Data-mined similarity function between material compositions

Lusann Yang and Gerbrand Ceder
Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139, USA

(Received 10 October 2013; revised manuscript received 14 November 2013; published 30 December 2013)

A new method for assessing the similarity of material compositions is described. A similarity measure is important for the classification and clustering of compositions. The similarity of the material compositions is calculated utilizing a data-mined ionic substitutional similarity based upon the probability with which two ions will substitute for each other within the same structure prototype. The method is validated via the prediction of crystal structure prototypes for oxides from the Inorganic Crystal Structure Database, selecting the correct prototype from a list of known prototypes within five guesses 75% of the time. It performs particularly well on the quaternary oxides, selecting the correct prototype from a list of known prototypes on the first guess 65% of the time.

DOI: 10.1103/PhysRevB.88.224107 PACS number(s): 61.50.Ah, 61.68.+n, 61.66.Fn

I. INTRODUCTION

Growing materials databases and the availability of more computational power have lead to considerable development in computational materials design.[1,2] Expanding databases of materials knowledge necessitates developing methods to organize such knowledge. For instance, the Inorganic Crystal Structure Database (ICSD) now contains 161 030 entries. Given a promising compound with certain properties, how can we systematically search for similar compounds? Such a definition of similarity must be with respect to a given property; in this paper, we develop a similarity function between two compositions that reflects the likelihood with which two compounds will have the same crystal structure.

Traditionally, materials scientists have relied upon structure mapping methods combined with heuristic design rules derived from the physical properties of individual ions to predict the structure of novel materials. For example, the Hume-Rothery rules relate the ratio of valence electrons per atom to the crystal structures formed,[3] while the Pauling rules relate the structure of ionic materials to the radii of the ions involved,[4] and Pettifor maps demonstrate clustering of similar crystal structures in a space where the coordinate of each element is simply related to its position in the periodic table.[5] Miedema *et al.* related the electron concentration at the boundary of a Wigner-Seitz cell to the formation enthalpy of binary metallic systems.[6] While these methods provide some physical insight, their predictive quality is limited,[7] and no obvious extension of these methods exists to make them more accurate or extend them to higher component systems.

Modern structure prediction methods may involve an unbiased search through the vast space of possible atomic arrangements;[8–12] others use chemical knowledge, including geometric data such as expected bond lengths,[13] secondary building units,[14] and structure prototype databases in conjunction with thermodynamic data[15] to reduce the search space. The problem with modern structure prediction methods is that they require a large number of energy evaluations, making them costly.

Motivated by Pettifor's structure maps, which displayed a correlation between ions with a similar Mendeleev number and the binary structure prototypes in which they formed, we generalize Pettifor's idea to incorporate information from not only binary, but ternary, quaternary, and more complex compounds. Following the ideas of Hautier *et al.*,[16,17] we use a data-mined quantitative likelihood with which two ions will substitute for each other within the same prototype to develop a composition similarity function. This composition similarity function has the advantage over traditional, heuristic methods that it is general: any two compositions, regardless of the number or identity of the components, can be compared to each other. Indeed, knowledge gleaned from the binary and ternary systems is used to inform our knowledge of the quaternary and more complex systems. This generality can be used to impart a distancelike structure to the database of knowledge, clustering together compounds of similar composition. The composition similarity function has the advantage of speed over that of modern structure prediction methods; informed by knowledge gleaned from current structure databases, composition similarity correctly classifies a new composition by selecting the correct prototype for an oxide structure from a list of known prototypes within five guesses 75% of the time.

The search through the space of possible atomic arrangements becomes much more difficult when considering complex materials, such as the quaternary oxides. This vast, sparsely sampled search space, with its enormous number of combinatoric possibilities, represents a rich area for the development of new materials. We present, in this paper, composition similarity, a data-mined function on the composition of a material that uses chemical knowledge from binary and ternary compounds and successfully applies it to quaternary structure prediction. Data mining across not only structural but also chemical similarity allows us to counteract the sparsely sampled nature of the space of quaternary materials; indeed, when limited to the quaternary subset of the oxides, composition similarity correctly selects the correct prototype from a list of known prototypes on the first guess 65% of the time.

II. METHODS

We begin with a few definitions. A composition is a set of ions and the associated ratios in which they appear. A crystal structure is given by a lattice and a basis of ions that decorates it; a crystal structure describes an infinite arrangement of ions in three-dimensional (3-D) space. By definition, the information given in a crystal structure includes the composition. A structure prototype is given by a lattice and an anonymized basis of ions that decorates it. A prototype contains information about the ratios in which differing ions appear, and their arrangement in 3-D space but not the identities of the ions involved. We use the word compound to refer to a specific entry in the ICSD, which includes information about the crystal structure of the compound, and, by definition, the composition of the compound. For example, the compound $LiFePO_4$ appears in our database with the composition Li^+, Fe^{2+}, P^{5+}, 4 O^{2-} in an olivine structure prototype.

All compounds containing over 20% oxygen by ion count in the ICSD 2012 were cleaned of peroxides, superoxides, and high-temperature and high-pressure phases and duplicates. Data cleaning details can be found in the Appendix. The final oxide data set consisted of 5695 compounds. After cleaning, the complete data set was randomly split into two sets for cross-validation. The first set, called the training set, consists of 95% of the compounds, representing the database of known compounds to be data mined. The remaining 5%, called the test set, mimics a set of as-yet-unseen compounds that we use to evaluate the efficacy of composition similarity. We performed the cross-validation process a total of five times.

In this section, we develop a similarity function between two compositions. Note that similarity refers to a property for which these compositions behave similarly. In this case, that property is crystal structure. We begin by data mining from the oxide training set an ionic substitution similarity function between two ions, per Hautier *et al.*[17] This function grows with the probability of the two ions substituting for each other within the same prototype within the database. Finally, we use the ionic substitution similarity as an input to the composition similarity function, defining a composition similarity function via the best matching of ions in composition c_1 to ions in composition c_2. The desired function should increase monotonically from 0 to 1, with the probability that the two compositions take the same prototype. It should achieve its maximum value of 1 when the two compositions are identical.

A. Ionic substitution similarity

In this section, a data-mined ionic substitution similarity function $\mathrm{Sim}_{\mathrm{ion}}(i_1,i_2)$ between any two ions i_1 and i_2 that satisfies the following constraints is developed:

(1) $0 \leqslant \mathrm{Sim}_{\mathrm{ion}}(i_1,i_2) \leqslant 1$ for all ions i_1 and i_2

(2) $\mathrm{Sim}_{\mathrm{ion}}(i_1,i_2)$ grows monotonically with the probability that i_1 and i_2 substitute for each other within a given prototype.

We use a typical approach used in data mining of formulating a model and then determining the parameters of the model by requiring that the known data is reproduced with maximum probability. Following the work of Hautier *et al.*,[17] the probability with which two compounds X and X' take the same prototype is modeled as a function of the ion-ion substitutions $i = (i_1,i_2)$ required to map the crystal structure of X onto that of X':

$$p(\mathrm{prototype}(X) = \mathrm{prototype}(X')) = \frac{e^{\sum_i \lambda_i f_i(X,X')}}{Z}, \qquad (1)$$

where f_i(X, X′), is a series of binary indicator functions:

$$f_i(X,X') = \begin{cases} 1, & \text{if ion } i_1 \text{substitutes for ion } i_2 \\ 0, & \text{else} \end{cases}. \qquad (2)$$

The λ_i indicate the weight that is assigned to each feature function f_i(X, X′) and encapsulate the heart of the model; ion-ion substitutions i that occur often are indicated by larger values of λ_i, whereas substitutions that are rare are indicated by smaller values of λ_i. Lastly, Z is the partition function necessary to ensure that all the probabilities sum to 1.

This binary substitution model allows us to learn from our database D the values of the weights λ_i for each substitution i. Representing the prototyped database D as a set of pairs of compounds that share the same prototype, the probability of the data in the training set is

$$P(\{(X,X') \in D\}) = \prod_{(X,X')\in D} \frac{e^{\sum_i \lambda_i f_i(X,X')}}{Z}. \qquad (3)$$

Solving for the λ_i that maximize this probability is known as calculating the maximum-likelihood model.[18] For those λ_i which correspond to unobserved substitutions, λ_i is set to -10.

Finally, letting X_i indicate the ith site of compound X, we extract the probability p(b|a) of ion b substituting for ion a in the same prototype, given that ion a is already known to exist in that prototype:

$$p(b|a) = \frac{p(X_1 = a, X'_1 = b|\mathrm{prototype}(X) = \mathrm{prototype}(X'))}{p(X_1 = a)} \qquad (4)$$

$$= \frac{e^{\lambda_{a,b}}}{1 + e^{\lambda_{a,b}}} \frac{1}{\sum_j \frac{e^{\lambda_{a,j}}}{1+e^{\lambda_{a,j}}}}, \qquad (5)$$

which leads naturally to the following proposed definition of a symmetric similarity of ion a to ion b:

$$\mathrm{Sim}_{\mathrm{ion}}(a,b) = \max(p(b|a), \quad p(a|b)) \qquad (6)$$

$$= \frac{p(X_1 = a, X'_1 = b|\mathrm{prototype}(X) = \mathrm{prototype}(X'))}{\min(p(X_1 = a), p(X_1 = b))}. \qquad (7)$$

The ionic substitution similarity function is defined to be the maximum of the two conditional probabilities to allow for sampling deficiencies in the data set. For example, consider the scenario in which ion a is rare and appears in only 1% of the data set, ion b is common and appears in 50% of the data set, and for every compound in which ion a appears, another compound appears in the same prototype and composition, only with ion b substituted for ion a. Choosing the minimum of the two conditional probabilities penalizes the probability of substitution because ion a is rare.

Using the similarity on the training set described above yields the similarity shown in the image below. Specifically, it shows the similarity of ion a to ion b for the 60 most common ions in our data set.

Figure 1 shows an ionic substitutional similarity function that is heavily weighted towards zero, with the vast majority

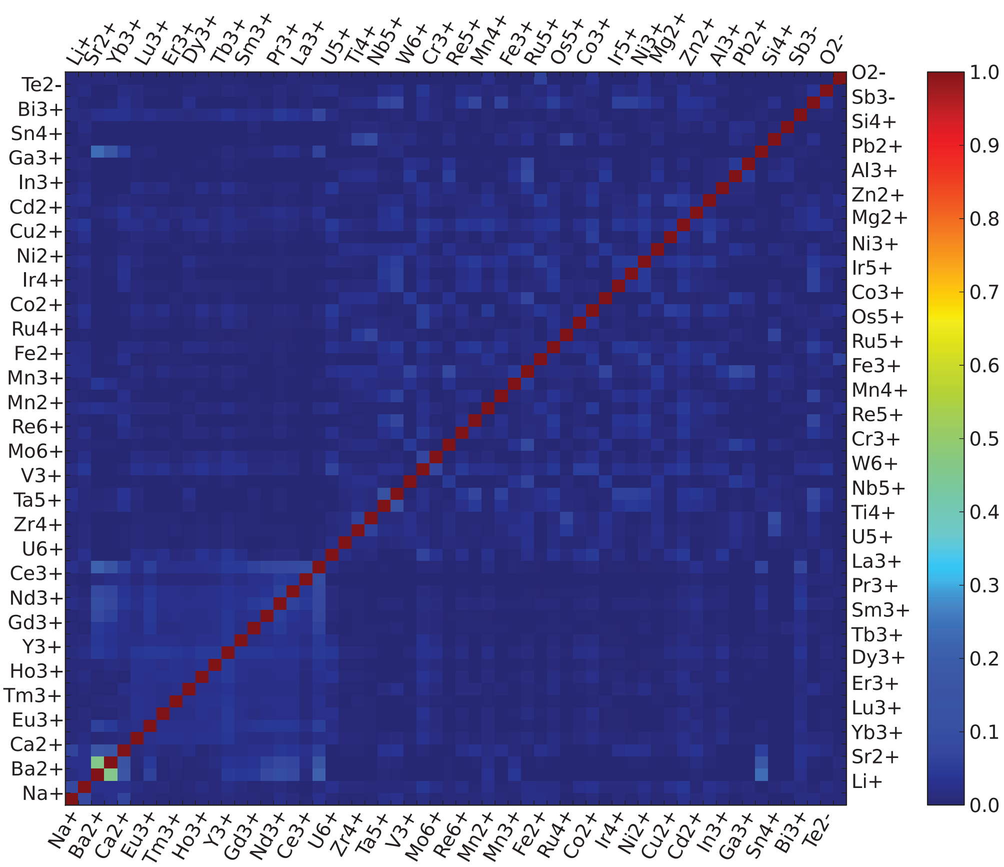

FIG. 1. (Color online) Unscaled ionic substitutional similarity. The probability of ion a substituting for ion b within the same structure prototype as given by the maximum likelihood model applied to the oxides in the ICSD.

of our values occurring below 0.30. Taking a logarithm and rescaling our similarity linearly such that all values lie within the interval $[1,0]$, we produce the following similarity between ions.

This rescaled similarity, shown in Fig. 2, differentiates ion substitution probabilities strongly, resulting in better structure prediction, while still growing monotonically with increasing probability of ionic substitution. As in previous work,[17] we find two areas with high substitution rates consistent with intuition. The rare earths are clearly distinguished in the bright yellow lower left-hand corner, and the transition metals, roughly in the center of the diagram, also substitute with each other with high probability.

B. Composition similarity

Given the ionic substitutional similarity above, it is now possible to define a quantitative data-mined similarity rating between two compositions. Calculating composition similarity involves finding the best possible matching of the ions in one composition to the ions in the other such that the average ionic substitutional similarity between each pair is maximized.

A composition is defined as a set of ions $\{i\}$ together with the number of times each ion appears $\{n\}$. We represent compositions by their reduced versions, where the reduced version has the smallest integers $\{n\}$ that preserve the correct ratios between the ions. The sum Σn of the number of ions in the reduced composition is the total number of ions n_{total} of a given composition.

Given two compositions c_1 and c_2, we find the lowest common multiple n_{lcm} of n^1_{total} and $n^2{}_{\text{total}}$. We cap n_{lcm} at a maximum value of 100 to limit computational complexity. Two sets of ions s_1 and s_2 of length n_{lcm} are created by enumerating the ions of c_1 and c_2 the appropriate number of times. Searching through all the possible matchings (e_1,e_2) of the ions e_1 in s_1 to the ions e_2 in s_2, the matching that maximizes the average similarity of the two sets is found. This maximal average similarity is defined as the composition similarity between c_1 and c_2.

$$\text{Sim}_{\text{comp}}(c_1,c_2) = \max_{\text{all matching}} \frac{\sum_{(e_1,e_2 \in \text{matching})} \text{Sim}_{\text{ion}}(e_1,e_2)}{n_{\text{lcm}}} \tag{8}$$

The composition similarity yields a rating between 0 and 1 for every pair of compositions c_1 and c_2, with identical compositions having similarity 1. Based on data-mined values for the probability with which each ion will substitute for another within the same prototype, this composition similarity provides a quantitative method to evaluate the likelihood with which two compounds will form in the same prototype.

III. RESULTS

The ionic substitution similarity function was evaluated based upon the compounds in the training set. Using the ionic substitution similarity, composition similarities were

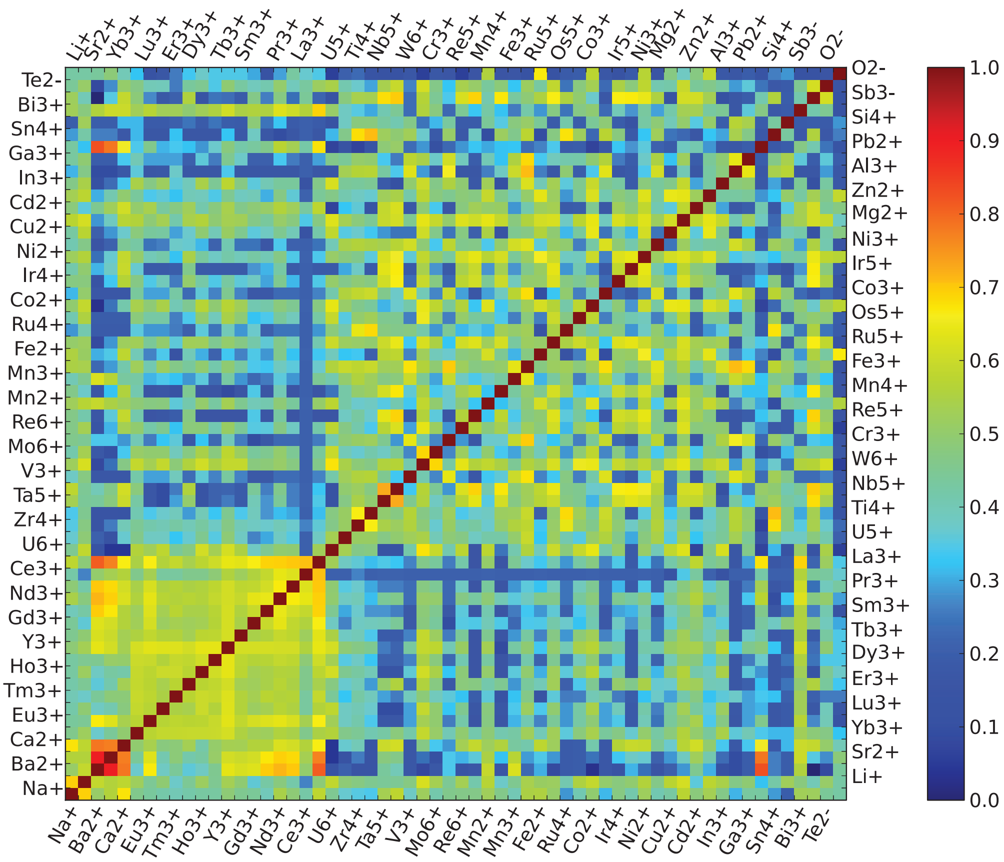

FIG. 2. (Color online) Rescaled ionic substitutional similarity function. The probability of ion *a* substituting for ion *b* within the same prototype, rescaled to better differentiate substitution probabilities.

calculated between every composition in the test set and every compound in the training set. Finally, for each composition in the test set, the compounds in the training set were ordered by composition similarity. This cross-validation process was completed with five different partitions of training and test sets; here, we present the cumulative results from all five partitions.

A. A few examples

We begin by examining the behavior of clustering by composition similarity upon three sample compounds from the test set.

1. Example 1

Table I and Fig. 3 summarize the results of sorting by composition similarity to Ca_2FeWO_6, which appears in the test set in the double-perovskite crystal prototype. Ranking the compounds in the training set by composition similarity to Ca_2FeWO_6, we find that composition similarity begins by finding compounds that differ by one ionic substitution per formula unit. The first most similar compound substitutes Mo for W, yielding Ca_2FeMoO_6, which forms in a distorted double-perovskite prototype. The next two most similar compounds are two polymorphs, Ca_2NiWO_6, forming in the distorted double-perovskite prototype and an experimentally reported prototype featuring square planar-coordinated Ni^{2+} ions. The next two most similar compounds, Ca_2MnWO_6, Ca_2MgWO_6, both form in the distorted double-perovskite prototype. Finally, the next guess, Ba_2FeWO_6, representing a substitution of two Ba^{2+} ions for Ca^{2+} ions per formula unit, appears in two polymorphs; the distorted and perfect double-perovskite prototypes. In this example, the composition similarity method finds the correct structure prototype for the compound in question on the third guess, though all guesses are structurally similar to Ca_2FeWO_6.

2. Example 2

Table II below summarizes the results of sorting by composition similarity to $BaLa_2Ti_3O_{10}$, which appears in the test set. The most similar compound in the training set is $La_4Ti_3O_{12}$, followed by $La_2Ti_2O_7$. With the third guess,

TABLE I. Compounds with high composition similarity to Ca_2FeWO_6. The first five compounds differ by one ion substitution per formula unit; the second two differ by two substitutions per formula unit.

Composition	Prototype	Similarity to Ca_2FeWO_6
Ca_2FeMoO_6	1	0.968
Ca_2NiWO_6	1	0.965
Ca_2NiWO_6	2	0.965
Ca_2MnWO_6	1	0.961
Ca_2MgWO_6	1	0.961
Ba_2FeWO_6	1	0.956
Ba_2FeWO_6	3	0.956

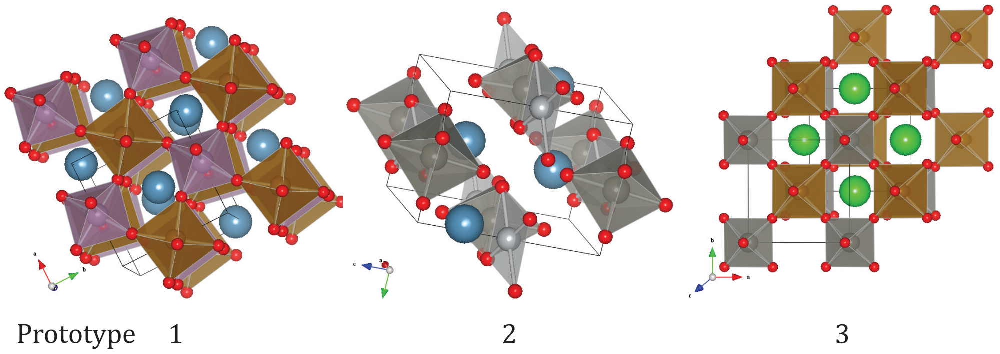

FIG. 3. (Color online) Three crystal structure prototypes suggested by the composition similarity method for the crystal structure of Ca_2FeWO_6. The first and third prototypes are both double perovskites, with the octahedron in the first prototype being slightly distorted. The second prototype represents an experimentally determined polymorph of Ca_2NiWO_6.

composition similarity returns to the original stoichiometry and finds $BaPr_2Ti_3O_{10}$, which forms in the same prototype as $BaLa_2Ti_3O_{10}$.

3. *Example 3*

Table III and Fig. 4 below summarize the results of sorting by composition similarity to a polymorph of $DyMnO_3$ that appears in the test set. The $LnMnO_3$ structures, where Ln is a lanthanide, form primarily in two structural prototypes: a distorted orthorhombic perovskite and a hexagonal structure.[19] The target compound in our training set forms in the distorted orthorhombic perovskite structure, shown in Fig. 4.

Examining the results of ordering the test set by composition similarity shown in Table III, we find that the three most similar compounds in the test set are polymorphs of $DyMnO_3$: two distorted versions of the distorted orthorhombic perovskite structure and the hexagonal structure. Finally, composition similarity substitutes Y for Dy, guessing two polymorphs of $YMnO_3$; the first polymorph is the tetrahedral arrangement, and the second is the correct distorted orthorhombic perovskite.

Again, composition similarity selects crystal structures that have similar structural components. Three out of the first four structures are remarkably similar distorted orthorhombic perovskite structures; the last is a well-known polymorph of $DyMnO_3$. The final structure only differs from previous ones by a slight shift in the placement of the central Dy^{3+} ion.

IV. APPLICATION TO STRUCTURE PREDICTION

To quantitatively assess the ability of composition similarity to cluster compounds with similar structure, we consider the application of composition to structure prototyping. For each composition in our test set, we ask, in what structure prototype would this compound form? We answer this question by referring to the list of compounds in our training set, ordered by similarity to the test set composition. Structure prototypes are guessed from that list, progressing from the most similar compounds downward, subject to the following rules:

(1) If a compound in the test set forms in a structure prototype that is unrepresented in the training set, we do not consider this compound, as it is impossible to guess this structure prototype. Composition similarity does not have the ability to suggest as-yet-unseen crystal structures.

(2) If the test set composition is a binary, ternary, or quaternary or more complex composition, only appropriate binary, ternary, or quaternary or more complex prototype guesses are permitted, matching the number of chemical components in the compound.

(3) No prototype is guessed twice.

(4) No training set compounds with the same composition as the test set composition are considered. Recall that the entire dataset, consisting of both test set and training set, is constructed such that no two entries share the same composition and prototype; such entries would be considered duplicates. Thus, guessing compounds with the same composition would be by definition suggesting polymorphs with incorrect prototypes.

In the following analysis, our structure prototype prediction algorithm, the composition similarity algorithm, is compared against a control in which the list of suggested prototypes is ordered by the frequency with which these prototypes appear in the training set, called the most common prototypes algorithm.

TABLE II. Compounds with high composition similarity to $BaLa_2Ti_3O_{10}$. The first two compounds have differing stoichiometry than the target compound.

Composition	Similarity to $BaLa_2Ti_3O_{10}$
$La_4Ti_3O_{12}$	0.973
$La_2Ti_2O_7$	0.965
$BaPr_2Ti_3O_{10}$	0.962

TABLE III. Compounds with high composition similarity to $DyMnO_3$. The first three are polymorphs of DyMnO3; they have the same composition but differing structure. All five compounds have similar crystal structures.

Composition	Prototype	Similarity to $DyMnO_3$
$DyMnO_3$	1	1.000
$DyMnO_3$	2	1.000
$DyMnO_3$	3	1.000
$YMnO_3$	3	0.927
$YMnO_3$	4	0.927

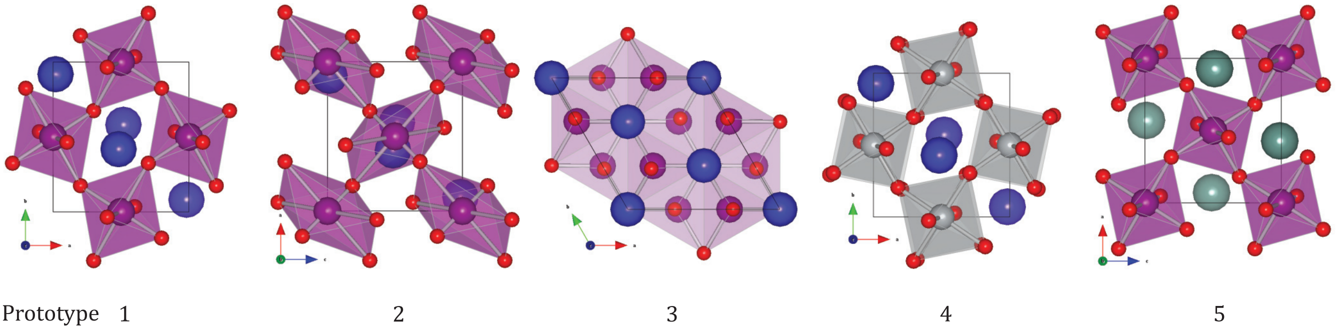

FIG. 4. (Color online) Four structure prototypes suggested by the composition similarity method for the crystal structure of $DyMnO_3$. Numbers 1, 2, and 4 are all distorted orthorhombic perovskite structures. Structure prototype number 3 is a hexagonal crystal structure commonly found in the $LnMnO_3$ family, where Ln is a lanthanide.

Prototypes are guessed, subject to the same rules, from the most frequently observed prototypes to the least. This is a similar benchmark to what was used previously by Fischer *et al.*[15]

Figure 5 depicts the performance of crystal structure prototyping via composition similarity aggregated across all five cross-validated test sets. The horizontal axis depicts the probability with which the correct prototype is among our guesses against the vertical axis, which depicts the number of guesses made. The black line shows the performance of prototyping via composition similarity. The dotted line shows the performance of prototyping via the most common prototypes method. Prototyping via composition similarity consistently outperforms prototyping via the most common prototypes method, requiring a smaller number of guesses at every confidence level. Overall, prototyping via composition similarity achieves 75% accuracy within five guesses. The correct prototype is guessed within the first three guesses 67% of the time.

The strengths and weaknesses of composition similarity become evident when the data is broken down by the number of components in the compound. Dividing the data set into groups of binary, ternary, and quaternary or higher compositions, a clear trend in favor of the prediction of more complex compound prototypes emerges. Figure 6 below shows the number of guesses necessary to find the correct prototype, broken down by number of components in the compound, for prototypes ordered via the composition similarity and most common prototypes methods. The composition similarity rating fares poorly in the binary compounds, performing comparably to the most common prototypes method. Its performance improves markedly in the ternaries, while the most common prototypes method suffers from the large number (~1100) of candidate ternary prototypes. The trend continues into the quaternaries, where the most common prototypes ranking fares remarkably well, predicting the correct structure prototype out of over 1600 candidate structure prototypes on the first guess 65% of

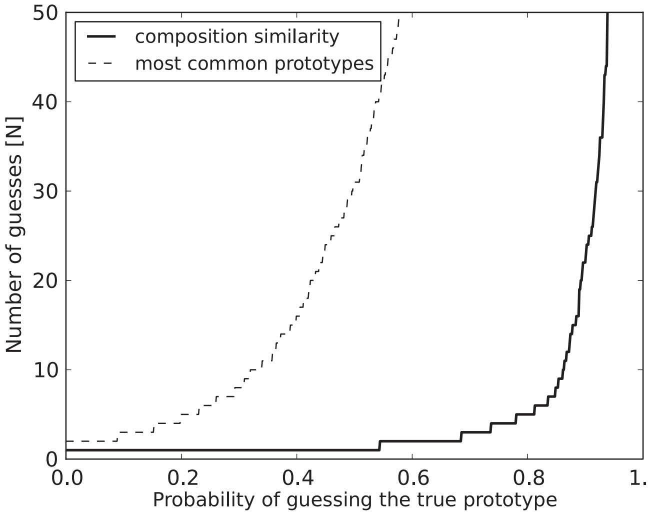

FIG. 5. Prototyping by composition similarity as applied to the oxides in the ICSD. The black line shows the number of guesses necessary to select the correct prototype by composition similarity. The dotted line shows the number of guesses necessary if those guesses were ordered by the frequency with which that prototype appears.

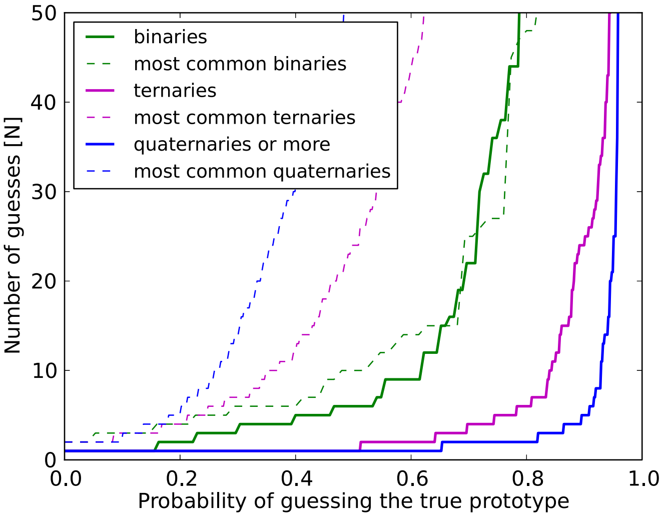

FIG. 6. (Color online) Prototyping by composition similarity broken down by the number of components in a compound. The bold lines show the number of guesses necessary to select the correct prototype by composition similarity. The dotted lines shows the number of guesses necessary if those guesses were ordered by the frequency with which that prototype appears. The performance of composition similarity increases dramatically with the number of components in a compound.

the time, within two guesses 80% of the time, and within ten guesses 90% of the time.

V. DISCUSSION

We have presented a data-mined, quantitative composition similarity function that reflects the probability of two compositions taking the same crystal structure prototype. This composition similarity function is obtained in two steps; the first is to data mine an ionic substitutional similarity function that reflects the probability that two ions will substitute for each other within the same prototype. The second is to use this ionic substitutional similarity to find the most similar matching of the ions in two given compositions; the average similarity of this matching is the composition similarity.

We have used structure prototype prediction as a means of evaluating the efficiency with which composition similarity groups similar compounds. Using composition similarity, we ordered the oxides in a training set of over 4000 compounds versus each compound in the test set. The compounds that appear first on the ordered list represent the most similar compounds to the test set compound in question. We have shown that these most similar compounds are very likely to share the same prototype as the test set compound. Additionally, we have found that composition similarity orders the possible prototype structures more effectively as the number of components of the compound increases, finding the correct prototype in remarkably few guesses.

The correlation between performance of composition similarity and the number of components of the compound in question can be attributed to the relative lack of prototypes in the quaternary structures. While the binary compounds contain only one nonoxygen ion and form in over 300 distinct prototypes, the ternaries contain two nonoxygen ions and form in 1100 prototypes, and the quaternaries contain three nonoxygen ions and form in 1600 prototypes. Considering the number of possible ionic combinations, with over 200 species of ions represented in our database, the number of combinations grow 200-fold going from the binaries to the ternaries and 4000-fold from the binaries to the quaternaries. However, the number of prototypes the algorithm must order in this data set grows far more slowly to the benefit of the predictive ability of our algorithm. While there undoubtedly exist ternary and quaternary prototypes that are as yet unrepresented in the ICSD—50% of the quaternary prototypes in the test set were not represented in the training set—this study shows that composition similarity takes advantage of the higher ratio of ionic combinations to number of structural prototypes available in a quaternary composition to better order the current list of available prototypes. For those quaternary prototypes that were represented in the training set, we find the correct prototype on the first guess 65% of the time.

We refer to example 2, in which we rank training set compounds by their composition similarity to $BaLa_2Ti_3O_{10}$ to highlight one of the weaknesses of composition similarity. In this example, it takes composition similarity three guesses to obtain the correct structure prototype. The first two guesses, $La_4Ti_3O_{12}$ and $La_2Ti_2O_7$, exhibit starkly differing stoichiometries from the desired compound, which would make them unlikely candidates for sharing the same structure prototype. However, because composition similarity is given by the highest average chemical similarity between pairs of ions from both compounds, high chemical similarity between a majority of the ions can outweigh a few improbable substitutions. Looking carefully at this example, it would seem improbable for the ions from $BaLa_2Ti_3O_{10}$ to map stoichiometrically onto the ions of $La_4Ti_3O_{12}$; however, taking the lowest common multiple, the 16 ions per formula unit in $BaLa_2Ti_3O_{10}$ and the 19 ions per formula unit in $La_4Ti_3O_{12}$, we find our similarity function is forced to compare 247 ions of each composition. In this case, one of the O^{2-} ions from $BaLa_2Ti_3O_{10}$ must eventually be mapped onto either La^{3+} or Ti^{4+}, a rather improbable substitution. However, with over 247 comparisons in total, there are enough overwhelmingly good substitutions to outweigh a few improbable ones.

In future work, it would be possible to address this weakness through a simple algorithmic variation. The substitution of ions in differing charge states could be strictly disallowed by automatically setting their similarities to $-\infty$. Such a modification, while computationally straightforward and physically meaningful, would also result in a loss of information. The current algorithm data mines chemical similarity, while allowing for multiple substitutions within the same prototype; it is not uncommon for ions of differing charge states to substitute for each other when accompanied by another, simultaneous substitution, which offsets the charge imbalance. The information gleaned from charge-imbalanced, multiple-ion substitutions would be lost if we implemented this variation.

Compared to other structure prediction algorithms, ordering via composition similarity has distinct strengths and weaknesses. Unlike direct optimization search methods, composition similarity does not have the ability to predict new structure prototypes that are not represented in current databases. Some direct optimization search methods, for example, simulated annealing, can attempt to systematically search through the infinite-size space of possible structures given infinite time, while composition similarity is strictly limited to searching the data set at hand. However, composition similarity effectively orders structure prototype candidates prior to any energetic evaluation and is thus quite computationally cost-effective. Furthermore, the growing efficacy of composition similarity with the complexity of the compound makes structure prediction via composition similarity much more attractive in the quaternary or even quintenary compounds. We expect the performance of composition similarity to improve as more quaternary compounds are discovered.

Composition similarity has broader potential for application than the example of prototype prediction discussed above. The composition similarity function takes as its input any two compositions and outputs a number that reflects the chemical and structural similarity between them. Such a function is useful with respect to the classification: we demonstrated the performance of composition similarity when classifying compounds by structure prototype. However, in a broader sense, composition similarity is useful because it is an effective clustering method, grouping together compounds that are similar and providing a mechanism for the mapping of composition space. Defining the distance between two

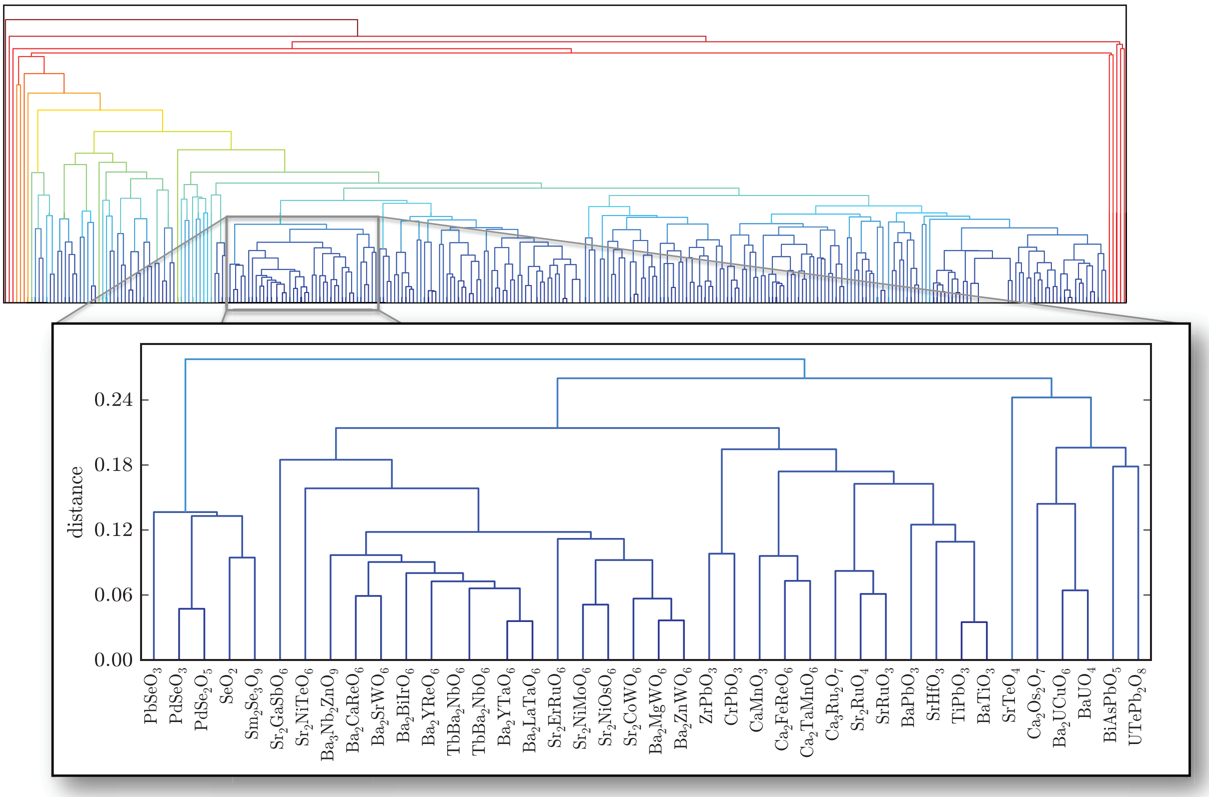

FIG. 7. (Color online) Clustering of a sample set of 300 oxides from the ICSD by composition similarity.

compositions as one, the composition similarity, we now have a semimetric that imposes a structure upon the space of compositions. Note that this distance does not satisfy the triangle inequality and thus is only a semimetric.

Figure 7 below shows a small sample set of 300 compounds drawn randomly from the test set, clustered by composition similarity. The vertical axis represents the distance between two compounds or the largest possible distance between two clusters. Identical compounds (SiO_2) have distance zero. The horizontal axis represents the clustering of similar compounds; generally speaking, similar compounds will be drawn closer to each other on the horizontal axis.

Figure 7 represents the ability of composition similarity to provide structure to a well-explored but as yet relatively unmapped space; the space of all compounds. It can form a hierarchical grouping of similar compounds across all chemistries, comparing binary compounds to ternaries and quaternaries in a quantitative, physically meaningful way. We suggest that the organizational value of composition similarity may prove useful, allowing the designers of new compounds a new mechanism by which to search for compositionally similar compounds.

Finally, it is possible to extend the composition similarity method such that it is no longer based upon structural similarity. This paper describes a two-step algorithm; the first part, following the work of Hautier *et al.*,[17] data mines an ionic substitutional similarity between two ions that reflects the tendency of those ions to form within the same structure prototypes. The second part describes how to use that ionic similarity function to compute a composition similarity function. Appropriately, we use the ionic similarity that reflects the tendency to form within the same prototype to predict crystal structure. However, the two parts are modular; the derivation of composition similarity is independent of which ionic similarity function is used. If the end goal were not the prediction of crystal structure but another property, using another ionic similarity function may prove more direct and yield a better prediction.

VI. CONCLUSIONS

A data-mined composition similarity function that combines both chemical and structural knowledge is presented. Compounds with high composition similarity have similar structures. In particular, composition similarity is particularly efficacious at predicting structure prototypes of quaternary oxides, an area in which the available data is notoriously sparse.

ACKNOWLEDGMENTS

We acknowledge technical support, discussion, and advice from G. Hautier, S. P. Ong, A. Qi, and Y. Zhu. This work was supported in part by the US Department of Energy under Contract No. DE-FG02-96ER45571 and by the Navy-Office of Naval Research under Contract No. N00014-11-1-0212. The formulation of ideas leading up to this work was supported by the National Science Foundation/Cyber-enabled Discovery and Innovation under Contract No. ECCS-0941043.

APPENDIX: DATA CLEANING

All of the compounds in the ICSD 2012 were searched for compounds that satisfied the following criteria:

(a) Compounds must be oxides, as indicated by at least 20% oxygen content by ion count.

(b) Compounds must not be peroxides or superoxides, as indicated by O-O bond lengths L<1.50 Å.

(c) Compounds must not be marked high pressure, HP, high temperature, or HT.

(d) Compounds must not have improbably short (<1 Å) bond lengths.

(e) Compounds must not have a mismatch between the reported composition and the ions given in the crystal structure.

(f) Compounds must not contain hydrogen. The reported crystal structures of compounds containing hydrogen are often unreliable.

The resultant oxides were sorted into structure prototypes using an affine mapping algorithm.[20,21] The data set was further cleaned by removing duplicates, defined as compounds with the same composition and the same structure prototype, resulting in a final data set of 5694 oxide compounds.

[1] Y. S. Meng and M. E. Arroyo-de Dompablo, Energy Environ. Sci. **2**, 589 (2009).

[2] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Comput. Mater. Sci. **50**, 2295 (2011).

[3] W. Hume-Rothery, R. E. Smallman, and C. W. Haworth, *The Structure of Metals and Alloys* (Metals and Metallurgy Trust, London, 1969).

[4] L. Pauling, J. Am. Chem. Soc. **51**, 1010 (1929).

[5] D. G. Pettifor, J. Phys. C **19**, 285 (1986).

[6] A. R. Miedema, P. F. de Châtel, and F. R. de Boer, Physica B + C **100**, 1 (1980).

[7] D. Morgan, J. Rodgers, and G. Ceder, J. Phys.: Condens. Matter **15**, 4361 (2003).

[8] C. W. Glass, A. R. Oganov, and N. Hansen, Comput. Phys. Commun. **175**, 713 (2006).

[9] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[10] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Phys. Rev. B **82**, 094116 (2010).

[11] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Comput. Phys. Commun. **183**, 2063 (2012).

[12] M. Amsler and S. Goedecker, J. Chem. Phys. **133**, 224104 (2010).

[13] A. Le Bail, J. Appl. Crystallogr. **38**, 389 (2005).

[14] C. Mellot Draznieks, J. M. Newsam, A. M. Gorman, C. M. Freeman, and G. Férey, Angew. Chem. Int. Ed. **39**, 2270 (2000).

[15] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).

[16] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).

[17] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).

[18] J. Aldrich, Stat. Sci. **12**, 162 (1997).

[19] S. M. Woodley, P. D. Battle, J. D. Gale, and C. R. A. Catlow, Chem. Mater. **15**, 1669 (2003).

[20] R. Hundt, J. C. Schön, and M. Jansen, J. Appl. Crystallogr. **39**, 6 (2006).

[21] H. Burzlaff and Y. Malinovsky, Acta Crystallogr. Sec. A **53**, 217 (1997).