Structure classification and melting temperature prediction in octet AB solids via machine learning

G. Pilania, J. E. Gubernatis, and T. Lookman

Phys. Rev. B **91**, 214302 — Published 15 June 2015

DOI: 10.1103/PhysRevB.91.214302

Structure classification and melting temperature prediction in octet AB solids via machine learning

G. Pilania,[1] J. E. Gubernatis,[2] and T. Lookman[2]

[1] *Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

[2] *Theoretical Divison, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

(Dated: May 29, 2015)

Abstract

Machine learning methods are being increasingly used in condensed matter physics and materials science to classify crystals structures and predict material properties. However, the reliability of these methods for a given problem, especially when large data sets are unavailable, has not been well studied. By addressing the tasks of classifying crystal structure and predicting melting temperatures of the octet subset of AB solids, we performed such a study and found potential problems with using machine learning methods on relatively small data sets. At the same time, however, we can reaffirm the potential power of such methods for these tasks. In particular, we uncovered an important new material feature, the excess Born effective charge, that significantly increased the accuracy of the predictions for the classification problem we defined. This discovery leads us to propose a new scale for the degree of ionicity and covalency in these solids. More specifically, we partitioned the crystal structures of a set of 75 octet solids into those that are ionic and covalent bonded and thus performed a binary classification task. We found that using the standard indices (r_σ, r_π), suggested by St. John and Bloch several decades ago, enabled an average success in classification of 92%. We found that using just r_σ and the excess Born effective charge ΔZ_A of the A atom enabled an average success of 97%, but we also found relatively large variations about these averages that were dependent on how certain machine learning methods were used and for which a standard deviation was not a proper measure of the degree of confidence we can place in either average. Instead, we calculated and report with 95% confidence that the traditional classification pair predicts an accuracy in the interval $[89\%, 95\%]$ and the accuracy of the new pair lies in the interval $[96\%, 99\%]$. For melting temperature predictions, the size of our data set was 46. We estimate the root-mean-squared error of our resulting model to be 11% of the mean melting temperature of the data, but we note that if the accuracy of this predicted error is itself measured, our estimated fitting error itself has root-mean-square error of 50%. In short, what we illustrate is that classification and regression predictions can vary significantly, depending on the details of how machine learning methods are applied to small data sets. This variation makes it important, if not essential, to average the predications and compute confidence intervals about these averages to report results properly. However, when properly used, these advanced statistical methods can advance our understanding and improve our predictions of material properties even for small data sets.

I. INTRODUCTION

In this paper, we apply several specific machine learning algorithms to study the crystal structure classification and melting temperature prediction of the octet subset of AB solids. Octet AB compounds are defined by $A^N B^{8-N}$ where N refers to the number of valence electrons. For the octets these classification and prediction tasks are ones visited by theorists and experimentalists off and on for over 50 years. Initially, the tasks were aimed at understanding the nature of chemical bonding in solids and the identification of new semiconductors. Today, these materials represent a well studied group on which to test methods for similar classification and prediction tasks to be applied to other classes of materials. These types of analyses provide models from which we can predict the expected properties of a proposed new material.

Historically, for the classification task for the octets and other materials, the analysis has been very simple, the construction of a structure plot which is just an xy-plot with the values of select physical features of the materials placed along the x and y axes. The challenge has been identifying two features amongst atomic radii, electronegativity, ionization potentials, etc. which enable the drawing of lines with a pencil and ruler that cluster materials with the same crystal structure. For the melting temperature prediction task, the search has been for more than two features amongst such quantities as the above, plus the bulk modulus, atomic number, nearest neighbor distance, etc., for use in a simple least-squares fit to known values of the melting temperature. For the classification task machine learning allows us to seek a hyper-dimensional classification model by using more than two features. Our principal objective is seeing whether doing so improves the classification and if so, identifying the features we need to include. With machine learning we can also provide more conveniently a measure of the accuracy of both our classification and melting temperature predictions.

The octet solids exhibit five crystal structures (rocksalt, zincblende, wurtzite, cesium chloride, and diamond), with rocksalt being the most common. What is difficult about the classification of these solids is drawing the boundary between a small set of rocksalts, zincblendes, and wurtzites (particularly between some zincblendes and wurtzites) whose ground state energy differences are small and positions in a structure plot are close. Past work by a number of authors[1–5] identified a very powerful pair of material features that classify the octet solids very well. These features are the r_σ and r_π pair first used in a

structural plot analysis by St. John and Bloch several decades ago.[2]

Saad et al.[6] and Ghiringhelli et al.[7] recently revisited the octet AB solids and used machine learning methods and more than two features as the basis for crystal structure classification. The work of these groups nicely illustrates the challenge in the classification problem: The differences in the ground state energies of a number of octet solids are small; further, the differences in the ground state energies of the same AB chemistry in different crystal structures is often smaller. Each group mitigated the challenge by redefining the classification problem. Saad et al. started by grouping 67 materials into four crystal structures (rocksalt, zincblende, wurtzite, diamond) plus two "dual structures" (zincblende-wurtzite, wurtzite-rocksalt) where the latter acknowledge the closeness in ground states of some materials in these crystal structures. Ghiringhelli et al.[7] also addressed the closeness of ground state energies of many octet AB solids and the closeness of these energies for the same AB chemistry in three different crystal structures (rocksalt, zincblende, and wurtzite). They chose not to distinguish between zincblende and wurtzite. With material features similar to that of Saad et al., they preformed a regression analysis that used a large number of functional combinations of a modest-sized set of features that included the energy difference between rocksalt and zincblende structure for 82 AB chemistries to refine the predictions of the energy difference between rocksalt and zincblende/wurtzite structures. From these results they inferred the expected crystal class, rocksalt or zincblende/wurtzite.

As Saad et al. and Ghiringhelli et al., we redefine the classification problem but do so in a quite different manner. As Ghiringhelli et al., we created a binary classification problem. In structure plots of the octets, the few cesium chloride solids sit near the rocksalts but apart from the rest, and the diamond structured materials sit near the zincblendes and wurtzites but apart from the rest. Accordingly, we designated the rocksalt and cesium chlorides as "rocksalt" and the remaining three structures as "non-rocksalts." This grouping separates the solids into two classes: those whose bonding is strongly ionic and those whose bonding ranges from predominantly covalent to very strongly covalent. From a machine learning perspective, we transformed the problem from a multi-class classification one (predicting all five crystal structures) into a binary classification one (rocksalt or not) as opposed to transforming a classification problem into a regression problem. The computational advantage of generating a binary classification task is the number of well developed machine learning methods designed for this task that are now available for our use. With the modified crystal

classification task, we can also more readily study issues associated with the proper use of machine learning methods on material science problems. With the St. John-Bloch pair as our baseline classifying pair and support vector machines[9,10] as our machine learning classifier, our average classification success rate is 92% with a 95% confidence interval of $[89\%, 95\%]$. We found the novel result that adding the excess Born effective charge[11] to the pair increases the average rate to 96% with a 95% confidence interval of $[93\%, 99\%]$. However, using just r_σ and the Born charge, we found an average accuracy of 97% and a confidence interval of $[95\%, 99\%]$. *When used alone, the excess Born effective charge classifies the solids as rocksalt or non-rocksalt (that is, ionic or covalent) with a remarkable accuracy of* 88%. The outstanding success accompanying the use of this novel feature made it difficult to move the success rate higher by using additional or other features. In general, using other features, unless one includes the excess Born effective charge, degrades the accuracy.

While machine learning methods have been used to predict the melting temperatures for classes of AB solids other than the octets, most recently, for example, by Saad et al.[6] for the suboctet AB solids, apparently little work has focused on predicting these temperatures for this special class of materials. What is also remarkable about the octets is the melting temperature data shows a 50% root-mean-square variation about its mean value. This variation is a challenge to any statistical inference method. What we found was that the small number of octets made the challenge even greater.

The challenge appears as we compare and contrast our melting temperature analysis with the very recent work of Seko et al.[8] These investigators made melting temperature predictions for a set of 248 binary compounds that included 46 octets we used. For each method of the four methods they used, their accuracy estimates for their training and testing sets were very consistent, and the accuracy predictions among three of their four methods were also very consistent. The fourth method, ε-support vector machines, produced a distinctively better fit to the data. As we report, for ε-support vector machines, the features, and data we used, we found consistency between the training and testing sets predictions hard to achieve. When combined with the standard machine learning method of cross-validation, the small number of octets with known melting temperatures makes the error estimate of the fit to the data itself subject to large errors. The average root-mean-square error of our fit is 11% (225°K), but that the standard deviation of this estimate was 67% (150°)K.

While our machine learning methods and feature sets have differences with those used by

Saad et al., Ghiringhelli et al., and Seko et al., the key differences between our work and theirs is our reporting the excess Born effective charge as a significant new feature for the boosting accuracy in the classification problem and also our reporting the significant sensitivities of the values predicted by our melting temperature models when cross-validation[9,10] was used as part of our machine learning. Finding the utility of the Born effective charge for classification is a novel result. In fact, we propose the excess Born effective charge[11] as a new scale for ionicity for these materials to replace the one proposed by Van Vechten and Phillips[12] several decades ago: It is measurable, better defined, and more easily and accurately calculated than an average ionic energy gap. While various materials machine learning applications have used cross-validation, including the three relevant to the octet AB solids,[6–8] we believe we are the first to report that this convenient and useful statistical method can experience difficulties with small data sets. Given the growing use of such methods in materials science, noting this possibility is important. We note that in bioinformatics, it has been realized that cross-validation can be unreliable when used on small datasets.[13–17] There, the data and feature sets tend to be at least an order of magnitude larger than those used in materials science.

In the next section, we discuss the AB solids we consider. We mainly review past classification and melting temperature efforts to underscore further differences between our and past work, and then we discuss the classification, regression, and cross-validation methods we used to model the data. In Section III, we first discuss the results for our classification task. Here we demonstrate the utility of using the excess Born effective charge as a feature. Next, we report results of melting temperature predictions. Stating a confidence interval for these results was important and difficult to provide. The estimated error of the fitting error had a one sigma variation about its mean of 67%. This makes knowing just the error of the fit (11%) a less useful result. In Section IV, we conclude with an assessment of our results and suggestions for future work, including its extension to formability and functionality studies of other classes of AB solids and to ABO_3 solids. In the Supplementary Material, we give tables of our Born charges.

II. BACKGROUND

A. Structure classification

Mooser and Pearson[18] appear to be the first to use a two-dimensional structure plot to classify AB solids. Their two features (that is, their x and y axes) were the average principal quantum number $\bar{n}$ of the two atoms and the difference ΔX in the Pauling electronegativity between the two atoms. With these features they were 90-95% successful in separating fourfold and sixfold co-ordinated octet AB compounds, which was a major breakthrough. Ten years later, Phillips and Van Vechten[12,19,20] introduced two new quantum-mechanical coordinates, namely the average covalent energy gap E_h and the average ionic energy gap C, both of which were based on a microscopic dielectric theory and newly available experimental spectroscopic data.[20] Phillips and Van Vechten showed that ionicity of the bond, more than its electronegativity difference, is a critical factor in classifying the crystal structure. With the two energy gaps as their features, they were able to separate exactly fourfold and sixfold coordinated binary compounds.

Perhaps the simplest and yet most efficient feature pair is the one proposed by St. John and Bloch[2] based upon the linear combinations of s and p orbital dependent radii r_s and r_p of the A and B atoms. In the notation of Chelikowsky and Phillips,[3] these linear combinations are

$$\begin{aligned} r_\sigma &= |(r_p^A + r_s^A) - (r_p^B + r_s^B)| \\ r_\pi &= |r_p^A - r_s^A| + |r_p^B - r_s^B| \end{aligned} \tag{1}$$

In the St. John-Bloch proposal, the r_l^X are the locations of the l-th orbital maxima of the eigenfunctions of a Simons-Bloch pseudo-potential for atom X.[1] With the r_l radii, Simons and Bloch argued that $S = (r_p - r_s)/r_p$ was a "structural" index for elemental solids with *sp*-bonding. St. John and Bloch subsequently argued that $X_l = 1/r_l$ was a measure of "orbital electronegativity" and defined the total atomic electronegativity to be[2]

$$X \equiv a \sum_{i=0}^{2} X_i + b \tag{2}$$

With a particular set of a and b, they found this expression fitted well both Pauling's[21] and Phillips's[20] electronegativity scales. The St. John-Bloch radii (1) are an unnormalized

extension of the above structural index and electronegativity difference to broader classes of solids with r_σ proposed as a measure of the electronegativity difference between atoms A and B, and r_π, as a measure of the average *sp*-orbital hybridization.

As shown by St. John and Bloch,[2] Chelikowsky and Phillips,[3] and others, with demonstrations perhaps pinnacling with the work of Zunger,[4,5] these combinations of radii have provided a solid foundation for an excellent structural classification of AB solids over a wide range of chemistries, including many AB solids with prominent *d*-shells. Chelikowsky and Phillips examined various correlations between this feature pair and the earlier pair proposals. While they found loose correlations, their general conclusion was these correlations were at best qualitative. In short, the St. John-Bloch pair is a distinctively different pair.

In addition to the above, we also note that Pettifor[22] put forward another chemical scale that evolved into $\mathcal{M}$, the Mendeleev number. This number allows a single two-dimensional structure plot for all AB solids. Each point in the plot is the pair ($\mathcal{M}_A$, $\mathcal{M}_B$). Although empirical, Pettifor's scale separates octet as well as non-octet AB solids structurally and separates fourfold and sixfold coordinated octet *sp*-*sp* bonded solids almost perfectly.

Despite the success of the St. John-Bloch pair and the Mendeleev numbers for structure classification, recent machine learning analyses have taken different paths to improve the accuracy of this classification pair by using hyper-dimensional structure plots. Instead of using the features in (1), Saad et al.[6] used the individual r_i^X. For other features, they used such quantities such as the valences and the ionization energies of the s and p orbitals of the A and B atoms. Ghiringhelli et al.[7] used machine learning methods to propose and assess a set of about 4500 composite features that are various functions of 23 "primary" features, including the St. John-Bloch pair, to find a pair, triplet, etc. that was the most effective for performing the classification. They claimed to find a new feature pair that was as good as the St. John-Bloch pair. They concluded however that in general more than two features are necessary to fit well the energy differences between the computed ground state energies.

We choose to start with the St. John-Bloch pair and study what happens if we add and subtract features from this pair. As we show in the Results section, we did not have to expand much to boost the accuracy of our classification to close to 100%. We gained accuracy because of the identification of a novel feature, the excess effective Born charge, as opposed to finding a larger set of features generated from functions of standard ones. We note that neither Saad et al. and Ghiringhelli et al. used the excess Born effective charge or

the Mendeleev numbers as features.

Born effective charges, also called dynamical, anomalous, and transverse charges, are measures of the local polarization density developed for one atom at the expense of the other.[11,33] The Born effective charge for a given atom is defined as the change in electric polarization divided by the amount a periodic sublattice of equivalent atoms is displaced. For sublattice k the effective Born charge is[11]

$$Z^k_{\alpha\beta} = \Omega_0 \left. \frac{\partial P_\beta}{\partial \tau_{k\alpha}} \right|_{E=0} \tag{3}$$

where P_β is the macroscopic polarization along the β direction, with collective nuclear displacements $\tau_{k\alpha}$ of sublattice k along the α direction. Ω_0 is the unit cell volume. The derivative is evaluated in zero electric field. The Born effective charge is a tensor object. In our analysis, because of our specific choices of the crystal structures, this tensor was simply a constant times the identity matrix. The value of the Born effective charge usually differs from the nominal valency of the atom. The excess charge ΔZ_A is simply the difference between the computed effective charge for the A atom and its nominal valence charge. *We found that the excess Born effective charge is almost always positive for rocksalt solids and negative for most non-rocksalt ones.*[23] When it misclassifies, its magnitude is small. *The sign of the excess is more effective in the classification than its magnitude.*

Density functional perturbation theory[24,25] combined with the modern theory of polarization,[26–28] as implemented in Vienna *ab initio* simulation package (VASP)[30] within the local density approximation (LDA),[29] was used to compute the Born effective charges. The electronic wave functions are expanded in plane waves up to a cut-off energy of 500 eV. The pseudopotentials based on the projector augmented wave (PAW)[31] method and Monkhorst-Pack sampling[32] of the Brillouin-zone integrations were used. To obtain a geometry-optimized equilibrium structure, atomic positions and the lattice parameters were fully relaxed using the conjugate gradient method until the stress components in all directions were less than 1.0×10^{-3} GPa. Our computed Born effective charges are found to be in good agreement with the corresponding experimental values[34–36] for a number of binary systems that exist in RS and ZB crystal structures. Interestingly, we found that a functional form proposed by Harrison[33] correlates with the computed values. However, as a general and well known trend, tight-binding tends to overestimate the effective charges.[36] In particular, this underestimation was noted by Bennetto and Vanderbilt[37] and subsequently mitigated by several

other authors.[38,39] Further details about the Born effective charges are discussed in the Supplemental Material.[23]

B. Melting temperature prediction

It can be argued that Chelikowsky and Phillips[3] were among the first to use machine learning methods to predict melting temperatures. Performing least-squares fits of functions to data is a simple form of machine learning. Viewing the melting temperature as being a function of the St. John-Bloch and then the Mooser-Pearson feature pairs, they fitted the data of 44 suboctet solids (A^NB^{N-P}, $3 \leq P \leq 6$) to quadratic polynomials of these different feature pairs. Recently, Chelikowsky and coworkers[6] revisited this same set of AB solids with more advanced machine learning methods, used a larger set of features in their fitting, and sought to identify which features were the most relevant. About half of the features they used for melting temperature predictions differed from those they used for classification.

As noted in the Introduction, Seko et al.[8] even more recently used machine learning to predict melting temperatures and studied the effectiveness of four regression methods (ordinary least-squares, partial least-squares, support vector machine, and Gaussian process) for a set of 248 single and binary compounds that included our 46 octets (and the 44 suboctets of Saad et al.). Their 23 features emphasized elemental features (atomic number and mass, the group and period in the periodic table, van der Waals and covalent radii, etc.) supplemented with either those computable with DFT or those available from other sources. These quantities included cohesive energy, bulk modulus, volume, and nearest-neighbor distance. They concluded that their regression was more accurate with an elemental and DFT-computed combination than with an elemental and measured combination. As Saad et al., they sought to find which of their features were the most relevant to the accuracy of the fit.

Our methods of analysis are much more similar to those of Seko et al. than to those of Saad et al.: We both use ε-support vector machines and similar cross-validation techniques. They however used four different machine learning methods while we used several different forms of one method. Although we used the same machine learning method, the ε-support vector regressor, with which they clearly produced their smallest root-mean-square error, we obtained our best results with a different kernel, a third-degree polynomial instead of a

Gaussian. As also previously noted, unlike them we had more difficulty producing acceptable results. We had to address more explicitly issues of underfitting versus overfitting of the data. The 11% root-mean-square error of our fit corresponds to an error of 225°K. Seko et al. estimate their error to be approximately 265°K. The melting temperatures in their larger data set span a similar range of values as our much smaller data set. We used the same features for melting temperature predictions that we used for classification. Seko et al. did not attempt any structural classification.

C. Machine learning

Support vector machines are commonly used in binary classification problems.[9,10] For binary classification, each instance of our data is described by a vector of features $\vec{x} = (f_1, f_2, \ldots, f_n)^T$ and a label y. The label has a value of $+1$, say for rocksalt, and -1, for non-rocksalt. A support vector machine finds a function that for any given $\vec{x}$ has a value of ± 1.

Ideally a support vector machine generates a hypersurface (decision boundary) in the space of features that maximizes the distance of the closest instance from either class from it.[9,10] This maximal distance is called the margin. Instances on the margin are called support vectors. Instances are points in the hyperspace of features and lie on one side or the other of this hypersurface. Depending on which side they lie they are classified as $+1$ or -1.

In general a clear separation of the data via a finite margin is not possible so a soft margin support vector machine is constructed instead. This classifier allows misclassification of instances; that is, it allows points in the margin. If we represent our input data by the set of labeled instances $\{(\vec{x}_i, y_i)\}$, then a soft margin support vector classifier determines the hypersurface in the space of features by solving

$$\alpha_1^*, \ldots, \alpha_n^* = \underset{\alpha_1,\ldots,\alpha_m}{\arg\min} - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\mathcal{K}(\vec{x}_i, \vec{x}_j) + \sum_{i=1}^{m}\alpha_i \tag{4}$$

subject to

$$0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{m}\alpha_i y_i = 0. \tag{5}$$

Adjusting C controls the number of misclassifications. In the minimization the competition is between the size of the margin and the degree of misclassification acceptable. The support vectors are now those $\vec{x}_i$ for which $0 < \alpha_i < C$.

$\mathcal{K}(\vec{x}_i, \vec{x}_j)$ is called the kernel. Three are three common choices: A linear kernel

$$\mathcal{K}(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j \tag{6}$$

a polynomial kernel

$$\mathcal{K}(\vec{x}_i, \vec{x}_j) = (\gamma \vec{x}_i \cdot \vec{x}_j + r)^d \tag{7}$$

and a Gaussian kernel (radial basis function)

$$\mathcal{K}(\vec{x}_i, \vec{x}_j) = \exp(-\gamma |\vec{x}_i - \vec{x}_j|^2) \tag{8}$$

Unless otherwise stated, we used the software in scikit-learn[40] for all the machine learning procedures used in this paper.

If the kernel is linear, the decision boundary is always a hyperplane. If the number of features is two, the linear kernel support vector machine draws a straight line through the data and hence is analogous to the pencil-ruler method used on a structure map. Past structure maps, however, were multi-class classifiers as they used several straight lines to separate the materials into more than two crystal structures. While with several lines and the eye, separating the data quite cleanly into multiple classes was possible, using one line to separate the data into two classes is not possible. The principal reason we choose support vector machine over other classification methods is a linear kernel and data with just two features mimics what was done in the past.

For predicting melting temperatures, we used the ε-soft-margin support vector machine regressor, the same as used by Seko et al. but with different kernels. This method adds an additional constraint, scaled by ε, to the minimization problem. The additional constraint introduces a new parameter α_i' for each instance and allows misclassified instances only if they are within a distance ε of the decision boundary.

We used the machine-learning technique of cross-validation[9,10] to assist in estimating the level of confidence, that is, the errors, of our results. In cross-validation the model is not fitted to the entire data set but rather the data is first spilt into training and testing sets. The model is fitted to the training set and then is validated by using the test set. As discussed below, often we nested cross-validations: In fitting the model to the *training data*, we would use a k-fold cross validation. This procedure randomly divides the training data into k subsets of roughly equal size. Of the k subsets, one is used as the test set with the remaining $k-1$ subsets used as the training set. Each subset is used once as the test set.

The average of the k test-set scores produces estimates of the accuracy of the fit on the training data and on the testing data.

Cross-validation produces a model fitted to the data that is more predictive of what to expect if new data is added to the data set. This type of model is most germane to the design and discovery of new materials.

III. RESULTS

Our dataset has 75 instances of octet AB solids, each described by the same 10 features. Our core feature set is: (1) r_σ, (2) r_π, (3) the valence of the A atom v_A, (4) the excess Born effective charge of the A atom ΔZ_A, (5) $\mathcal{M}_A$, (6) $\mathcal{M}_B$, (7) the difference between the Pauling electronegativities of A and B atoms ΔX, (8) the difference between the ionization potentials of A and B atoms $\Delta\chi$, (9) the nearest-neighbor distance $d_{\rm DFT}$ for the crystal structure, and (10) a bond polarity measure α_p.[33] For each instance, we also know its rocksalt/non-rocksalt label. For the r_i^X, we used the values in Table I of Saad et al. to compute r_σ and r_π and used their Table IV for ionization potentials. We used Pauling's values for the electronegativity. Our Born effective charges and nearest neighbor distances were computed with density functional theory (DFT), using a rocksalt structure for all the solids we grouped into this class and with a zincblende structure for all the materials we grouped as non-rocksalt. In the Supplemental Material,[23] we note our computation of these charges compares well both with measured values[34] and with tight-binding predictions from a formula due to Harrison.[33] These agreements point to the possibility of using effective Born charges obtained by means other than DFT calculations.

We used the Ghiringhelli et al. 82 octet AB solids but without CuF, which seems not to exist,[6] and without BSb, GeC, SnC, GeSi, SnSi, and SnGe whose crystals structures are unlisted by both Zunger and Pettifor. We note Zunger classified 112 octets, including CuF, into six crystal structures. CuF was one of his five misclassifications. The other four were: BeO, MgS, MgTe, and MgSe. Of these 75 solids, the melting temperatures of 46 are known. We used the melting temperatures from Table V of Seko et al.

A. Binary classification

To use a support vector machine classifier, we have to select a kernel and set its parameters. To aid in doing these, we used a grid-search cross-validation[40] that generates for each of four kernels we studied (linear, polynomial of degree 2, polynomial of degree 3, and the radial basis function (RBF)), a one, two or four dimensional grid. These numbers are one (for C) plus the number of parameters in the kernel. For each kernel at each grid point, we used a 5-fold cross-validation on a 0.9/0.1 training/testing split of the dataset. We set $k = 5$ and the initial 0.9/0.1 split after some experimentation. Our metric of success is the accuracy, that is, the number of instances in the test set predicted correctly divided by the number of instances in the test set. For this metric the grid often had a number of points with nearly identical values. For each kernel, however, grid points with any of C, γ, and r less than 1 performed noticeably poorer. Instead of choosing the parameters values at the grid point with the best value of the metric for a given kernel, we simply choose $C = \gamma = r = 1$ to define the models for whatever kernel we used. With it and the right combination of features, we were able to achieve excellent classification for all four kernels.

With the models set, we classified the data using four kernels and the (r_σ, r_π) feature pairs and no cross-validation. The predictions of the models applied to the *entire* dataset are shown in Fig. 1. The linear and third degree polynomial kernels misclassify 4 instances; the second degree polynomial and the radial basis function kernels, 3. The misclassifications in Figs. 1 involve 3, 4, or 5 instances, with 3 consistently being MgS, MgTe, and MgSe and the others depending on the subtle shifts in the decision boundaries. The three consistently misclassified were 3 of the 4 misclassified by Zunger.[5]

Next we redid the classification using the same four kernels and nested 5-fold cross-validation on a 0.9/0.1 training/testing split of the data. As the parameters of the kernel are set, cross-validation here is being used to quantify the expected accuracy of the four models. The resulting predictions for the entire dataset are in Fig. 2. The results look very similar to the plot of the non-cross-validated case, but here the linear and the second and third degree polynomial kernels misclassify 4 instances; the radial basis function kernel, 5. The encircled points are the members of the test set. We note that in this analysis the training set size is slightly smaller than the one in the previous figure.

The results in this figure and the previous one suggest relatively accurate classifications,

but variations in the predictions from one analysis to another exist. In particular, we found variations of 10 or more percentage points in the predictions of a 0.9/0.1 training/testing split with 5-fold cross-validation compared to a 0.8/0.2 split with 5-fold cross-validation, of a 0.9/0.1 split with 5-fold cross-validation compared to a 0.9/0.1 split with 10-fold cross validation, a 0.9/0.1 split and 5-fold cross validation repeated with a different random number sequence, etc. While it is expected that the predictions would not be identical, this large of a variation made it difficult to state the accuracy of the models fitted to the data. Undoubtedly, because of the small size of the data set, the different random training/testing and cross-validation splits generate data subsets that are not statistically equivalent.

To state the accuracy of the predictions with a level of confidence, we decided to use

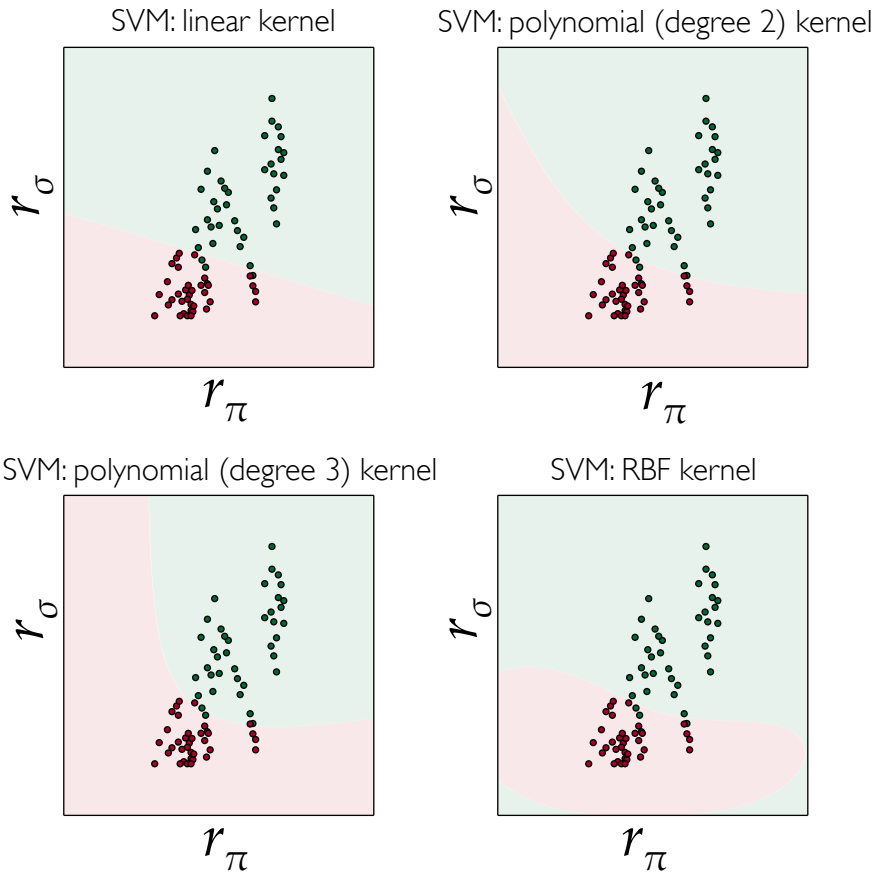

FIG. 1. Comparison of different optimizers for classification based on just the feature pair (r_σ, r_π) but using all the instances. The green region is rocksalt.

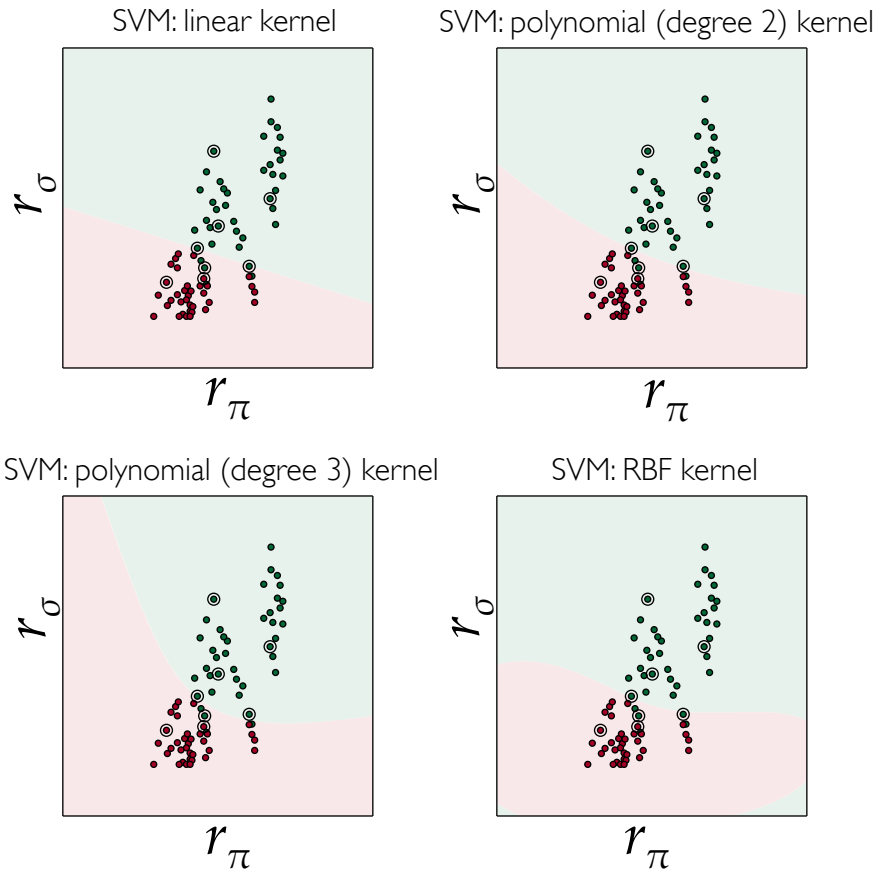

FIG. 2. Comparison of different cross-validated optimizers for classification based on just the feature pair (r_σ, r_π). Instances double-circled are the test set. The green region is rocksalt.

just the linear classifier, a 0.9/01 split, and a 5-fold cross-validation. We then repeated the analysis multiple times, collecting statistics of the predictions. For a modest number of repetitions, say 30, we found the mean and median of the predictions were unequal. The implied skewness of the predictions indicates that the computation of a standard deviation would mislead as an indicator of the statistical error. We thus choose to repeat the analysis for 10000 times and create empirically the probability distribution of the results. For this large number of repetitions, the mean and median were approximately equal, but as can be seen in Figs. 3 and 4, the histograms in general did not always approximate a Gaussian. One issue is that for several of our feature subsets, the accuracy was so high that fluctuations above the mean were bounded within a few percentage points by 100% while those below

the mean could range more. This situation skewed the distributions.

To do this statistical analysis, we first switched our cross-validation to a 5-fold stratified cross-validation.[40] This differs from 5-fold cross-validation in that the relative proportions of positive and negative labeled data in the training and test sets are roughly the same as in the entire set. To reduce statistical correlations among the predictions, we shuffled the data before doing any 5-fold stratified cross-validation. Although the accuracy metric lies in $[0, 1]$, we plotted it in the reduced interval $[0.85, 1]$, divided this interval into 30 bins, histogramed our 10000 samples, and normalized the area under this curve to be unity. In Fig. 3, we show the results for the St. John-Bloch pair and that pair when the excess Born effective charge is included. Besides the average, we also report the median. The averages and medians are nearly the same. For reference, we also report the standard deviation, but chose another way to assign a confidence interval for the above results: We chose to state an interval $[x_{\min}, x_{\max}]$ of minimum width, positioned so that it captures 95% of the area under the generated distribution. The discreteness of the curves in Fig. 3 enables doing this only with some subjectivity. On the basis of the results in these figures, we claim with 95% confidence that the accuracy of of the St. John-Bloch pair classification lies in the range $[0.89, 0.95]$ while that of the feature triplet lies in the range $[0.93, 0.99]$. We rank the

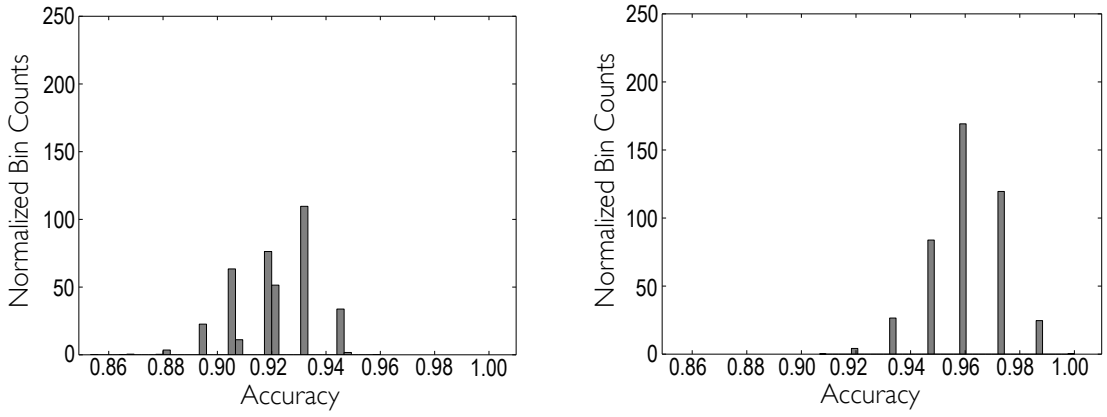

FIG. 3. The normalized histogram of accuracy of the binary classification using a linear soft-margin support vector machine. Left: (r_σ, r_π). The average accuracy is 0.922; the median accuracy, 0.920; and the root-mean-square error, 0.015. Right: $(r_\sigma, r_\pi, \Delta Z_A)$. The average accuracy is 0.961; the median accuracy, 0.960; and the root-mean-square error, 0.014.

accuracy of feature tuples in Fig. 3 as $(r_\sigma, r_\pi) \prec (r_\sigma, r_\pi, \Delta Z_A)$ and claim that adding the specified feature increases the accuracy of the classification.

While we did not explore all possibilities, it does seem that adding single or double features to the St. John-Bloch pair generally degrades performance relative to just using the St. John-Bloch pair. Exceptions mostly appear if the excess Born effective charge is present. We propose the following rankings: $(r_\sigma, r_\pi, v_A) \preccurlyeq (r_\sigma, r_\pi) \prec (r_\pi, r_\sigma, v_A, \Delta Z_A) \preccurlyeq (r_\sigma, r_\pi, \Delta Z_A)$. We also claim that $(\mathcal{M}_A, \mathcal{M}_B) \prec (r_\sigma, r_\pi, \mathcal{M}_A, \mathcal{M}_B) \preccurlyeq (r_\sigma, r_\pi)$ and that $(r_\pi, \Delta Z_A) \prec (r_\sigma, r_\pi) \preccurlyeq (r_\sigma, r_\pi, \Delta Z_A) \prec (r_\sigma, \Delta Z_A)$. For the purposes of comparison, we give in Fig. 4 the accuracy histogram for the Mendeleev pair and for the Mendeleev pair when the excess Born effective charge is added. These results individually and collectively identify ΔZ_A as an important new feature for crystal classification.

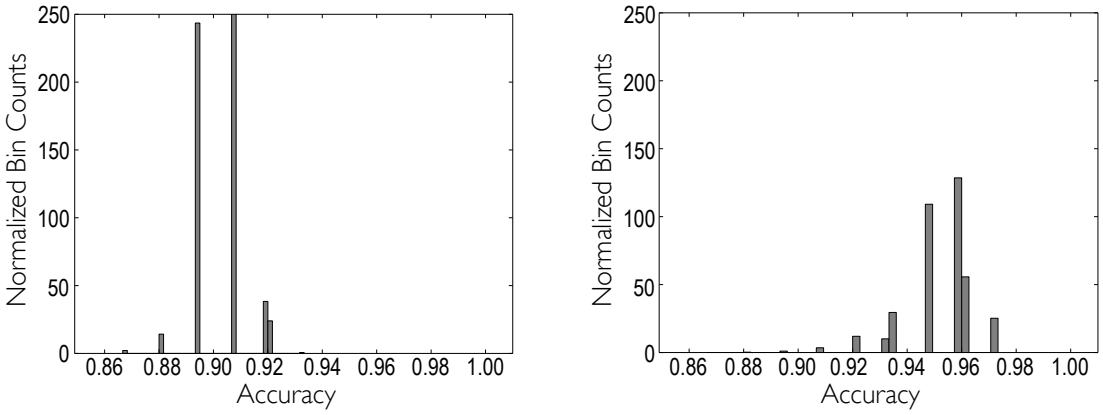

FIG. 4. The normalized histogram of the accuracy of the binary classification using a linear soft-margin support vector machine. Left: $(\mathcal{M}_A, \mathcal{M}_B)$. The average accuracy is 0.902; the median accuracy, 0.907; and the root-mean-square error, 0.010. Right: $(\mathcal{M}_A, \mathcal{M}_B, \Delta Z_A)$. The average accuracy is 0.952; the median accuracy, 0.960; and the root-mean-square error, 0.013.

B. Melting temperature predictions

For melting temperature predictions, our task shifts from classification to regression. For regression we want to learn a model $\hat{f}(\vec{x})$ that is as successful as possible in predicting a real number $\hat{y}$ associated with our data. We use the same training/testing split of the data

with k-fold cross-validation applied to the training data to help set our regression models, but use the root mean-squared metric to quantify fit quality.

The most important change in moving from classification to regression is the data set size being reduced from 75 to 46, as the melting temperatures are known for just this number. This smaller size made assigning confidence intervals for the predictions difficult and necessitated the adoption of additional cross-validation procedures. The challenge is properly balancing bias and variance in the predictions:[9,10] Bias makes the predictions inaccurate; variance makes them uncertain. Bias can occur when the data is underfit; variance, when it is overfit. Overfitting is typically caused by using too many parameters; underfitting, by using too few. Besides the parameters needed to specify the kernel in the support vector machine, as can be seen from (4), each instance adds a potential parameter. Which instance participates and the number of participants depends on the number of α_i and α_i' that satisfy $0 < \alpha_i, \alpha_i' < C$. This number can vary as the training data and kernel parameters change. It is unclear whether a higher degree polynomial kernel necessarily means there are more parameters to fit. In the constrained optimization problem represented by the support vector machine regressor, values of C and ε too small or too large by themselves can cause under and overfitting, likely by increasing and decreasing the number of support vectors.

Using our core feature set, we started setting the model in the same way we did for classification by using a grid search cross validation nested with 5-fold cross-validation on a 0.90/0.10 training/testing split. Besides a linear and a RBF kernel, we considered polynomial kernels of degree 2 through 7. A coarse grid lead us to focus on only polynomial kernels of degree 2 through 5.

The grid search results for the reduced data and core feature sets showed sensitivity to the random number sequences used in the training/testing split, the percentage of the split, and the choice of k in the k-fold cross-validations. Our metric was the root-mean-square error of the predicted melting temperatures of the test set. Normally we would want to adjust the model so this score is as small as possible. While the grid search generates a number of parameters giving seemingly good or bad scores, it provides little information about whether these scores are consequences of under or overfitting. To assess whether these issues were present, we experimented with various split ratios of the data and number of cross-validation folds. In many respects, the size of the training set became another parameter specifying the model.

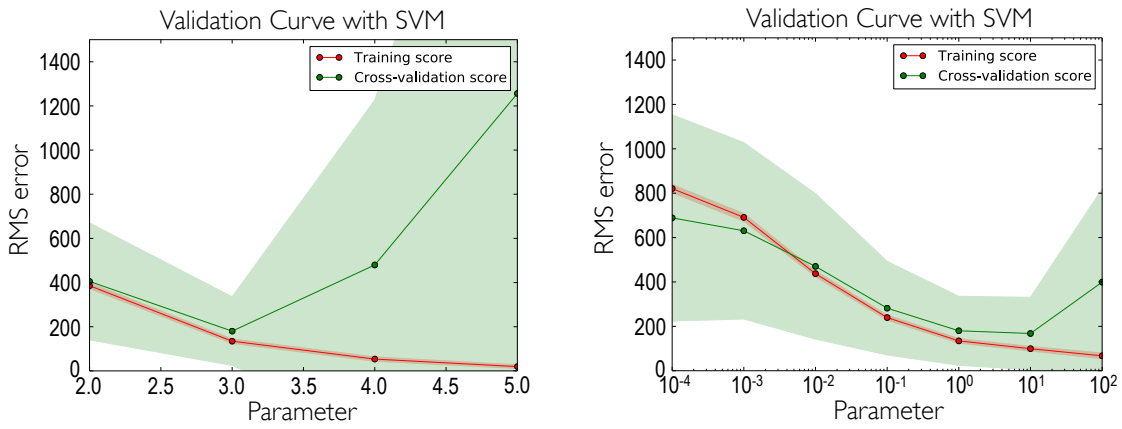

FIG. 5. Validation curves for the ε-soft-margin support vector machine regressor. On the left, $C = 1$ and the parameter being varied is the degree d of the polynomial kernel from 2 to 5. On the right, $d = 3$ and the parameter being varied is C

To reduce the importance of the training set size, we proceeded in the following manner: We set some of our parameters from the grid search, finding γ, ε, and r to have relatively large ranges of variation with nominal effect on results. We choose $\gamma = 1$, $\varepsilon = 0.01$, and $r = 0$. More variation was shown in the choice of the d for the polynomial kernel and the value of C.

To set d and C, we used a validation curve which is simply a plot of the cross-validation and test scores (the values of the metric) as one of these parameters is varied. Here we used various N/M training/testing splits. They were created by randomly selecting M instances to be in the test set. For cross-validation on the training data, we used the Leave P Out method.[40] Here there is no random selection of the training subsets but rather for the N instances in the training data all $N!/(N-P)!/P!$ subsets of $(N-P)$-sized training subsets are used and their scores on the P test subsets are evaluated and averaged to produce a training score. In this procedure, for small values of P, which are the only ones practical because of the exponential growth in the number of possibilities with increasing P, the model always fits the training data well as it is being fit to almost all the data. For each fit the scores for the M instances in the test set are calculated and averaged to produce a cross-validation score. This procedure was repeated 200 times to generate a mean test score and its standard deviation. In Fig. 5, on the left, we show the validation curve for $C = 1$ as we varied d; on the right we show the curve with $d = 3$ as we vary C. In both cases we used

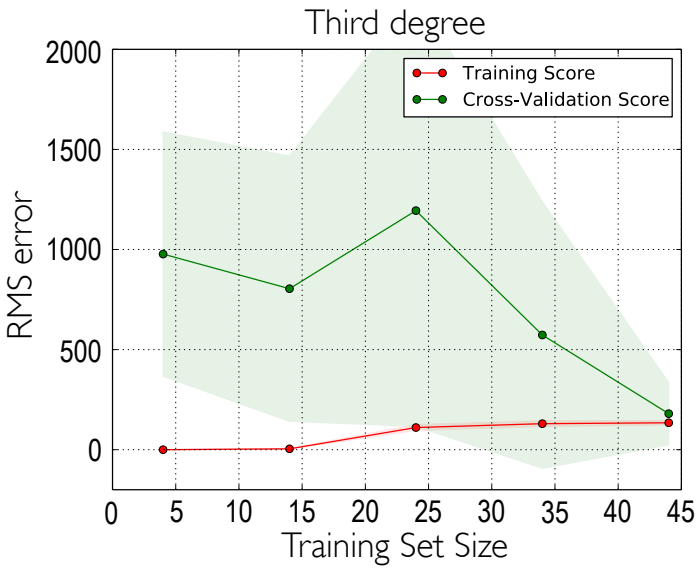

FIG. 6. The learning curve for a kernel polynomial of $d = 3$, $C = 1$, $\varepsilon = 0.01$, $r = 0$, and $\gamma = 1$.

$P = 2$. We see $d = 3$ has the smallest cross-validation score and it has the smallest standard deviation for the score. A similar result with slightly smaller root-mean-square error was obtained with $P = 3$. For $P = 1$, the Leave One procedure, we got a similar result but the variances as a function of parameter d were much larger.

We now use learning curves to study the sensitivity of our model to changes in the training set size.[40] A learning curve is simply a plot of the cross-validation and test scores of a model as a function of the training set size. In Fig. 6, we show these the cross-validation and training scores for $d = 3$ and $C = 1$. We used a nested Leave 2 Out cross-validation which allowed the training set size to vary from 3 to 45. When the training set size is small, the model overfits the data, and the training score is small and the cross-validation (testing) score is large. The scores converges around the maximum size of the training set. With convergence, the analysis has reached a point where adding additional data will not improve the results.[10] A more complex model, and possibly more features, are needed to decrease the error. In Fig. 7, we show the same analysis but for $d = 2$ (on the left) and $d = 4$ (on the right).

The lower and nearly equal cross-validation and test scores score for $d = 3$ and $C = 1$ is our basis for selecting this set of parameters as defining our model and saying with 68% confidence the root-mean-square error in its melting temperature predictions lie in the

interval $[25°\mathrm{K}, 325°\mathrm{K}]$.

To set C and d, we actually iterated a few times between using validation curves as a function of C and learning curves as a function of d and vice versa. We also replaced the Leave P Out cross-validation with a shuffle-split cross-validation procedure that we repeated for several hundred times. Previously, we shuffled only before the k-folds were selected. Here shuffling occurs before each training/cross-validation set is selected. While noisier than those from the Leave P Out method, the curves were similar.

For other models, the behavior of the cross-validation and testing scores in the learning curve generally behaved like those in Fig. 7. For a small training set sizes the cross-validation score is large, and the training score, low. As the training set size increases, the cross-validation score decreases and the test score increases. When the set size reaches its limit, either these scores were far apart, but within the statistical error of the cross-validation score, or they are beginning to meet. If they are meeting, their average scores are typically large. In all cases, the statistical error of the cross-validation score is nearly equal to the value of the score.

For melting temperature analysis, as for our classification analysis, it is important the distinguish the error computed with cross-validation, which is a prediction based on a subset of the data, and the error of the model computed for the entire data set. In Fig. 8, we plot our model predictions applied to the entire data set versus the experimental values. From

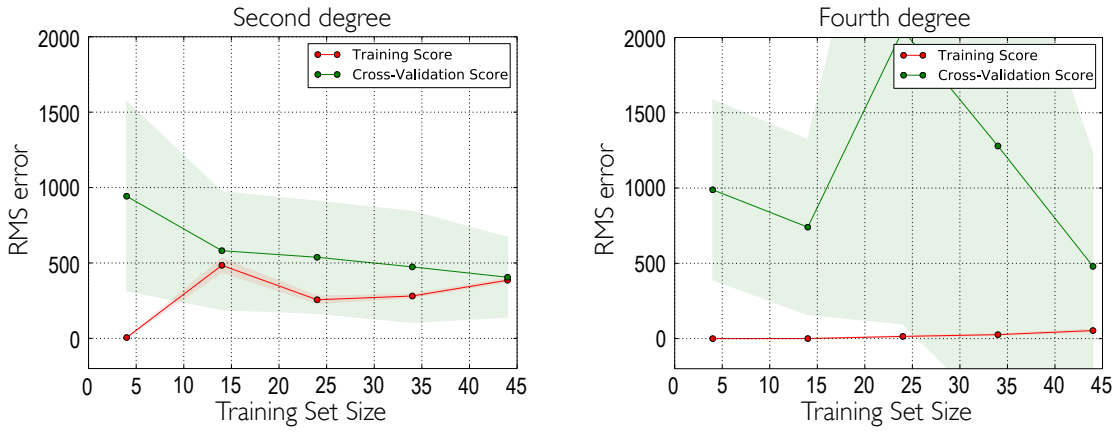

FIG. 7. The learning curve for a kernel polynomial of $C = 1$, $\varepsilon = 0.01$, $r = 0$, and $\gamma = 1$. On the left, $d = 2$; on the right, $d = 4$.

the figure, we see that we actually have an excellent fit except for about 4 to 6 compounds. Here the root-mean-square error of the predictions is 134°K, a bit more than half of the error predicted by cross-validation. We also note that different accuracy metrics portray this fit in an even more favorable manner: The average absolute deviation of the predictions is 47°K; the median absolute deviation is 0.18°K. The latter is the size of the deviation between the absolute values of the predictions and measurements that splits the absolute deviations in half. The error computed by cross-validation however is more indicative of the error expected if a new solid were added the the data set. It is the error more appropriate for materials design. The object is not producing fits to the data that remember the data well but predicting from the data with confidence in the predictions.

Our intent was to systematically add features to the St. John-Bloch pair and observe how the additions made a difference. We are reporting only our best result. With fewer features consistent results were harder to obtain.

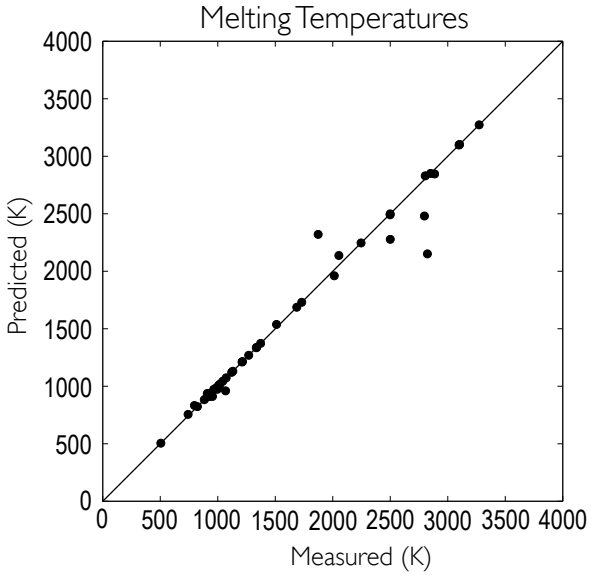

FIG. 8. The melting temperatures computed for a kernel polynomial of $C = 1$, $\varepsilon = 0.01$, $r = 0$, and $\gamma = 1$ plotted against the experimentally measured values.

IV. DISCUSSION

Our main new finding is that replacing r_π with the excess Born effective charge of the A atom improves the accuracy of our structural classification task significantly. Our application of machine learning methods to the classification of crystal structures and the prediction of the melting temperatures of the octet AB alloys produced models that in the best cases classify the crystal structure with a success rate of 99% and predict the melting temperature with an error of 2% of the data's mean. In the worst cases, they classify the crystal structure with a success rate of 96% and predict the melting temperature with an error of 21% of the data's mean. The worse case is at least as good as previously reported[6] averages.

The better success with classification has to do with the St. John-Bloch pair being an excellent classifier to build upon. With this pair alone and using the pencil-ruler method on the data in Fig. 2, the complete data set has at best only four misclassifications. With the number of instances being 75, this means that using this pair the best accuracy is 94.7%. It is interesting to note that adding extra features to this pair does not necessarily increase the accuracy. If one of the added features is the excess Born effective charge, then its presence generally helps. For our data, machine learning methods, and features, we gave an explicit demonstration that $(\mathcal{M}_A, \mathcal{M}_B) \prec (r_\sigma, r_\pi) \prec (r_\sigma, \Delta Z_A)$ and thus have found a more effective two-coordinate structure map of the type Chelikowsky and Phillips plus others were searching for. As r_σ defines an electronegativity scale and ΔZ_A defines an ionicity scale for the solids, our proposal updates the electronegativity and ionicity feature pair proposed by Van Vechten and Phillips[12] several decades ago, establishes this new scale as more effective (at least for this set of solids), and affirms their belief about the ionicity being a critical factor in the classification.

Our result of identifying the excess Born effective charge as an important feature is novel. To appreciate more fully its significance, we present in Fig. 9 the same type of structure plot for the $(r_\sigma, \Delta Z_a)$ feature pair as we did for the St. John-Bloch pair in Figs. 1. We clearly see that with the new feature pair the linear and cubic support vector machines, without the cross-validation training, misclassify only the same single solid (InN), while the quadratic and quartic machines misclassify this solid plus ZnO. Both solids are wurtzites and the Born effective charge was computed assuming they were zincblendes. The linear kernel is an automated version of the pencil-ruler method for constructing a structure map.

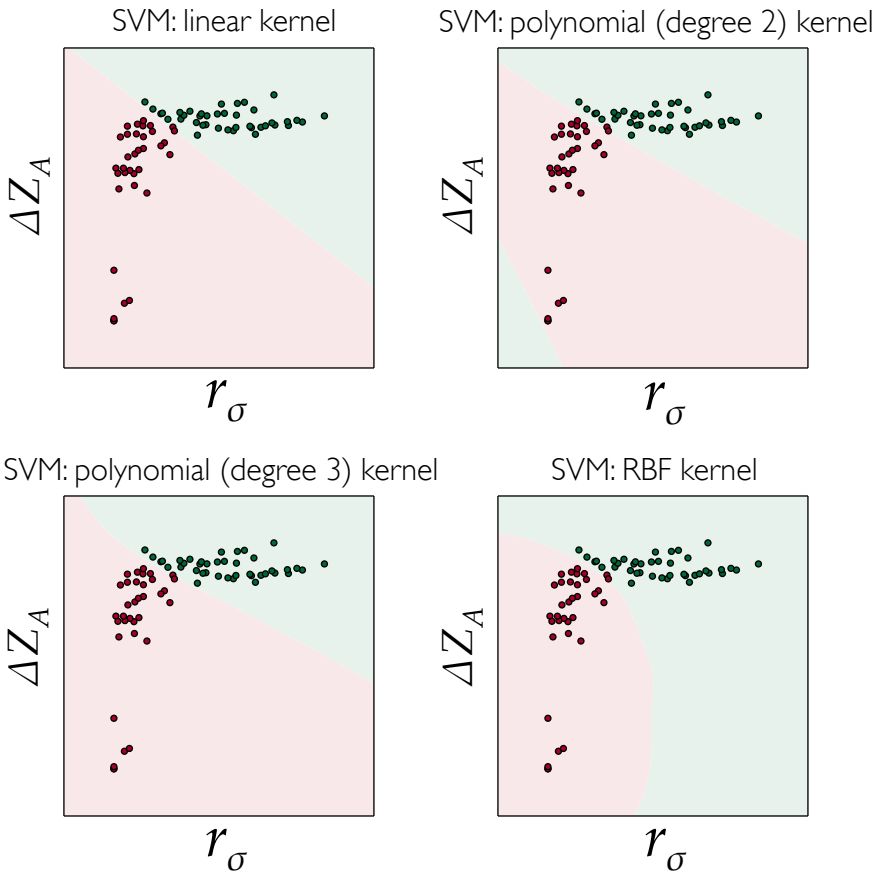

FIG. 9. Comparison of different cross-validated optimizers for classification based on the feature pair $(r_\sigma, \Delta Z_A)$. The green region is rocksalt.

We recall this automation of the drawing of the straight line for the St. John-Block pair structure map produced 3 misclassifications (Fig. 1a). With the new feature pair we have just one.

Born effective charge are central to the long wavelength LO-TO phonon splitting in polar crystals and to spontaneous polarization of materials. If a relative displacement of the sublattices is made (that is, an optical phonon), the energies of the bonds on one side of an atom are lowered relative to those on the other side. The generates a charge transfer from one side to the other and produces a polarization with an effective charge. In some materials this charge is anomalously large.[41] The Born effective charge reveals the mixed ionic and covalent charge of a bond.[42]

We offer that the excess charge succeeds as an effective classifier mainly because of its sign as distinguished from its magnitude. For example, if we used just the Born effective charge, the sign change would disappear, and instances of octets with similar values of the effective charge would have atoms with different nominal charges. Using the excess charge produces a feature more distinctively poised to function effectively as a classifier.

Why does the sign of the excess Born effective charge differ between rocksalt and non-rocksalt? (*c.f.* Table I, Supplemental Material).[23] This question would be best answered by doing a band-by-band calculation of the Born effective charge for each atom[42] in the solid. Doing this allows the identification of the charge different orbitals contributing to the atom's total dynamical charge. These types of calculations have been more actively pursued for perovskite materials than AB solids. For AB solids, except for non-octet ZnO, few calculations for cubic materials exist. For ZnO, a wurtzite, the excess charge is nearly zero for the various orbitals.[43,44] Hence the total excess is small. For $PbTiO_3$ and the non-octet α-PbO, the Pb $5d$ and $6s$+$2p$ orbitals show opposite trends with respect to the excess charge, making the first more negative and the second less negative.[45] With the apparent exception of a study of the non-octet cubics MgO, CaO, SrO, and BaO,[44] a systematic study of Born effective charges within and across various crystal structures apparently has yet to occur.

Despite not having direct calculations of this subtle process to infer from, we offer the following hypothesis as to why the excess charge changes sign: The origin of anomalous effective charge is linked to a polarization effect created by intrasite hybridizations of occupied orbits and charge transfer effects created by intersite hybridizations of unoccupied and occupied orbitals.[42] The strengths of the latter hybridizations are unimportant. What is important is the rate of change of the existing hybridizations between occupied and unoccupied or those induced by the optic-mode-like displacements of the A and B atoms relative to each other. Convention has the A atom to have a smaller electronegativity than the B atom. With a positive excess charge a fraction of unity, the ionic AB solids have an A ion that has almost completely transferred its nominal valence to the B atom. The small excess charge is mainly a polarization effect. For the ionic/covalent AB solids, the A ion has transferred most of its charge from s-orbitals to the p-orbitals of B ion but charge transfer from A to B through hybridizations with the d-orbitals of A to orbitals of B is incomplete. In short, the situation for the ionic/covalent AB solids mimics that of $PbTiO_3$ and α-PbO. The degree of

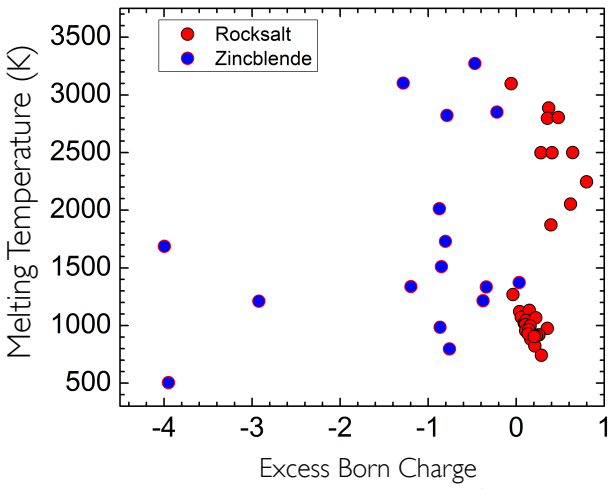

FIG. 10. Melting temperatures of the octet AB solids versus the excess Born effective charge. The rocksalts are green circles and the non-rocksalts are depicted in red.

incompleteness of the charge transfer serves as measure of the degree of ionicity in a bond tending to be otherwise covalent.

Our secondary finding is the degree of confidence that we could place in our machine learning predictions was strongly affected by our data sets being small. Here we rediscovered what is already known in bioinformatics[13–17] where data and feature sets are typically larger than those presently available in the materials sciences. The much larger error in predicting the melting temperature in a large part has to do with this data set having a large range in melting temperatures. The average melting temperature of the data is 1580°K with a standard deviation of 791°K. We note that if applied to the entire data set, that is, without the cross-validation training, our model predicts an average melting temperature of 1562°K and a standard deviation of 767°K. Hence, it faithfully reproduces the data. However, large error associated with this predicted error is mainly due to the small size of the data set. With cross-validation training, the standard deviation estimate of the average standard deviation is 150°K which decreases confidence in any melting temperature predictions.

What is not understood, and not explicitly adressed here, is why the octets show such a wide range in melting temperture. This dispersion in melting temperatures is generally the case for most classes of AB solids,[46] including the simpler subocets considered by Che-

likowsky and co-workers.[3,6,47,48] As noted by Chelikowsky and Anderson,[47,48] features useful for classification generally show little correlation with the melting temperature. We illustrate this in Fig. 10 for the excess Born effective charge. The excess Born effective charge is an indicator of the type of bonding; for melting temperatures predictions, it would seem we need at least some features that are indicators of bond strength.

The principal physical model of melting remains the venerable Lindemann's criterion.[49] For the octets, there is also the scaling theory of Van Vechtan. As noted by Van Vechtan,[50] using Lindemann's criterion requires an estimate of the Debye temperature which is often hard to infer from data. His scaling theory obviates the need to know this feature for other than one instance of data from which the melting temperatures of other instance are obtained by scaling. His theory has limited generality. His results for 24 zincblends octets showed about a 35% root-mean-square deviation relative to the mean. The mean temperature was 1785° K. The data had a standard deviation of 857° K. More recently, Kumar et al.[51] estimated the Debye temperatures from data fits to the zincblende octets, and with Lindemann's criterion, they reduced the average deviation between fitted and observed values by a factor of two relative to Van Vechtan's results.

Physical models such as those of Lindemann and Van Vechtan have a different basis than the statistical models of our work and the recent machine leaning works of Saad et al. and Seko et al. Both types of models are predictive, but the statisical models are simply fits to generic, convenient functions and not fits to explicit functional forms suggested by the physics. As we illustrated in the present work, fitting too few data to too many features can result in overfitting that leads to large variance in the predictions. In such fits, one would like to use features that are physically relevant in the sense that they correlate with the objective of the fit, such as the excess Born effective charge with the bonding type. It seems that such features have yet to be found for melting temperatures predictions.

Both Saad et al. and Seko et al. attempted to assess the importance of features used on their fits. The Seko et al. instances of AB solids spanned several quite different classes of solids. As noted and illustrated by Hullinger and Villiars,[46] an optimal uniform description of melting temperatures of AB classes has yet to be found. Two of Seko et al.'s features, the cohesive enegies and bulk moduli of the solids, are indicators of bond strength. A surprising result of the work of Saad et al. on the 44 suboctets was their conclusion that their 16 feature fit to the data was the least sensitive to the four features suggestive of bond strength

of the elements, such as their boiling points and heats of vaporization, as well as, their two atomic numbers. The atomic number relates to the atomic mass of the element. This mass influences the amplitudes of atomic vibration, the key ingredient in a Lindemann criterion. As Seko et al., we estimated confidence intervals on our results. Saad et al. did not. It would have been useful if they had so one could assess how well they controlled overfitting of the data. Our work has illustrated the importance of reporting such confidence intervals when fitting to small data sets. We note the Saad et al. did not create one statisical model for their 44 suboctets, but created a statistical model for each suboctet, one at a time, using an unstated-sized subset of the 44 that was judged similar to the octet under consideration. We remark that none of our features were indicators of bond strength. Still, our statisical model had a reasonably high predictive accuracy and stated confidence limits tempering the value of that accuracy. Clearly, there is more to understand about building statistical and physical models of melting temperatures of the octet solids.

Virtually all machine learning methods optimize something, generally a cost function. Accordingly, it is appropriate to mention the No Free Lunch Theorems for optimization.[9] These theorems state that a universal optimizer, that is, one that is optimal for all optimization problems, is not possible. With respect to machine learning, these theorems say that one can learn only what is in the data and that while for a given task, data, and feature set we can tune our algorithm so it performs better than the others we choose to consider, if we change the tasks, data, or features, there is no guarantee that the chosen algorithm still works the best. Even more specifically, an algorithm optimized for the training data might not be optimal for the test data. Because of these theorems, we should not be surprised to find variations in our results depending on the subset of data selected to validate our predictions via cross-validation. What was surprising however was the degree these results changed as this subset changed without any change in the algorithms being used. The variations we found, which we believe are due to the small sizes of the data sets, shift the use of the machine learning methods from being deterministic to being probabilistic. Instead of executing several such methods and comparing the results for consistency, cross-validation became part of the method as opposed to being an ancillary procedure. The procedure we reported to estimate the confidence interval for our results is a version of the bootstrapping method.[9,10]

We do believe the machine learning methods such as those used here will be useful for

pursuing similar tasks for materials that have larger data sets. The octets are only part of the over 545 known AB solids.[5,22] The full group has many more crystal structures, with some having a small number of solids per structure. This situation could make a complete multi-class classification task challenging. The classification task however would have an advantage because of the effectiveness of the St. John-Bloch pair and excess Born effective charge to build on. Solids of the type A_nB_m are also numerous, and in many cases are classified by using pairs of averages of the Mendeleev numbers. Would adding averages of the excess Born effective charge of the A atoms help? Many materials have chemistries of the type ABO_3, and whether they are perovskites or not is an important question. Perovskites exhibit a spectrum of functionalities, so the prediction of the Curie temperatures for ferromagnetism or ferroelectricity is important. These materials would not have the St. John-Bloch pair on which to build. However, for nearly a century, the tolerance factor has played an analogous role, and many perovskite materials have anomalously large effective Born charges.[42,45]

ACKNOWLEDGMENTS

We thank P. Balachandran and R. M. Martin for helpful discussions. This work was supported by the Laboratory Directed Research and Development (LDRD) program of the Los Alamos National Laboratory.

---

[1] G. Simons and A. N. Bloch, Phys. Rev. B **7**, 2754 (1973).

[2] J. St. John and A. N. Bloch, Phys. Rev. Lett. 33, 1095 (1974).

[3] J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B 17, 2453 (1978).

[4] A. Zunger, Phys. Rev. Lett. 44, 582 (1980).

[5] A. Zunger, Phys. Rev. B 22, 5839 (1980).

[6] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B 85, 104104 (2012).

[7] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big Data of Materials Science - Critical Role of the Descriptor. Submitted to Phys. Rev. Lett. (May 9, 2014).

[8] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Phys. Rev. B 89, 054303 (2014).

[9] P. Flach, *Machine Learning: The Art and Science of Algorithms the Make Sense of Data* (Cambridge, Cambridge, 2012).

[10] Z. Ivezić, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, Machine Learning in Astronomy* (Princeton, Princeton, 2014).

[11] R. Martin, Electronic Structure: Basic Theory and Practical Methods (Cambridge University Press, New York, 2004).

[12] J. C. Phillips and J. A. van Vechten, Phys. Rev. 22, 705 (1969).

[13] E. R. Dougherty, Comparative and Functional Genomics 2, 28 (2001).

[14] L. Ein-Dor, O. Zur, and E. Domany, PNAS 103, 5923 (2006).

[15] U. Braga-Neto, IEEE Signal Processing 24, 91 (2009) and references therein.

[16] E. R. Dougherty, C. Sima, J. Haa, B. Hanczar, U. M. Braga-Neto, Current Bioinformatics 5, 53 (2010).

[17] E. R. Dougherty, A. Zollanvari, and U. M. Braga-Neto, Current Genomics 12, 333 (2011).

[18] E. Mooser and W. B. Pearson, Acta Crystallogr. 12, 1015 (1959).

[19] J. A. van Vechten, Phys. Rev. 182, 891 (1969).

[20] J. C. Phillips, Rev. Mod. Phys. 42, 317 (1970).

[21] L. Pauling, *THe Nature of the Chemical Bond* (Cornell University Press, Ithaca, 1960).

[22] D. G. Pettifor, Solid State Commun. 51, 31 (1984).

[23] See Supplemental Material at [URL will be inserted by publisher] for technical details about computation of Born effective charges, data visualization and machine learning based binary classification. Computed nearest neighbor distances and excess born charges have also been provided.

[24] X. Gonze, C. Lee, Phys. Rev. B 55, 10355 (1997).

[25] S. Baroni, S. de Gironcoli, A. D. Corso, P. Giannozzi, Rev. Mod. Phys. 73, 515 (2001).

[26] R. D. King-Smith and D. Vanderbilt, Phys. Rev. B 47, 1651 (1993).

[27] D. Vanderbilt and R. D. King-Smith, Phys. Rev. B 48, 4442 (1993).

[28] R. Resta, Rev. Mod. Phys. 66, 899 (1994).

[29] D. M. Ceperley and B. J. Alder, Phys. Rev. Lett., 45, 566 (1980).

[30] G. Kresse and J. Furthmller, Phys. Rev. B 54, 11169 (1996).

[31] P. E. Blöchl, Phys. Rev. B 50, 17953 (1994).

[32] H. J. Monkhorst and J. D. Pack, Phys. Rev. B 13, 5188 (1976).

[33] W. A. Harrison, *Electronic structure and the properties of solids: the physics of the chemical bond* (Courier Dover Publications, 2012).

[34] G. Lucovsky, R. M. Martin, and E. Burstein, Phys. Rev. B 4, 1367 (1971).

[35] W. A. Harrison, Phys. Rev. B 8, 4487 (1973).

[36] R. M. Martin, Phys. Rev. B 5, 1607 (1972).

[37] J. Bennetto and D. Vanderbilt, Phys. Rev. B 53, 15417 (1996).

[38] M. di Ventra and P. Ferandez, Phys. Rev B 56, 12698 (1997).

[39] U. Iessi, C. Parisi, M. Bernasconi, and L. Migilo, Phys. Rev. 61, 4667 (2000).

[40] http://www.scikit-learn.org

[41] P. B. Littlewood and V. Heine, J. Phys. C: Solid St. Phys. 12, 4431 (1976).

[42] Ph. Gonze, J.-P. Michenaud, and X. Gonze, Phys. Rev. B **58**, 6224 (1998).

[43] S. Massidda, R. Resta, M. Posternak, and A. Baldereschi, Phys. Rev. B **52** R16977 (1995).

[44] M. Pasternak, A. Baldereschi, H. Krakauer, and R. Resta, Phys Rev. B **55**, 15983 (1997).

[45] M. Veithen, X. Gonze, and Ph. Ghosez, Phys. Rev. B **66**, 235113 (2002).

[46] F. Hulliger and P. Villars, J. Alloys Compounds **197**, 197 (1993).

[47] J. R. Chelikowsky and K. E. Anderson, Phys. Lett. **114A**, 482 (1986).

[48] J. R. Chelikowsky and K. E. Anderson, J. Phys. Chem Solids **48**, 197 (1987).

[49] J. M. Ziman, *Principles of the Theory of Solids* (Cambridge University Press, Cambridge, 1965).

[50] J. A. van Vechtan, Phys. Rev. Lett. **29**, 769 (1972).

[51] V. Kumar, V. Jha, and A. K. Shrivastava, Cryst. Res. Technol. **45**, 920 (2010).