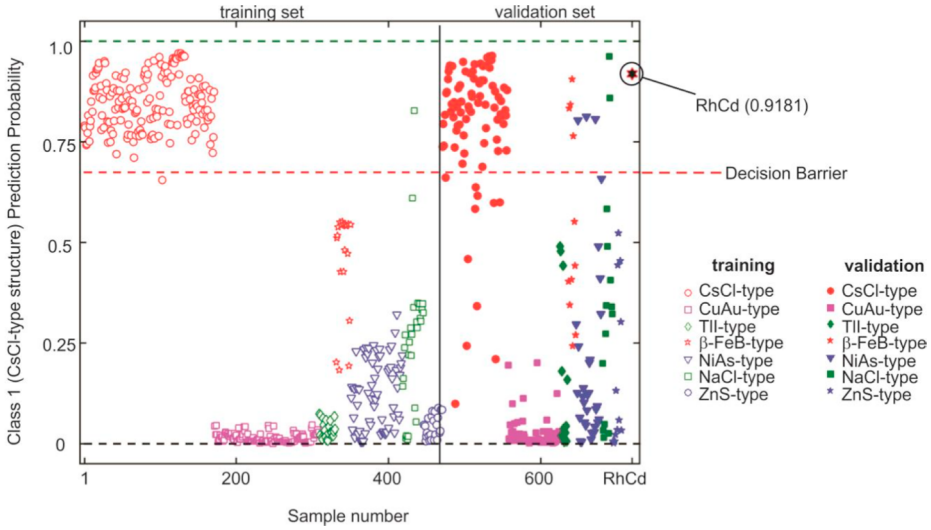


Figure 3. Predicted probability for CsCl-type structures for SVM model using 31 selected features.

forward selection were average numbers of valence electrons and Zunger radii sums (and some expressions derived from them).

3.2. PLS-DA Prediction. To predict the structure type that a compound is likely to adopt, PLS-DA was applied as a classification technique. Plots of the scores of PLS-DA, namely, the latent values, provide information about the underlying patterns in the data; that is, they serve as structure maps in which compounds with similar properties are projected close to each other in latent values score spaces (Figure 2a). This guidance can be very valuable in cases where the experimental synthesis is high-risk (e.g., with radioactive elements like Tc) or expensive (e.g., with precious metals like Rh). As a test, we arbitrarily chose a hypothetical compound RhCd, which has not been previously reported and for which no phase diagram investigations (in the Rh−Cd system) have been conducted. The hypothetical compound RhCd is located at the point marked by the black hexagram in Figure 2a and predicted by the PLS-DA model to adopt a CsCl-type structure. Samples lying within the confidence ellipse (95%) of the model indicate that they can be predicted with a higher degree of confidence. However, this point also falls at the peripheries of the predicted probabilities of CuAu- and NiAs-type structures, which overlap slightly with CsCl-type structures. Note that the CuAu-type structure is essentially a tetragonally distorted version of the cubic CsCl-type structure with the inequivalent *a*- and *c*-parameters being only slightly different. It is not surprising that these two structure types are difficult to distinguish experimentally (as they have similar X-ray diffraction patterns) and theoretically.

The results can also be visualized as plots of the sample number on the abscissa and the prediction probability on the ordinate, as shown for the CsCl-type structure using the variables selected (Figure 2b). The probability should be close to unity for samples predicted to belong to a given class and close to zero for all other samples. The PLS-DA model predicted the training set data with sensitivity of 95.9% and specificity of 66.6%. Although the model predicts the CsCl-type structure largely correctly, the false positive rate is very high and the overall model accuracy was 77.2%. When the model was applied to the validation set (containing 236 data points), the sensitivity was 96.5%, the specificity was 66.0%, and the accuracy was 77.1%. Even though there seemed to be some improvement in predicting the validation set data, the prediction probability for the test compound RhCd is 0.669, which is only slightly higher than the decision boundary. A better classifier is desired.

3.3. SVM Prediction. We present here for the first time an application of SVM to inorganic crystal structure prediction. With the same training and validation set data used as in the PLS-DA model, a SVM classification model was generated to predict various structure types. The prediction probabilities for the CsCl-type structure were much starker (Figure 3). For the training set data, the sensitivity was 100.0%, the specificity was 99.3%, and the accuracy was 99.6%; for the validation set data, the sensitivity was 94.2%, the specificity was 92.7%, and the accuracy was 93.2%. Thus, the model performance was significantly better with SVM than with PLS-DA methods.

To evaluate the need for feature selection, a SVM model was constructed on the basis of all 56 variables using the full training data set. This model led to prediction sensitivity, specificity, and accuracy of 42.7%, 100%, and 79.2%, respectively, on the training data. When this model was applied to the validation set, prediction sensitivity, specificity, and accuracy were found to be 44.2%, 98.0%, and 78.4%, respectively, thus demonstrating that feature selection is essential.

The feature selection process with the CR-FS algorithm took about 12 h to complete, involving the simultaneous optimization of CR for seven classes in three-dimensional score space and 720 pairwise CR calculations required at each step of the optimization. However, once the feature selection is completed, the process of training, validation, and prediction was extremely fast, taking less than 1 min to complete for either PLS-DA or SVM models.

Figure 3 reveals a handful of false negatives (compounds that are experimentally found to adopt CsCl-type structures notwithstanding a low predicted probability) in the SVM model, and it may be interesting to see if any insight can be gained by examining them. Starting from the worst predictions (lowest probability), these are CaPd, CsI, CuY, CuEu, and YRh. Although there do not seem to be any common features among them, the most glaring outlier is CsI. The cesium halides (CsCl, CsBr, CsI) also stand out as false negatives in the PLS-DA

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

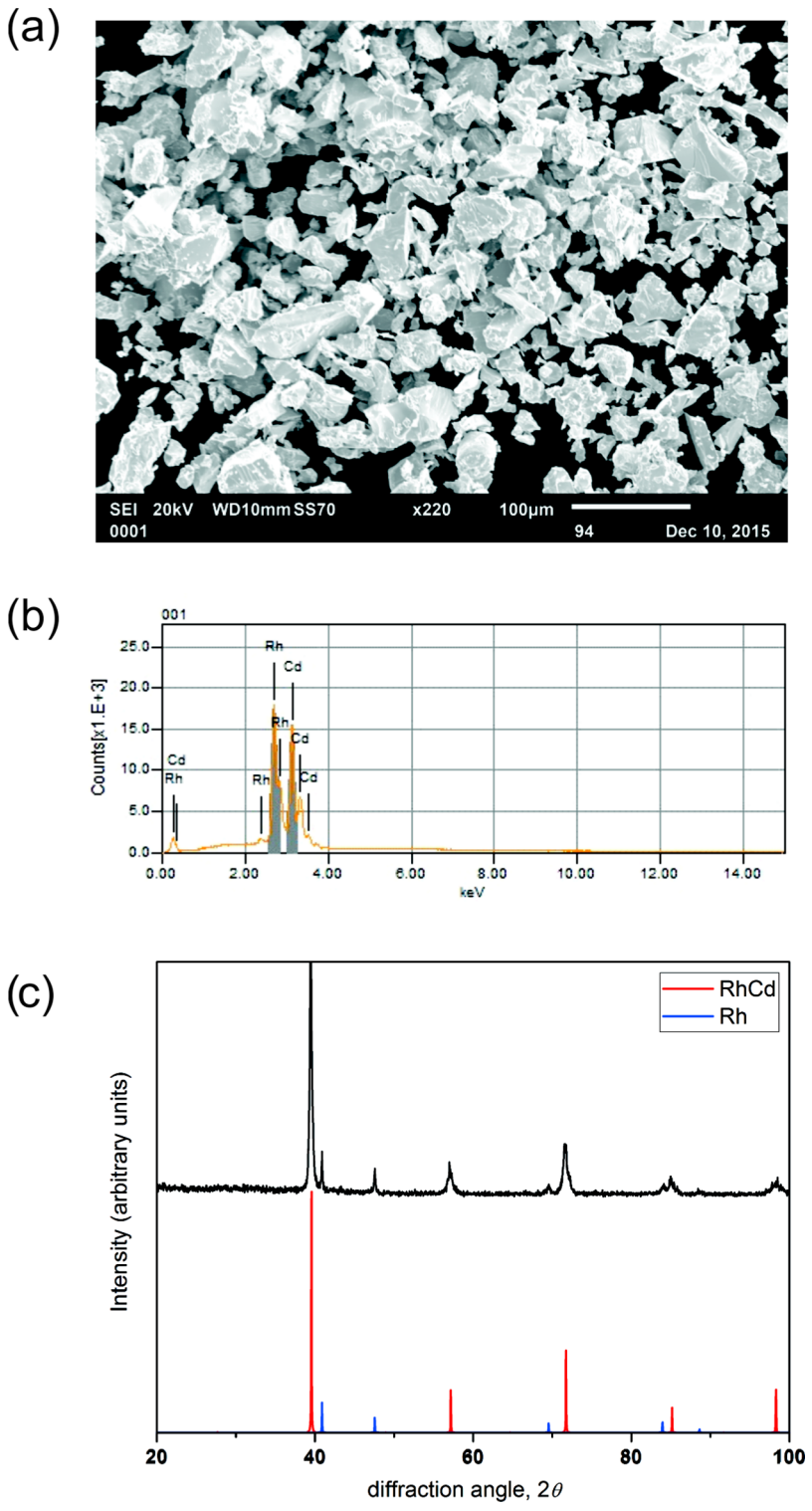


Figure 4. New binary compound RhCd. (a) SEM image of crystals, (b) EDX spectrum indicating presence of equal ratios of Rh and Cd in crystals, and (c) powder XRD pattern confirming CsCl-type structure.

model. As has been discussed previously,[53] the CsCl-type structure is actually very rare for highly ionic compounds, and when it does occur, it is over a narrow range of stability in which size factors must compete with electrostatic factors.

It is important to emphasize that, like all statistical methods, the reliability of these crystallographic predictions depends on the size and quality of the experimental data set. First, for AB compounds belonging to structure types that are rare or unique, it may be difficult to train the model because the variance cannot be captured adequately. For this reason, we examined structure types containing a relatively large number of representatives (30 or more), but it may be interesting in future iterations of this model to include less common structure types. Second, AB compounds for which no experimental information was available about their existence were excluded from the model. Including such unknown compounds would

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

bias the model with unreliable information because we do not know if they really do not exist or were just missed; moreover, the overwhelming amount of data belonging to the class of hypothetically nonexistent compounds could lead to a problem of overfitting with the SVM method and thus worsen the quality of the predictions. Third, we recognize that many AB compounds may have a homogeneity range so that deviations from exact 1:1 stoichiometry are possible and that there are binary phases with nearly 1:1 stoichiometry adopting other structure types. The problem of dealing with solid solutions is a difficult one even from the standpoint of assigning what structure type they belong to. For example, at what point should solid solutions derived from CsCl-type be designated as W-type or those from NaCl-type as Cu-type? Although there remain challenges in extending this predictive model to more complicated situations, there are major advantages of our approach. The CR-FS algorithm (implemented in conjunction with the SVM model) almost always converges to a solution giving a subset of features that are important to help distinguish between different classes. However, if the cluster resolution converges to a low value, then the model quality (as measured by sensitivity, specificity, and accuracy) will also be poor. Predictions of the structure type of an unknown compound are expressed quantitatively in the form of a probability (a number between 0 and 1).

3.4. Prediction and Experimental Verification of RhCd.

From the analysis above, 31 out of 56 variables were important for separating CsCl-type structures from others (Figure 2a). As has been emphasized in the past, the CsCl-type structure is, notwithstanding the ionic character of the prototype compound CsCl itself, essentially a metallic one adopted by hundreds of intermetallics, exhibiting the highest coordination geometries (cubic, CN8) among AB-type structures.[53] Bond character (as gauged by electronegativity differences) and radius ratios are thus key factors in the formation of the CsCl-type structure. Although PCA/PLS-DA has been used elsewhere to classify structures over limited types of compounds,[54] its application to the broader set of data here was not as successful. For hypothetical RhCd, the predicted probability that it adopts the CsCl-type structure was only 0.669 and the overall quality of the PLS-DA model was not great (Figure 2b). The SVM model yielded significant improvement in sensitivity, specificity, and accuracy (>92%) after feature selection and gave a much higher probability of 0.918 of a CsCl-type structure for RhCd (Figure 3). The probabilities of RhCd adopting other types of structures were extremely low or essentially nil (Figures S2–S7). Thus, SVM achieves a clearer separation between structure types and gives more definitive predictions in this case.

The synthesis of RhCd was attempted by reaction of the elements at 800 °C. The products were examined by SEM, EDX, and powder XRD (Figure 4). Small single crystals, <50 μm in their longest dimension, were obtained. Their average composition is 47(2)% Rh and 53(2)% Cd, in excellent agreement with the formula RhCd. The powder XRD pattern confirms that RhCd adopts the CsCl-type structure. Small amounts of Rh metal (<9%) were found as a byproduct; this is understandable given that Cd metal is volatile and a small amount was found sublimed on the walls of the fused-silica tube. The structure was refined from single-crystal diffraction data (Table S2). With an assignment of fully occupied Rh at 0, 0, 0 and Cd at 1/2, 1/2, 1/2 in space group $Pm\bar{3}m$, an excellent agreement factor (R_1 = 0.008) was obtained. (Note that, because there are only 13 unique reflections and 4 refinable parameters, a low data-to-parameter ratio is unavoidable.)

During the review of this manuscript, we became aware of unpublished information giving evidence for the existence of RhCd.[55] A recent report has also now appeared describing a second binary phase, Rh_2Cd_{15}, in the Rh−Cd system.[56]

3.5. Factors Influencing Structures of AB Compounds.

It is instructive to compare the variables selected by CR-FS with those used in earlier schemes to derive structure maps of AB compounds. Previously, Villars noted that the most common variables used in such structure maps can be grouped according to the pattern of behavior with position in the periodic table and represented by five prototypical properties: (A) radius, (B) atomic number, (C) atomization energy, (D) electronegativity, and (E) number of valence electrons.[8] Of these, excellent separation of structure types was achieved using expressions involving radius, electronegativity, and number of valence electrons. Because these earlier structure maps were deduced by trial-and-error and chemical intuition, it was not certain if other combinations of properties could give better separation; however, inclusion of additional variables from classes B and C (atomic number, atomization energy) could be ruled out. Our results confirm that cluster resolution is optimized by properties related to radius and electronegativity, which were high-ranked variables, and by number of valence electrons, which was, surprisingly, a low-ranked variable.

Of course, there are many scales of radii and electronegativities and different ways of expressing the number of valence electrons. In structure maps, an arbitrary decision had to be made in selecting one of these scales, based on the subset of AB compounds being examined. In CR-FS, the selection of these scales is performed in an unbiased manner. It may appear that introducing too many different scales conveying similar information could confuse the learning algorithm. However, as in all statistical methods, some redundancy is desirable to provide stability in the iterative selection of variables; thus, variables are eliminated not because they are low-ranked but because they do not contribute meaningfully to model quality.

3.5.1. Electronegativity. Among the ∼20 different scales that have been developed for electronegativity, 5 were chosen that are appropriate for intermetallics (which constitute the majority of AB compounds, given that the periodic table consists of mostly metals): Pauling, Martynov-Batsanov, Gordy, Mulliken, and Allred- Rochow.[42−46] Only two, Martynov-Batsanov and Allred-Rochow, survived the model used to optimize cluster resolution, in the form of electronegativity differences ($\Delta\chi$) or ionic character ($f = 1-\exp(-1/4(\chi_A - \chi_B)^2)$, where χ_A and χ_B are electronegativities of A and B atoms, respectively). It is interesting that the Pauling scale, which is the most familiar and widely used among chemists, is simply not as effective. The Allred-Rochow scale relates the attraction of valence electrons in an atom to electrostatic force, evaluated from effective nuclear charge (estimated using Slater's rules) and covalent radius (obtained experimentally); it differs from the Pauling scale largely with respect to the precious metals, which have corrected values that are not the same as in sulfur and phosphorus. The Martynov-Batsanov scale is evaluated from average ionization energies of valence electrons; because it was specifically developed for crystalline inorganic substances, it is reassuring that it works well to separate structure types of AB compounds, as was also concluded by Villars.[8]

3.5.2. Size. Since the earliest days of crystal chemistry, size factors were intuitively believed to be crucial in determining

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

crystal structures. However, even for the very limited subset of alkali-metal halides among AB phases, radius ratio rules (based on ionic radii) fail miserably (notwithstanding misrepresentations in some introductory chemistry textbooks). If a much larger variety of AB phases is considered, exhibiting diverse types of bonding, it is not obvious which scales of radii (atomic, ionic, covalent, metallic, and others) would be most appropriate. Even if one assumes that metallic bonding is predominant because AB phases are mostly intermetallic compounds, the bonding interaction between metal atoms can have different degrees of polar character. As a brute force way to evaluate these different scales of radii ("the proof of the pudding is in the eating"), we compared the sums of radii with the actual A−B bond lengths found in all AB phases with CsCl-, NaCl- and ZnS-type structures reported in Pearson's Crystal Data, including metastable phases and theoretical calculations (Figure 5). Overall, the observed bond lengths agreed much

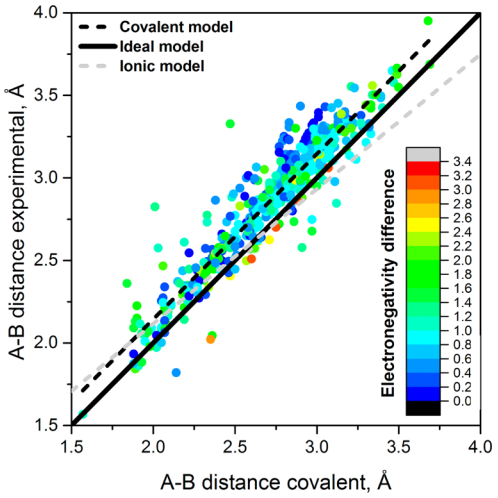


Figure 5. Comparison of experimental distances in AB compounds with sums of covalent radii (or other scales of radii, as indicated).

better with sums of covalent radii than with ionic or atomic radii; the average deviation was less than 5%, which is even smaller than the deviation of cell parameters for multiple reports of the same compound at the same conditions. Introducing a Porterfield correction[57] to account for polar character in A−B bonds (r_{AB} (in Å) = $r_A + r_B - 0.07(\Delta\chi)^2$) did not lead to improvement; in fact, the change was negligible, with the average deviation barely changing from 4.9% to 5.0%. Surprisingly, the calculated distances did not agree better with experimental values even in highly ionic compounds, for which this correction was specifically designed to tackle. This analysis suggests that the covalent radii scale is expected to be a good variable choice to represent size within a large set of compounds exhibiting a wide variety of chemical bonding interactions. Although the sum of covalent radii tends to be slightly smaller than the experimental distances, the trends (as measured by the slope) are the same.

The size variables selected by CR-FS were actually a combination of atomic, covalent, and ionic radii, reflecting a compromise to capture the diversity of bonding interactions in AB compounds. However, we also considered Zunger pseudopotential radii.[7] In this scale, orbital radii are obtained by quantum calculations within a pseudopotential (Simons-Bloch) in which core electrons are frozen. For a single atom A, the radii sum $(r_s + r_p)^A$ can be defined. Although the difference of Zunger radii sums, $(r_s + r_p)^A - (r_s + r_p)^B$, was chosen by Villars as a coordinate in his structure map,[8] it did not survive in the model optimization. Instead, the sum of Zunger radii sums, $(r_s + r_p)^A + (r_s + r_p)^B$, was a high-ranked variable that was effective according to cluster resolution.

3.5.3. Electron Count. Electron count is an important factor for normal valence compounds following the octet rule.[58] Thus, the total number of valence electrons, ΣVE_{AB}, was a third coordinate in Villars' structure map.[8] In our study, the average number of valence electrons, $\overline{VE}_{AB}$, was an initially low-ranked variable selected in model optimization. These two expressions convey similar information originating from position of elements in the periodic table, but the average is more effective in separating structure types for compounds formed from disparate vs closely related elements. To expand on this idea, we introduced a family number that classifies elements into: (1) alkali metals, (2) alkaline-earth metals, (3) f-block metals, (4) d-block metals, (5) p-block metals, (6) p-block metalloids, (7) p-block nonmetals, (8) chalcogens, and (9) noble gases. This classification is not the same as group number (1−18 or IA−VIIIA/IB−VIIIB), but it reflects better the drastic differences in chemical behavior in the p-block in which elements in the same group can form quite different compounds and structures. (The concept is comparable to that of Mendeleev numbers, which are sequential integers assigned to each element so that those of similar chemical properties are grouped close together.[9]) As expected, variables based on this family number make a significant contribution to separating structure types.

4. CONCLUSIONS

The problem of predicting structures *a priori* is a challenging one that pervades all of chemistry. Although it can be addressed directly through variational quantum mechanical calculations, a semiempirical approach is attractive because it makes use of chemical concepts (such as atomic size, bond character, and electron count) which we intuitively believe must be important. However, relying on user-selected variables poses the risk of introducing bias. In quantum mechanical calculations, the crystal structure that is obtained is the one in which the total energy is minimized when atomic orbitals interact with each other. In the semiempirical approach, the implicit assumption is that there is a correspondence between the total energy of a crystal structure with higher-level properties such as radii and electronegativity. The relationship is undoubtedly complex, but it is reassuring that most of the spread in the structure types adopted by AB compounds can be captured in a few number of principal components that depend on combinations of variables that were previously suspected to be critical: Zunger pseudopotential radii, Martynov-Batsanov electronegativities, and numbers of valence electrons.

The size factor plays a greater role in formation of CsCl-type structures than of other cubic structures (NaCl- and ZnS-type) for AB phases. This can be appreciated by comparing nearest-neighbor heteroatomic A−B contacts vs next-nearest-neighbor homoatomic A−A and B−B contacts (Figure 6). The homoatomic distances are typically only slightly longer (0.4−0.6 Å) than the heteroatomic distances in CsCl-type structures, but they are considerably longer (>1.0 Å) in NaCl- and ZnS-type structures. The interpretation is that a more complex balance of long-range heteroatomic (A−B) and homoatomic (A−A and B−B) interactions influences the formation of CsCl-type structures, whereas heteroatomic interactions dominate in

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

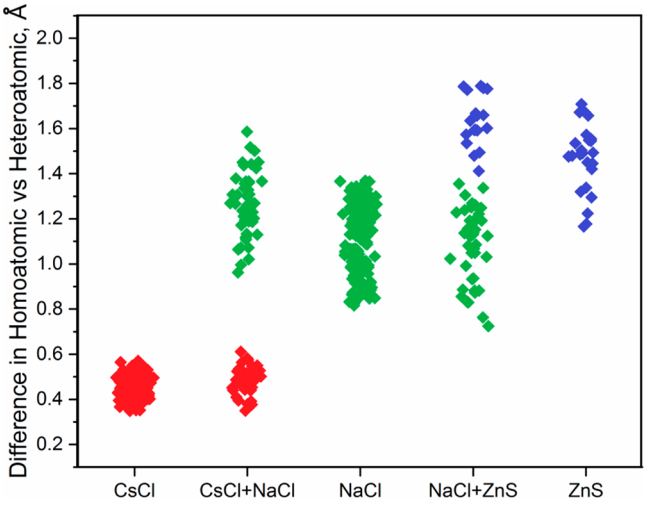


Figure 6. Difference between homoatomic and heteroatomic distances, $r_{A-A} - r_{A-B}$, in the most common structure types formed by AB phases.

NaCl- and ZnS-type structures. This disparity persists even when polymorphism occurs (e.g., CsCl/NaCl or NaCl/ZnS) for a fixed combination of elements.

The successful preparation of RhCd adopting a CsCl-type structure has been achieved, suggesting that the SVM model shows promise as a powerful predictive tool in crystallography. However, we wish to emphasize that these predictions do not preclude overcoming experimental difficulties; the synthetic chemist must still grapple with practical considerations such as choice of starting materials and reaction temperatures. Moreover, critical inspection of existing data and addition of experimental data in the form of new compounds can be fed back into the SVM model to help improve it, so that the experiment and prediction synergically benefit each other. Efforts are in progress to extend the use of SVM to predict the structures of other binary and ternary phases.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemmater.6b02905.

Structure types of all binary compounds AB (full data matrix), single-crystal XRD analysis for RhCd, Fisher and selectivity ratio scores for 56 variables, and additional plots of predicted probability for other structure types (PDF)

Structure types of binary compounds AB in Excel worksheet (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: james.harynuk@ualberta.ca.

*E-mail: arthur.mar@ualberta.ca.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, through the Discovery Grants (A.O.O., A.M., L.A.A., J.J.H.) and the Collaborative Research and Training Experience Programs (A.O.O., A.M.), and by Genome Canada/Genome Alberta (J.J.H., L.A.A.).

■ REFERENCES

(1) Pauling, L. "The empirical information made it clear that there was some possible systematization to the interatomic distances and also to other aspects of the crystal structures, and I thought perhaps we had come to the time when we could predict what the structures are without x-ray diffraction patterns." In *Structure and Bonding in Crystals*; O'Keeffe, M., Navrotsky, A., Eds.; Academic Press: New York, 1981; Vol. *1*, pp 1−12.

(2) Pauling, L. *The Nature of the Chemical Bond*, 3rd ed.; Cornell University Press: Ithaca, NY, 1960.

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

(4) Phillips, J. C. Structure and properties: Mooser-Pearson plots. *Helv. Phys. Acta* **1985**, *58*, 209−215.

(5) Phillips, J. C.; van Vechten, J. A. Dielectric classification of crystal structures, ionization potentials, and band structures. *Phys. Rev. Lett.* **1969**, *22*, 705−708.

(6) Pettifor, D. G. A chemical scale for crystal-structure maps. *Solid State Commun.* **1984**, *51*, 31−34.

(7) Zunger, A. Systematization of the stable crystal structure of all *AB*-type binary compounds: A pseudopotential orbital-radii approach. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1980**, *22*, 5839−5872.

(8) Villars, P. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *J. Less-Common Met.* **1983**, *92*, 215−238.

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

(10) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chem. Mater.* **2010**, *22*, 3762−3767.

(11) Oganov, A. R.; Lyakhov, A. O.; Valle, M. How Evolutionary Crystal Structure Prediction Works − and Why. *Acc. Chem. Res.* **2011**, *44*, 227−237.

(12) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191−201.

(13) Gautier, R.; Zhang, X.; Hu, L.; Yu, L.; Lin, Y.; Sunde, T. O. L.; Chon, D.; Poeppelmeier, K. R.; Zunger, A. *Nat. Chem.* **2015**, *7*, 308−316.

(14) Wold, S. Chemometrics; what do we mean with it, and what do we want from it? *Chemom. Intell. Lab. Syst.* **1995**, *30*, 109−115.

(15) Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, 2nd ed.; Wiley-VCH: New York, 2007.

(16) Johnson, K. J.; Synovec, R. E. Pattern recognition of jet fuels: comprehensive GC × GC with ANOVA-based feature selection and principal component analysis. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 225−237.

(17) Doble, P.; Sandercock, M.; Du Pasquier, E.; Petocz, P.; Roux, C.; Dawson, M. Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks. *Forensic Sci. Int.* **2003**, *132*, 26−39.

(18) Sandercock, P. M. L.; Du Pasquier, E. Chemical fingerprinting of unevaporated automotive gasoline samples. *Forensic Sci. Int.* **2003**, *134*, 1−10.

(19) Sandercock, P. M. L.; Du Pasquier, E. Chemical fingerprinting of gasoline: 2. Comparison of unevaporated and evaporated automotive gasoline samples. *Forensic Sci. Int.* **2004**, *140*, 43−59.

(20) Sandercock, P. M. L.; Du Pasquier, E. Chemical fingerprinting of gasoline: 3. Comparison of unevaporated automotive gasoline samples from Australia and New Zealand. *Forensic Sci. Int.* **2004**, *140*, 71−77.

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

(21) 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.
(22) Li, X.; Xu, Z.; Lu, X.; Yang, X.; Yin, P.; Kong, H.; Yu, Y.; Xu, G. Comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry for metabonomics: Biomarker discovery of diabetes mellitus. *Anal. Chim. Acta* **2009**, *633*, 257−262.
(23) Beckstrom, A. C.; Humston, E. M.; Snyder, L. R.; Synovec, R. E.; Juul, S. E. Application of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry method to identify potential biomarkers of perinatal asphyxia in a non-human primate model. *J. Chromatogr. A* **2011**, *1218*, 1899−1906.
(24) Villars, P.; Cenzual, K. *Pearson's Crystal Data − Crystal Structure Database for Inorganic Compounds* (on DVD), Release 2015/16; ASM International: Materials Park, OH, 2016.
(25) Villars, P., Okamoto, H., Cenzual, K., Eds. *ASM Alloy Phase Diagrams Database*; ASM International: Materials Park, OH, 2016 (http://www.asminternational.org).
(26) Srinivasan, S.; Rajan, K. "Property Phase Diagrams" for Compound Semiconductors through Data Mining. *Materials* **2013**, *6*, 279−290.
(27) Pettersson, F.; Suh, C.; Saxén, H.; Rajan, K.; Chakraborti, N. Analyzing Sparse Data for Nitride Spinels Using Data Mining, Neural Networks, and Multiobjective Genetic Algorithms. *Mater. Manuf. Processes* **2009**, *24*, 2−9.
(28) Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Novel Approach for Clustering Zeolite Crystal Structures. *Mol. Inf.* **2010**, *29*, 297−301.
(29) Rajan, K. Materials informatics. *Mater. Today* **2005**, *8*, 38−45.
(30) Broderick, S.; Rajan, K. Informatics derived materials databases for multifunctional properties. *Sci. Technol. Adv. Mater.* **2015**, *16*, 013501-1−013501-8.
(31) Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166−173.
(32) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* **1992**, 144−152.
(33) Wang, L.; Brown, S. J. Prediction of RNA-binding residues in protein sequences using support vector mechines. *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society* **2006**, 5830−5833.
(34) Janda, J.-O.; Busch, M.; Kück, F.; Porfenenko, M.; Merkl, R. CLIPS-1D: analysis of multiple sequence alignments to deduce for residue-positions a role in catalysis, ligand-binding, or protein structure. *BMC Bioinf.* **2012**, *13*, 55−1−55−11.
(35) Redfern, O. C.; Harrison, A.; Dallman, T.; Pearl, F. M. G.; Orengo, C. A. CATHEDRAL: A Fast and Effective Algorithm to Predict Folds and Domain Boundaries from Multidomain Protein Structures. *PLoS Comput. Biol.* **2007**, *3*, 2333−2347.
(36) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.
(37) Rajalahti, T.; Arneberg, R.; Kroksveen, A. C.; Berle, M.; Myhr, K.-M.; Kvalheim, O. M. Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles. *Anal. Chem.* **2009**, *81*, 2581−2590.
(38) Chong, I.-G.; Jun, C.-H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103−112.
(39) Sinkov, N. A.; Johnston, B. M.; Sandercock, P. M. L.; Harynuk, J. J. Automated optimization and construction of chemometric models based on highly variable raw chromatographic data. *Anal. Chim. Acta* **2011**, *697*, 8−15.
(40) Gross, N.; Kotzyba, G.; Künnen, B.; Jeitschko, W. Binary Compounds of Rhodium and Zinc: RhZn, Rh_2Zn_{11}, and $RhZn_{13}$. *Z. Anorg. Allg. Chem.* **2001**, *627*, 155−163.
(41) *Scifinder*; Chemical Abstracts Service: Columbus, OH, 2015.
(42) Pauling, L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. *J. Am. Chem. Soc.* **1932**, *54*, 3570−3582.
(43) Martynov, A. I.; Batsanov, S. S. New approaches to determining the electronegativity of atoms. *Zh. Neorg. Khim.* **1980**, *5*, 3171−3175.
(44) Ghosh, D. C.; Chakraborty, T. Gordy's electrostatic scale of electronegativity revisited. *J. Mol. Struct.: THEOCHEM* **2009**, *906*, 87−93.
(45) Mulliken, R. S. A New Electroaffinity Scale; Together with Data on Valence States and on Valence Ionization Potentials and Electron Affinities. *J. Chem. Phys.* **1934**, *2*, 782−784.
(46) Allred, A. L.; Rochow, E. G. A scale of electronegativity based on electrostatic force. *J. Inorg. Nucl. Chem.* **1958**, *5*, 264−268.
(47) Emsley, J. *Nature's Building Blocks: An A-Z. Guide to the Elements*; Oxford University Press: New York, 2011.
(48) Lin, C.; Weng, R. C. *Simple probabilistic predictions for support vector regression*; National Taiwan University: Taipei, 2004.
(49) Sheldrick, G. M. *SHELXTL*, version 6.12; Bruker AXS Inc.: Madison, WI, 2001.
(50) Sinkov, N. A.; Harynuk, J. J. Three-dimensional cluster resolution for guiding automatic chemometric model optimization. *Talanta* **2013**, *103*, 252−259.
(51) Sinkov, N. A.; Sandercock, P. M. L.; Harynuk, J. J. Chemometric classification of casework arson samples based on gasoline content. *Forensic Sci. Int.* **2014**, *235*, 24−31.
(52) Adutwum, L. A.; Harynuk, J. J. Unique Ion Filter: A Data Reduction Tool for GC/MS Data Preprocessing Prior to Chemometric Analysis. *Anal. Chem.* **2014**, *86*, 7726−7733.
(53) Adams, D. M. *Inorganic Solids: An Introduction to Concepts in Solid-state Structural Chemistry*; Wiley: New York, 1974.
(54) Suh, C.; Rajan, K. Data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure-property relationships. *Mater. Sci. Technol.* **2009**, *25*, 466−471.
(55) Lidin, S.; Jana, P. P. Personal communication.
(56) Xie, W.; Liu, M.; Wang, Z.; Ong, N.-P.; Cava, R. J. Composite Icosahedron/Cube Endohedral Clusters in Rh_2Cd_{15}. *Inorg. Chem.* **2016**, *55*, 7605−7609.
(57) Porterfield, W. W. *Inorganic Chemistry: A Unified Approach*, 2nd ed.; Academic Press: San Diego, CA, 1993.
(58) Abegg, R. Die Valenz und das periodische System. Versuch einer Theorie der Molekularverbindungen. *Z. Anorg. Allg. Chem.* **1904**, *39*, 330−380.

DOI: 10.1021/acs.chemmater.6b02905
Chem. Mater. 2016, 28, 6672−6681

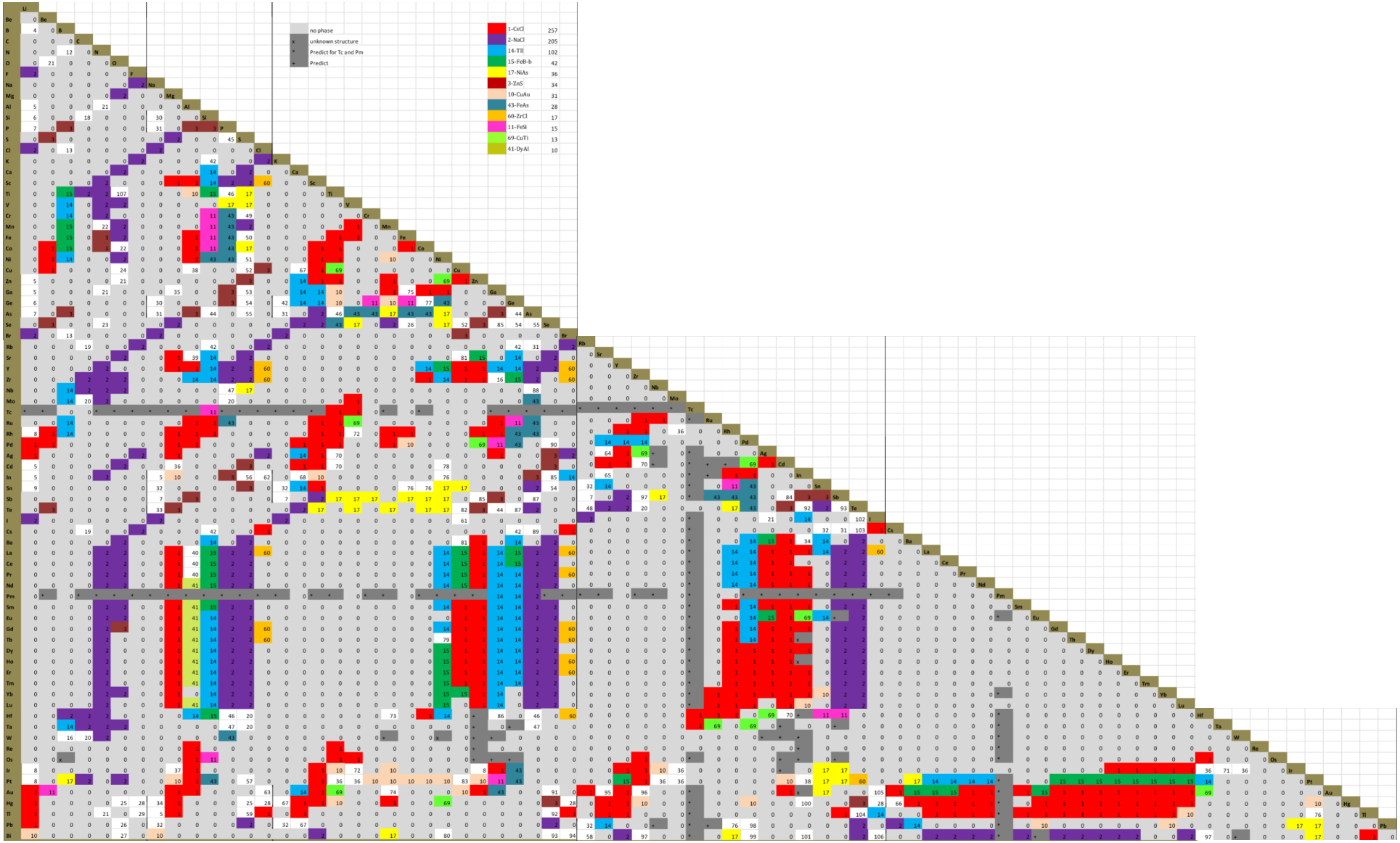
SUPPORTING INFORMATION

Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis

Anton O. Oliynyk, Lawrence A. Adutwum, James J. Harynuk, and Arthur Mar*

Department of Chemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2G2

Table S1. Structure Types of Binary Compounds AB [a]



[a] Legend:

0-does not exist	41-DyAl	82-CuTe	x-phase exists but structure is unknown (OsB, NiW, TbIn, HoIn, AuIn)
1-CsCl (Pm-3m)	42-KGe	83-CuPt	
2-NaCl (Fm-3m)	43-FeAs	84-CdSb	
3-ZnS (F-43m)	44-SiAs	85-GaSe	
4-LiB (P63/mmc)	45-AsS	86-ThIn	
5-NaTl (Fd-3m)	46-TiP	87-GeTe	*-contains Tc and Pm
6-LiGe (I41/a)	47-NbAs	88-NbAs	
7-LiAs (P121/c1)	48-CsTe	89-AsCs	+-uninvestigated
8-LiRh (P-6m2)	49-CrS	90-PdS	GaW
9-LiSn	50-FeS	91-AuSe	GaOs
10-CuAu	51-NiS	92-TlSe	ZnW
11-FeSi	52-CuS-b	93-BiSe	ZnRe
12-BN	53-GaS-a	94-BiBr	ZnOs
13-BCl	54-GeS	95-SrAu	ZnTa
14-TlI	55-AsS	96-AuZr	ZnHf
15-FeB-b	56-InS	97-ZrSb	MnW
16-MoB	57-PtS	98-PdPb	AgNb
17-NiAs	58-CsSb	99-PdBi	CdNb
18-SiC	59-TlS	100-HgIn	NbPb
19-RbC	60-ZrCl	101-InBi	GeTa
20-WC	61-CuI	102-TeI	GeOs
21-ZnO	62-InCl	103-CsTe	CdRu
22-CoO	63-AuCl	104-TlTe	InRu
23-SeN	64-SrAg	105-AuI	RuPb
24-CuO	65-SrIn	106-BiI	CdRh
25-HgS	66-KHg	107-TiO	AgW
26-PbO	67-CaCu		CdTa
27-BiO	68-PuGa		CdW
28-HgCl	69-CuTi		CdOs
29-TlF	70-CdTi		CdIr
30-NaSi-a	71- TaIr		HfIn
31-NaP	72-VIr		InW
32-NaPb	73-CdNi		InRe
33-NaTe	74-AuMn		InOs
34-NaHg	75-MnGa		SbEu
35-MgGa	76-CoSn		SbTa
36-AuCd	77-CoGe		SbOs
37-MgIr	78-CdNi		EuBi
38-CuAl	79-TbNi		BiW
39-InCl	80-NiBi		
40-CeAl	81-BaCu		

Table S2. Crystallographic Data for RhCd

Data collection and refinement	
formula	RhCd
fw (amu)	215.31
space group	$Pm\bar{3}m$ (No. 221)
a (Å)	3.2191(7)
V (Å^3)	33.358(13)
Z	1
ρ_{calcd} (g cm^{-3})	10.718
T (K)	296(2)
crystal dimensions (mm)	0.05 × 0.03 × 0.03
radiation	graphite monochromated Mo $K\alpha$, λ = 0.71073 Å
μ(Mo $K\alpha$) (mm^{-1})	27.489
transmission factors	0.285–0.666
2θ limits	17.96–65.48°
data collected	$-4 \le h \le 4$, $-4 \le k \le 4$, $-4 \le l \le 4$
no. of data collected	234
no. of unique data, including $F_o^2 < 0$	13 (R_{int} = 0.0152)
no. of unique data, with $F_o^2 > 2\sigma(F_o^2)$	13
no. of variables	4
$R(F)$ for $F_o^2 > 2\sigma(F_o^2)$ [a]	0.0086
$R_w(F_o^2)$ [b]	0.0185
goodness of fit	1.297
$(\Delta\rho)_{max}$, $(\Delta\rho)_{min}$ (e Å^{-3})	0.617, –0.339
Positional and displacement parameters [c]	
Rh at 1*a* (0, 0, 0)	
U_{iso} (Å^2)	0.02(2)
Cd at 1*b* (1/2, 1/2, 1/2)	
U_{iso} (Å^2)	0.016(14)

Interatomic distances (Å)	
Rh–Cd (×8)	2.7878(6)
Cd–Cd (×6)	3.2191(7)
Rh–Rh (×6)	3.2191(7)

[a] $R(F) = \sum \left| |F_o| - |F_c| \right| / \sum |F_o|$.

[b] $R_w(F_o^2) = \left[\sum \left[w(F_o^2 - F_c^2)^2 \right] / \sum w F_o^4 \right]^{1/2}$; $w^{-1} = \left[\sigma^2(F_o^2) + (Ap)^2 + Bp \right]$ where $p = \left[\max(F_o^2, 0) + 2F_c^2 \right] / 3$.

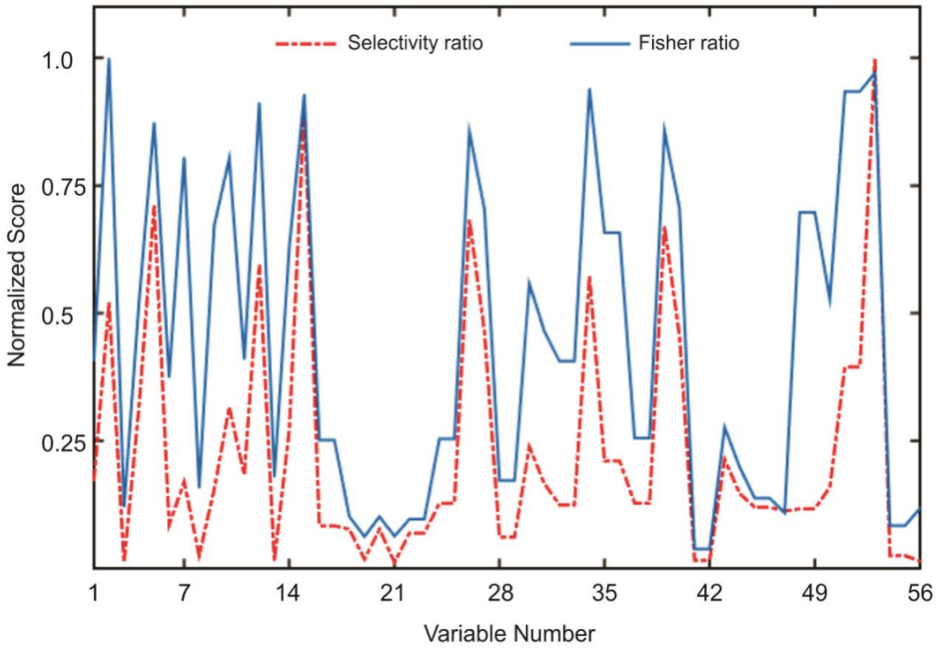


Figure S1. Fisher (blue solid line) and selectivity (red dashed line) ratio scores for 56 variables (defined in the legend in Figure 1 in the main text).

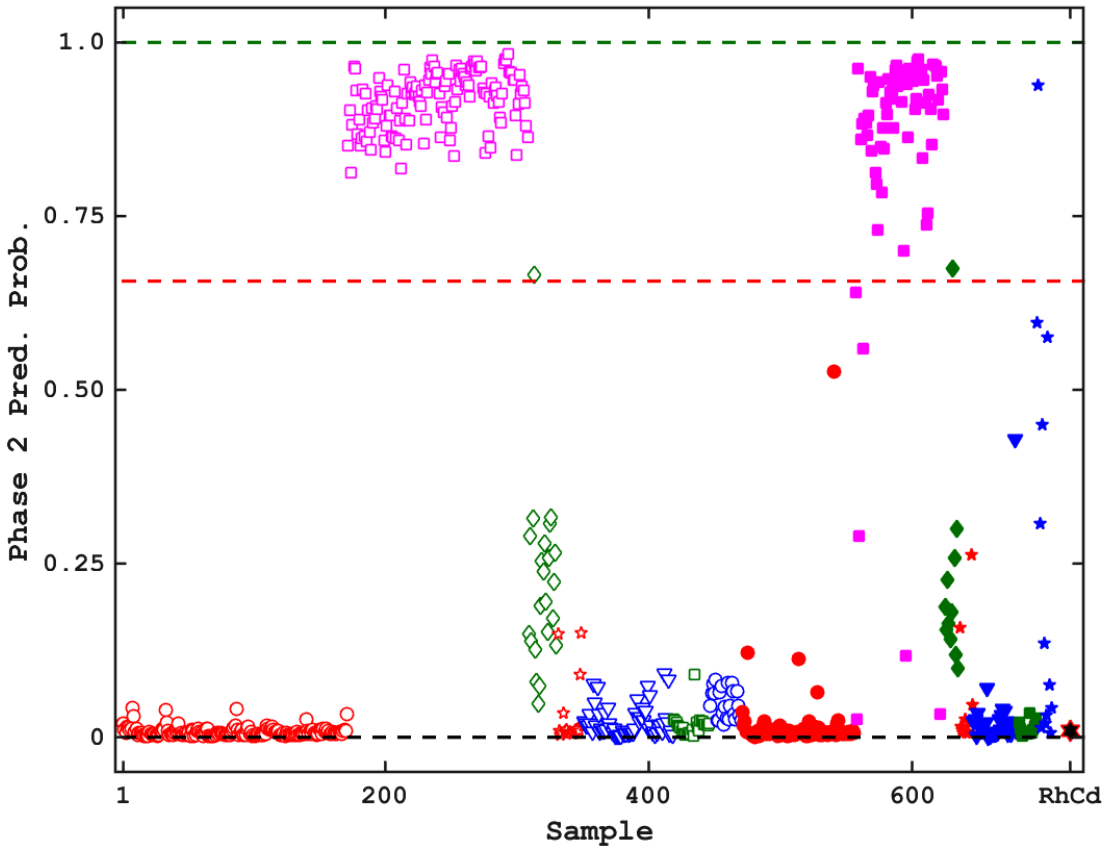


Figure S2. Predicted probability for NaCl-type structures for SVM model using 31 selected features.

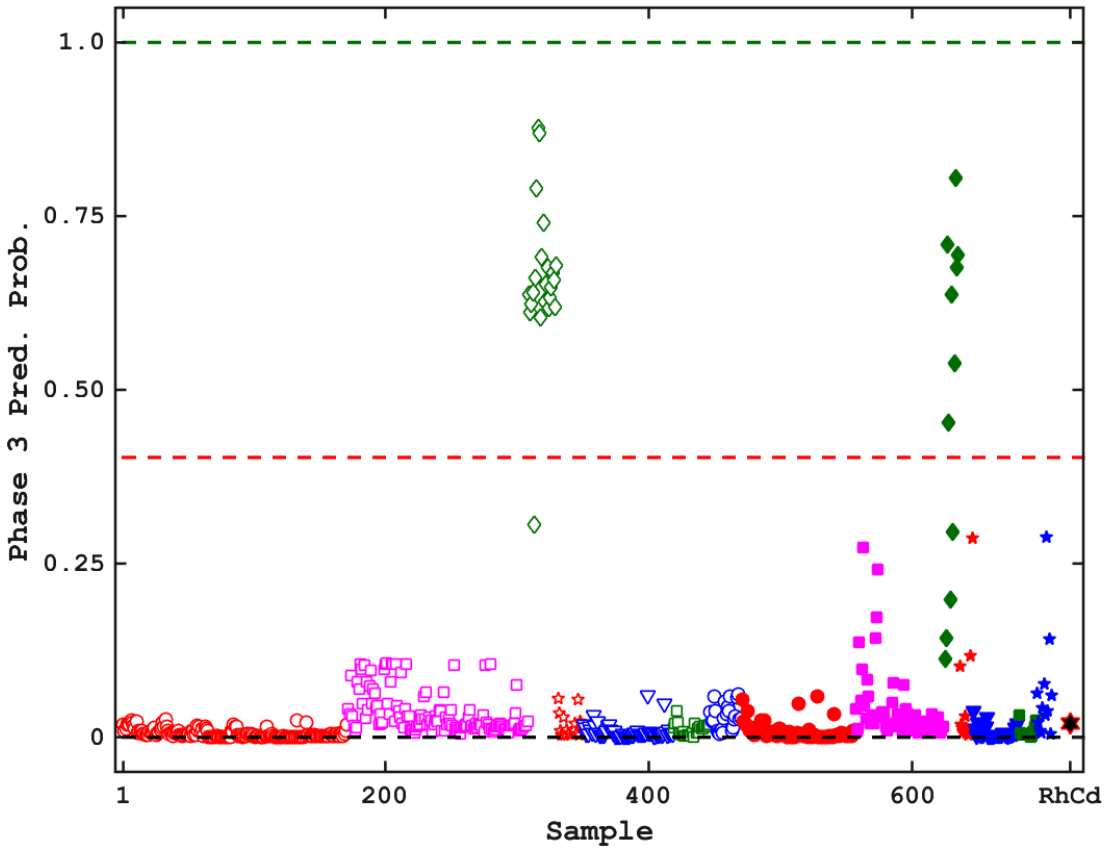


Figure S3. Predicted probability for ZnS-type structures for SVM model using 31 selected features.

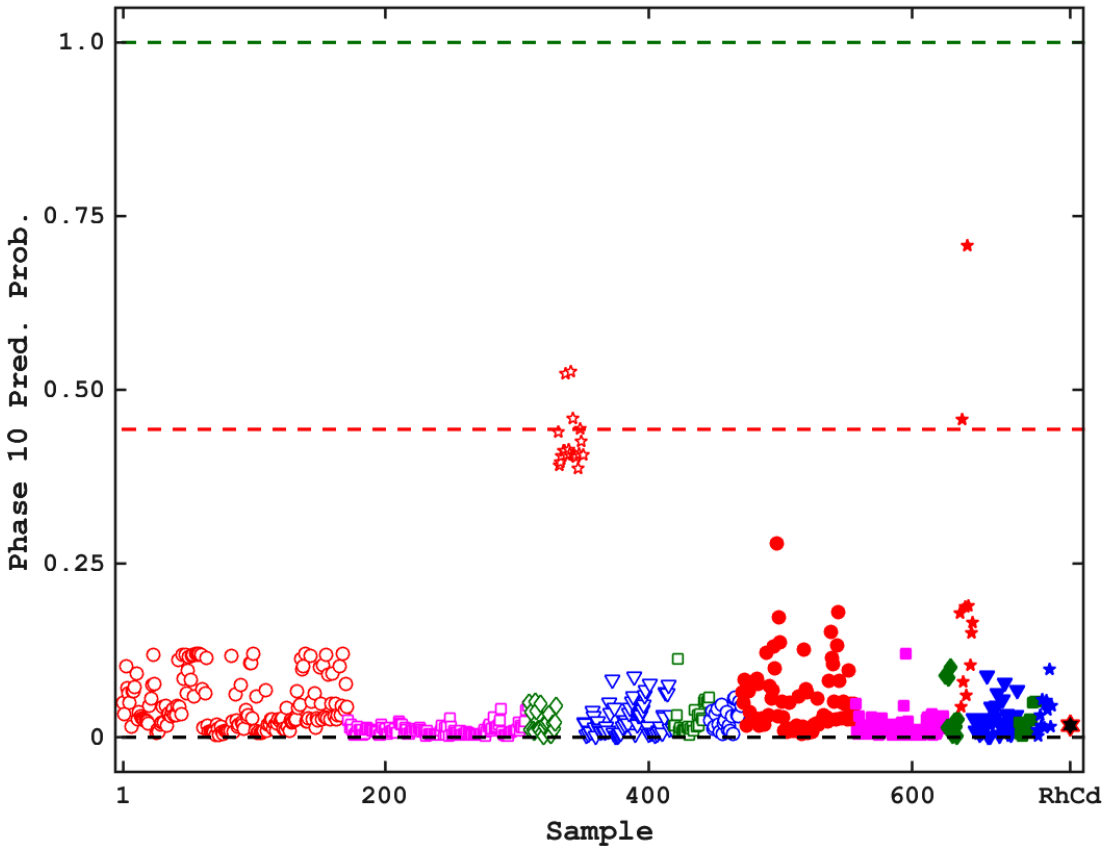


Figure S4. Predicted probability for CuAu-type structures for SVM model using 31 selected features.

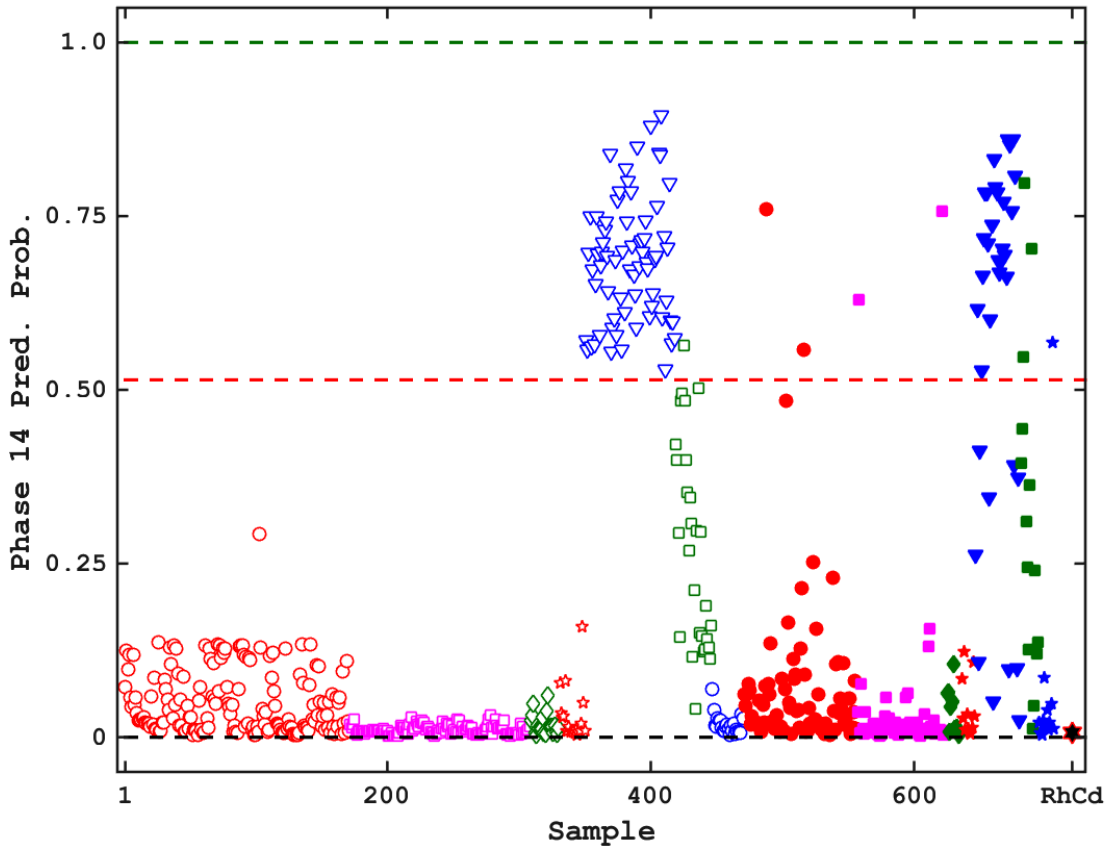


Figure S5. Predicted probability for TlI-type structures for SVM model using 31 selected features.

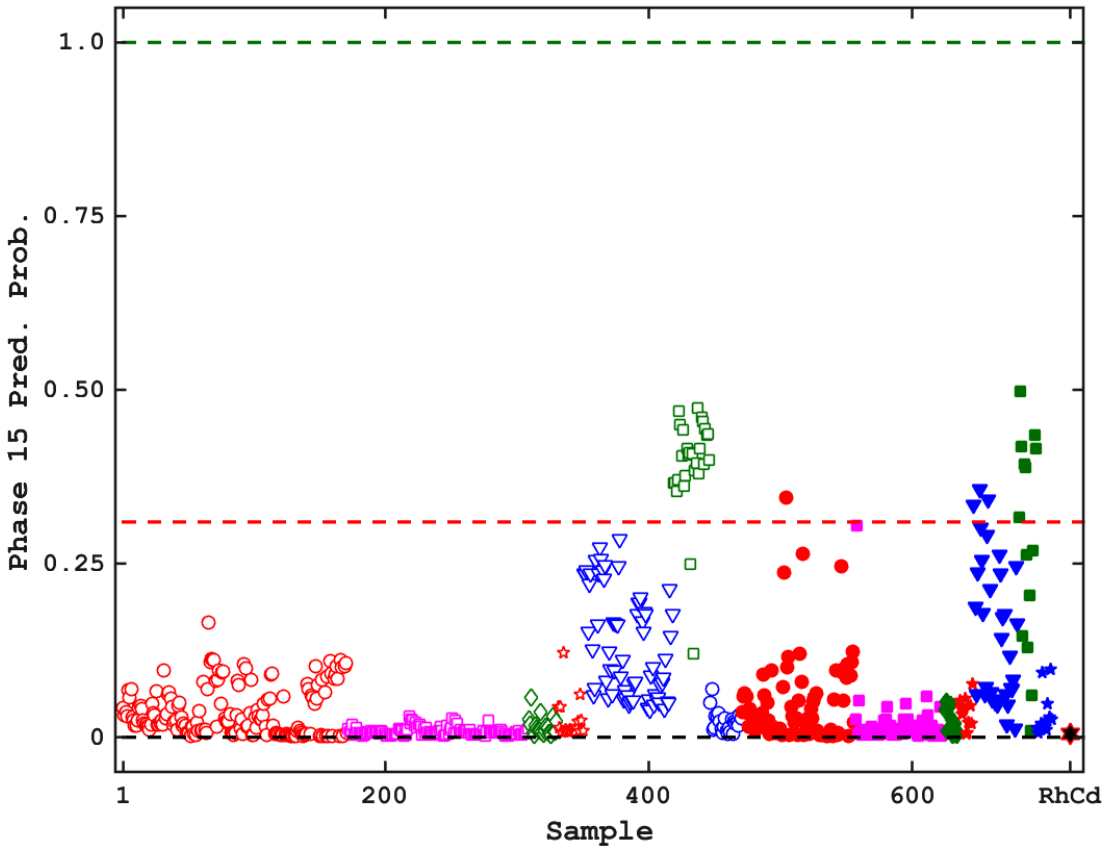


Figure S6. Predicted probability for β-FeB-type structures for SVM model using 31 selected features.

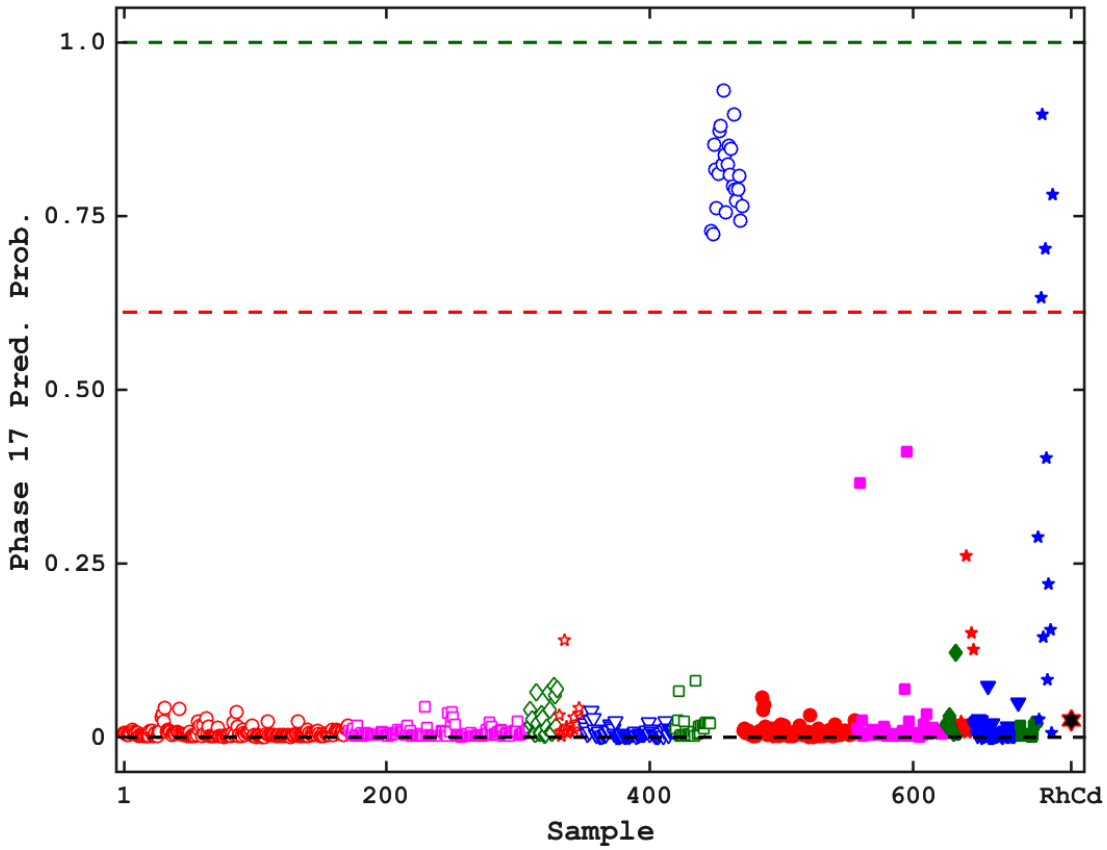


Figure S7. Predicted probability for NiAs-type structures for SVM model using 31 selected features.