Proposed definition of crystal substructure and substructural similarity

Lusann Yang, Stephen Dacek, and Gerbrand Ceder

Phys. Rev. B **90**, 054102 — Published 5 August 2014

DOI: 10.1103/PhysRevB.90.054102

A Proposed Definition of Crystal Substructure and Substructural Similarity

Lusann Yang, Stephen Dacek, and Gerbrand Ceder*
Department of Materials Science and Engineering
Massachusetts Institute of Technology

There is a clear need for a practical and mathematically rigorous description of local structure in inorganic compounds so that structures and chemistries can be easily compared across large data sets. A new method for decomposing crystal structures into substructures is given, and a similarity function between those substructures is defined. The similarity function is based upon both geometric and chemical similarity. This construction allows for large-scale data mining of substructural properties, and the analysis of substructures and void spaces within crystal structures. The method is validated via the prediction of Li ion intercalation sites for the oxides. Tested on databases of known Li-ion containing oxides, the method reproduces all Li-ion sites in an oxide with a maximum of 4 incorrect guesses 80% of the time.

PACS numbers: 61.50.Ah, 61.66.Fn, 61.68.+n

I. INTRODUCTION

Growing materials databases, computational power, and better computational techniques have made it an exciting time in computational materials science.[1,2] The Materials Project has published *ab initio* computations of nearly 50,000 compounds online, including over 20,000 band structures.[3] The Inorganic Crystal Structure Database, a database of experimentally determined compound crystal structures, now contains 161,030 entries.[4] Growing databases of materials incur the necessity to develop methods with which to organize such knowledge, and allow for the possibility of systematically mining this data for patterns.

When mining data for patterns, it is often useful to develop similarity or difference functions that quantify the relationships between structures. Such similarity functions provide the ability to cluster similar structures together and imposes a natural ordering on the space of crystal structures. A similarity function between two crystal structures typically consists of two components; a component that measures similarity in chemistry, and a component that measures geometric similarity. Unlike traditional symmetry or unit-cell based methods of structure description, methods of geometric comparison should be based upon continuous functions of ion position, which allows for structures with similar ionic positions to be grouped together. Additionally, these functions of ion position should be invariant under rotation, translation, and choice of unit cell.

There have been several definitions of continuous, quantitative similarity functions on the space of crystal structures. The works of Willighagen and Oganov combine radial distribution functions with ion-specific information such as charge state or neutron scattering length to describe crystal structure.[5,6] De Gelder derives his similarity function using weighted cross correlations of the powder diffraction pattern,[7] while Rupp et al have used a matrix representation of structure based upon the Coulomb interaction between ions.[8] All of the above methods combine two components: Each method features a description of the geometrc arrangement of ions (either embedded in the radial distribution pattern, the powder diffraction pattern, or the strength of the Coulomb interaction), and a chemical descriptor for each ion (embedded in the scattering pattern or the strength of the charges in the Coulomb interaction).

In this paper, we will develop a continuous, quantitative similarity function between substructures that allows chemically and topologically similar structures to be grouped together despite the fact that they may have different unit cells, composition, and symmetry. Substructure based analyses of materials properties have a rich history in materials science, as breaking crystal structures into substructures allows for the study of sites, defects, void spaces, and the packing of substructures. Linus Pauling's Principles determining the structure of complex ionic crystals[9], colloquially known as the Pauling rules, describe a set of rules for deriving ionic crystal substructures based upon geometry and local packing rules. Daams and Villars presented in 2000[10] an enlightening study in which they decomposed 200,000 inorganic crystal structures in 5,000 structural prototypes into chemistry-independent atomic environment topologies. Interestingly, they showed that only 20 of the most frequent atomic environment types were necessary to account for 80% of the prototypes seen; only another 70 rarer atomic environment types were necessary to account for their entire data set. Methods that quantify the similarity of crystal structures can be used to predict the existance and packing of substructures, which would prove useful in the prediction of novel crystal structures. Mellot-Draznieks et al. have published a number of exciting studies predicting the structures of inorganic materials by assembling secondary building units via simulated annealing methods.[11–13]

Given the high degree of structure that Daams and Villars have found in the atomic environments of inorganic compounds, alongside Pauling's intuitive and compelling physical arguments for the existence of a set of highly ordered substructures, decomposing crystal structures into substructural units is a natural next step. In previous work, we developed a similarity between ionic compositions and used it to validate the correlation between similar compositions and crystal structure prototypes.[14] The ionic similarity function measures to what extent pairs of ions behave similarly in terms of crystal structure formation. Hence, it can also be thought of as the probability with which one ion will substitute for another while retaining the same structure prototype. Breaking crystal structures down into substructures allows us to further subdivide the data set of known compounds and their structures, yielding more points in a richer database. A database of 5,500 oxides contains millions of substructures. In this paper we use the previously developed ionic substitutional similarity as our method of chemical comparison, and a weighted Voronoi construction by O'Keefe[15] as our method geometric comparison, to construct a similarity function between substructures.

The proposed similarity function provides a robust method to describe and quantify the differences between substructures. It allows for the organization of a database of substructures, and for the analysis of both sites and void spaces within crystal structure. We validate our work with the application to an important problem in the design of better lithium ion battery materials. As electrode materials function by reversibly inserting and extracting Li ions into their crystal structure, it is important when designing new electrode compounds to have algorithms that can identify potential Li sites. Using cross validation, we will evaluate how well the proposed substructure similarity function can be trained on known Li sites and then used to predict Li sites.

II. METHODS

In this section we will present a description of substructure and a similarity function between substructures that respects both geometric and chemical similarity. Materials databases such as the Inorganic Crystal Structure Database[4] contain hundreds of thousands of crystal structures which can be decomposed into millions of substructures. These substructures will vary in both composition and geometry. For example, the perovskite crystal structures shown in Figure 1 feature several common distortions of the coordination polyhedra around the central ion. These distorted polyhedra should be considered *geometrically* similar. Furthermore, it is not uncommon for differing crystal structures to share the same prototype, save for the substitution of one ion for another. In the cases in which similar ions inhabit similar substructures, these substructures must be considered chemically similar.

A. Defining Substructure

With the requirements for substructural similarity in mind, we present a definition of substructure that contains information regarding both chemistry and geometry. This definition of substructure should be well defined and continuous versus small perturbations in atomic position. In the field of machine learning, the term feature vector selection is used to describe the problem of finding and choosing the variables that influence the outcome of the problem at hand. The choice of feature vector should not only include all of the factors that may influence the outcome of the problem, but it should also be dense: To the greatest extent possible, it should not include information that does not influence the outcome of the problem.

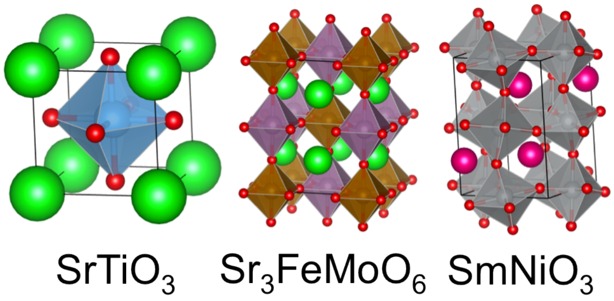

FIG. 1. Variations on the perovskite crystal structure. Three perovskite crystal structures are shown. The first structure $SrTiO_3$, is the perovskite crystal structure. The next two structures are variations on perovskite. The second features a slight orthorhombic distortion and alternating Fe and Mo ions, while the third features distorted octahedra. While the octahedral environments vary in both chemistry and geometry in all three structures, they are nonetheless similar.

With respect to the problem at hand, we believe that a good feature vector for substructure prediction and data mining should fulfill the following criterion.

1. The feature vector should include *geometric* information about a substructure. At a minimum, the feature vector should be able to distinguish between several commonly found coordinations such as tetrahedral, octahedral, cubic, etc. More geometric information, for example distortions in octahedral site, or the chemical identities of next-nearest neighbors, could also prove useful. As we are expecting millions of substructures, it is also useful to minimize computational complexity by limiting feature vector size.

2. The feature vector should be *continuous* with respect to small variations in geometry.

3. The feature vector should include *chemical* information, allowing for the subsequent similarity function to cluster similar chemistries together.

4. Lastly, the feature vector should be clearly, simply, and intuitively defined.

Motivated by the work of Villars[10], we design a feature vector based upon a substructure around a central ion. This choice allows for a clear decomposition of a given crystal structure into substructures; every ion is the center of it's own substructure. While Villars uses an atomic environment based upon Brunner and Schwarzenbach's maximum gap rule[16] for the radial distribution function, in this paper we chose to describe a feature vector based upon a weighted Voronoi polyhedron described by OKeeffe.[15] Both methods include geometric information that is independent of symmetry and continuous against small perturbations in crystal structure, but the clarity of the maximum gap rule breaks down when a maximum gap in the radial distribution function is not clearly discernable.

O'Keeffe's method is based upon the Voronoi decomposition of a crystal structure.[15] A crystal structure c can be decomposed into a set of Voronoi polyhedra with one central ion x inside each polyhedron.[17] The space enclosed by each polyhedron represents the set of points that are closer to the central ion x than any other ion. Each face of the polyhedron surrounding x is generated by the plane bisecting the line between x and a neighboring ion y, and subtends a solid angle Ω_y from the central ion. Following the work of OKeeffe, for a given polyhedron, the neighbor y corresponding to the greatest solid angle is assigned a weight $w_y = w_{\max} = 1$, and every other neighbor z is assigned a weight $w_z = \Omega_z/\Omega_{\max}$. O'Keeffe's method is illustrated in Figure 2.

Neighboring ions with greater Voronoi weights tend to be closer to the central ion i; they also tend to have fewer nearby neighbors within the same solid angle from the central ion. Correspondingly, neighbors with small Voronoi weights tend to be further away from the central ion i, and tend to have more nearby neighbors. O'Keeffe's method is elegant, continuous, and calculable with Barber's Quickhull algorithm.[18] Additionally, O'Keeffe's algorithm is intuitive and rigorously mathematically defined.

For the purposes of this paper, we represent the substructure s of a central ion i in a crystal structure c with the following data structure:

- We identify the central ion and its ionic specie i
- We keep an unordered list of neighboring ions, represented by (ionic species x and Voronoi weight w_x) pairs, $\{(x, w_x)\}$.

For example, a phosphate tetrahedron would be stored thus:

- Central ion: P^{5+}
- Peripheral ions: $(O^{2-}, 1)$, $(O^{2-}, 1)$, $(O^{2-}, 1)$, $(O^{2-}, 1)$

This choice of feature vector allows us to represent the atomic environment of an ion in a chemistry and geometrically sensitive manner. Furthermore, the feature vector will remain unchanged if the crystal structure is rescaled by a constant, allowing us to robustly compare the atomic environments of ions of differing radii. We note that the proposed definition of a substructure is mathematically rigorous in that it can be calculated for every crystal structure. It is independent of symmetry, and it is continuous against small variations in crystal structure with respect to both small perturbations of ion position and small changes in lattice vector.

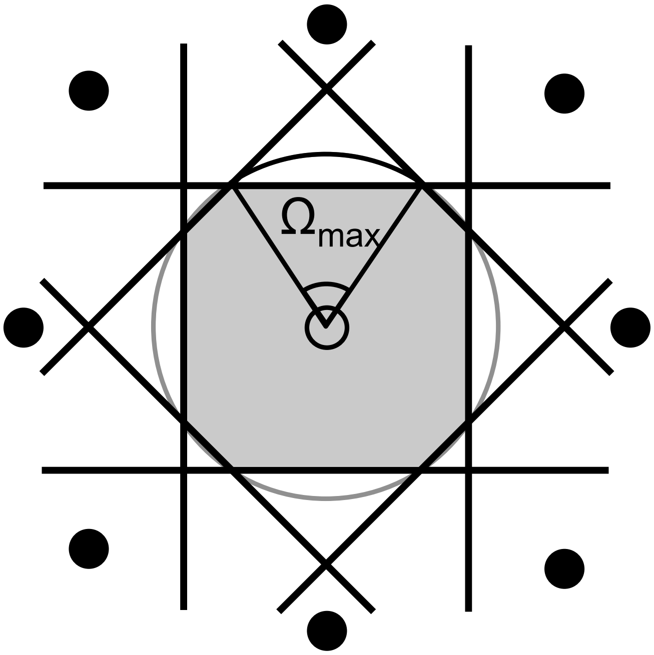

FIG. 2. The Voronoi polyhedron around a central ion is shown in gray. Faces of the Voronoi polyhedron are generated by the perpendicular bisectors of the lines between the central ion and all other ions in the crystal structure. Each face of the Voronoi polyhedron is thus caused by the existence of a neighboring ion y. The neighboring ion whose polyhedron face subtends the largest solid angle $\Omega_{\max}$ has weight 1; all other neighboring ions have weights proportional to the angle subtended by their polyhedron faces.

B. Similarity between Substructures

In this section we develop a similarity function between two substructures. This function should be higher if two substructures are chemically similar, and higher if two substructures are geometrically similar. This function should be 1 if two substructures are identical, and should always be greater than 0.

A similarity function between substructures must have a method of quantifying geometric similarity as well as a method of quantifying chemical similarity. In this section, we quantify the chemical similarity between ions via a previously developed data mined ionic substitutional similarity function[14,19]. This function, denoted $Sim_{\text{ion}}(i_1, i_2)$, grows with the probability of two ions substituting for each other within the same structure prototype and captures the similarity of two ions with respect to crystal structure formation. The ionic substitutional similarity of two identical ions is 1, and it decreases to a minimum of 0.

As a substructural similarity function balances chemical similarity versus geometric similarity, we introduce a tunable ionic similarity function that allows us to reweight the importance of chemical similarity. This tunable ionic similarity function is constructed by passing the ionic substitutional similarity $Sim_{\text{ion}}(i_1, i_2)$ through a sigmoid function.

$$\text{Sigmoid}_{\sigma,\,\mu}(x) = \frac{1}{\sigma\sqrt{2\pi\sigma}} \int e^{\dfrac{-(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x \qquad (1)$$

Setting μ sets the threshold above which ions are considered similar, and setting σ sets how sharply the function differentiates similar and dissimilar ions. Large σ values broaden the width of the Gaussian, minimizing the penalty for starkly differing chemistries. The function that quantifies the similarity between two ions i_1, i_2 is given by $Sim_{\text{ion}}^{\sigma,\,\mu}(i_1, i_2)$:

$$Sim_{\text{ion}}^{\sigma,\,\mu}(i_1, i_2) = \text{Sigmoid}_{\sigma,\,\mu}(Sim_{ion}(i_1, i_2)) \qquad (2)$$

Given the ionic substitutional similarity which quantifies chemical similarity two ions, we now develop a method for quantifying both the geometric and chemical similarity between two substructures. We will begin by developing a score that quantifies the geometric and chemical similarity between two peripheral ions in a substructure, and then define the similarity between two substructures via a best matching of the ions in one substructure to the ions in the other.

Two substructures s_i and s_j have central ions i and j and sets of neighbors N_i and N_j. Each neighboring ion $x \in N_i$ and $y \in N_j$ has an associated weight w_x or w_y that satisfies $0 \leq w \leq 1$.

We define a score between two peripheral ions x and y:

$$Score(x, y) = Sim_{\text{ion}}^{\sigma,\,\mu}(x, y) \min(w_x, w_y) e^{\frac{-(w_x, w_y)^2}{c^2}} \qquad (3)$$

This score satisfies the following properties:

- It is greater if the two ions are chemically similar, due to the contribution of the ionic substitution similarity function.
- It is greater if the weights of the two ions are higher, due to the contribution of the minimum of the two weights, and
- it is greater if the weights of the two ions are close to each other, due to the contribution of the Gaussian function.

The parameter c allows the user to tune the sensitivity with which the score penalizes different weights. A higher value of c yields a wider spread in the Gaussian, allowing for greater differences between the weights and lesser geometric sensitivity.

This score represents the similarity of the two neighboring ions to each other, taking into account both chemical and geometric similarity. It tends to be higher if the neighboring ions are more central to their respective substructures.

Next, we define a product between two substructures s_i and s_j:

$$Product(s_i, s_j) = \\ \max_{\text{all matchings}} \Sigma_{x,y \in \text{matching}} Score(x, y) \qquad (4)$$

where the sum is taken over the pairing of ions $x \in N_i$ to ions $y \in N_j$ that maximizes the product. If there are more ions in one substructure than the other, the excess ions in the larger substructure will remain unpaired, and will not contribute to the sum. This product between two substructures does not take into account the similarity of the central ion, and is thus appropriate for the analysis of void spaces or the comparison of substructures for which the central ion remains the same. We use this product throughout this paper as the application under consideration is the identification of Li sites.

An alternative product between two substructures that takes into account the similarity of the central ions should be used when comparing substructures with differing central ions. This product can be formulated by multiplying the product in equation 4 by the similarity of the central ions.

$$Product(s_i, s_j) = Sim_{\text{ion}}^{\sigma,\,\mu}(i, j) * \\ \max_{\text{all matchings}} \Sigma_{x,y \in \text{matching}} Score(x, y) \qquad (5)$$

Finally, we normalize the product of two substructures to obtain the substructural similarity function. This type of normalization is necessary to avoid assigning larger substructures larger similarities due to their greater number of neighboring ions.

$$Sim_{\text{substruct}}(s_i, s_j) = \\ \frac{Product(s_i, s_j)}{\sqrt{Product(s_i, s_i), Product(s_j, s_j)}} \qquad (6)$$

We define substructural similarity to be the resultant similarity function.

III. APPLICATION TO LI SITE PREDICTION

We now have the ability to parse crystal structures into substructures and to organize those substructures by similarity to each other. We have presented a tool kit for the systematic and quantitative study of substructures and void spaces. The methods in this paper can be used to extract databases of crystal substructures from existing materials databases, to mine for patterns in the crystal structures in which they form, to analyze and predict the environments of specific ions, and to study how substructures connect to each other.

To validate the capabilities of this set of tools, we chose an important and straightforward application to the problem of Li site identification. The search for better materials for Li ion batteries involves requires the identification of Li sites in potential electrode materials. We will use substructural similarity to characterize Li sites in the oxides, and use the database of Li sites to predict where Li ions can be inserted in new materials.

Oxides, defined as compounds consisting of over 20% oxygen by ion count, were extracted from the Inorganic Crystal Structure Database[4] (ICSD) 2012 and cleaned of high temperature and high pressure phases, peroxides and superoxides, and structures that were improperly reported. Details of the data cleaning procedure can be found in the appendix. The resultant oxides were sorted into structure prototypes using an affine mapping algorithm.[20] 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 5,509 oxides.

The complete data set was randomly split into 10 equally sized subsets for cross validation.[21] Each subset served in turn as a test set, while the other 9 subsets served as the training set. The training set, comprising 90% of the compounds, represents the database of known compounds to be data mined. The remaining 10%, called the test set, mimics a set of as-yet-unseen compounds from which we remove the Li ions and attempt to recover their positions. This test set is used to evaluate the efficacy of our site prediction algorithm.

A. Site Prediction Algorithm

The ionic substitution similarity functions was extracted from the training set. The parameters σ and μ, which were used in the tunable ionic similarity, and the parameter c, used in the substructural similarity, were set using nested cross-validation. Each training set was further subdivided into 5 partitions: Each partition was held apart as a test set in turn, creating 5 internal cross-validation sets for each external cross-validation set. The algorithm described below was run for each internal cross validation set while we varied each parameter σ, μ, and c in turn; the set of parameters that yielded the highest overall area under curve score (described below) was chosen. This optimal set of parameters $\sigma = 0.3$, $\mu = 0.7$, and $c = 0.05$ was then used to generate the results in the external cross-validation loop.

Each compound in the training set was searched for Li sites, and a database of Li substructures was compiled. The ionic substitution similarity and the database of Li substructures consist of all the information extracted from the training set.

For each Li-containing compound in the test set, the Li sites were removed. The resultant, unrelaxed, Li-free crystal structures represent artificially delithiated oxides for which we will recover the Li sites. The Voronoi polyhedra for the Li-free crystal structures were computed using the quickhull algorithm by Barber et al.[18] The set of points given by the corners of each Voronoi polyhedron and the centers of each Voronoi polyhedra face constitute a reasonable set of potential lithiation sites; we call this set of points the Voronoi points of the crystal structure. The Voronoi polyhedron corners represent the set of points that are equidistant from their four closest neighbors, and thus represent the set of points that are as far away as possible from any other point.[17] The face centers of Voronoi polyhedra are another potential site for inserted species. Finally, all sites and Voronoi points for each Li-containing compound were grouped together into symmetrically identical sites using pyspglib[22] to reduce computational complexity.

To test how well our data mined substructural similarity can predict favorable environments for Li ions, we attempted to rediscover the removed Li sites in each Li-containing compound in the test set. For each symmetrically distinct Voronoi point in a crystal structure, we calculated the substructural similarity between this Voronoi point and every known Li substructure obtained from our training set. We ranked each symmetrically distinct Voronoi point in the structure by its *distance* to the most similar Li containing substructure, where *distance* is given by $1 - \textit{similarity}$, to produce an ordered list. We evaluated the efficacy of this list in predicting Li sites via the following rules:

1. If the Voronoi point is within $r < 0.5$Å of an undiscovered Li site, it is considered a correct guess. That Li site and all of its symmetrically distinct neighbors are now marked as discovered, and cannot be discovered again.

2. If the Voronoi point is within $r < 0.25$Å of a previously guessed Voronoi point, this Voronoi point is not put forward as a possible interstitial site. This rule is necessary because Voronoi points are commonly found in compact clusters.

It is useful to compare the performance of ranking Voronoi points by substructural similarity against another ranking method. One simple and interesting ranking method involves ranking the sites by radius, where the radius is given by the distance from the Voronoi point to the nearest site. The average Li site radius in the oxide database was 2.1 Å. This is consistent with an oxygen ionic radius of 1.4 Å, and a lithium ionic radius of 0.7 Å[23,24]. The following section gives the results for ranking both via the substructural similarity method, and by

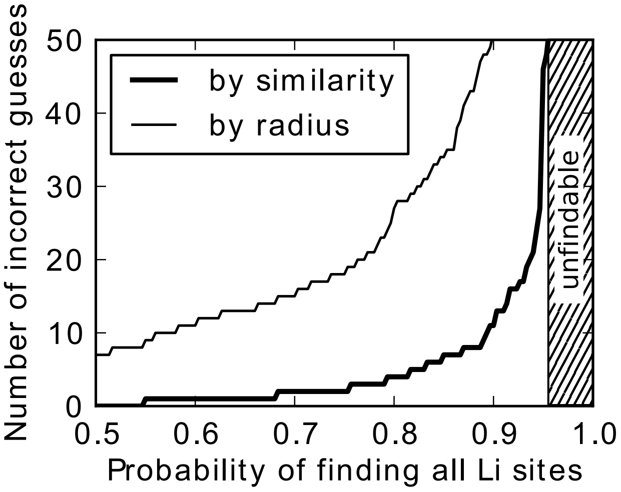

FIG. 3. The Li site identification rate. The solid line shows the number of incorrect guesses before identifying all the Li sites in a structure via the substructural similarity method; the dotted line shows the number of incorrect guesses using the radius of the site. Substructural similarity finds all the Li sites within a crystal structure with 9 incorrect guesses 90% of the time. The striped region on the right represents the 5% of Li sites that are not within 0.5Å of a Voronoi point, and thus cannot be found by this algorithm.

distance $d = |r - 2.1\text{Å}|$, where r is the site radius given by the distance from the site to the center of the closest ion.

B. Results

Figure 3 depicts the results of Li site prediction in the oxides by the two ranking methods given above. The x axis shows the probability of achieving a given result for a specific oxide selected from the data set; the y axis shows the number of wrong guesses necessary before finding all the Li sites. The data shown is aggregated across all 10 cross-validated sets, and thus represents Li site prediction across all unique 302 Li-containing oxides.

Ranking potential Li sites by substructural similarity fares better than ranking by site radius, consistently requiring 2-3 times fewer incorrect guesses. Substructural similarity finds all the Li sites within a crystal structure with 4 incorrect guesses 80% of the time. 50% of the time, it finds all the Li sites without any incorrect guesses. On the most difficult 5% of compounds the algorithm requires over 50 incorrect guesses to find all the Li sites. Upon investigation, we discover this is because approximately 5% of Li-ion sites are not within 0.5Å of any Voronoi point, and thus cannot be found by this algorithm. While 5% of Li-ion sites are not within 0.5Å of any Voronoi point, this does not mean that these 5% of Li-ion sites are unfindable via substructural similarity methods, but rather that future implementations should consider selecting potential Li sites via a method other than Voronoi decomposition. For example, it would be possible to discretize a given crystal structure into a 0.5Å grid, and use the resultant vertices as potential Li sites.

Figure 4 depicts the results of Li site prediction via a receiver operating characteristic (ROC) curve. A ROC curve depicts the true positive fraction versus the false positive fraction for a binary classifier as the predictive threshold is varied. In this case the binary classifiers at hand are classifying Voronoi points as Li sites or non Li sites. The true positive fraction or the sensitivity is the number of correctly identified Li sites divided by the number of actual Li sites. The false positive fraction is the number of non Li sites that have been incorrectly labeled as Li sites divided by the number of actual non Li sites. If a Voronoi point x has a similarity to the most similar Li site in the training set y, it's distance to the most similar Li site is given by $d = 1 - y$. The predictive threshold is the number p such that if $d \leq p$, point y is predicted to be a Li site. The ROC curve is generated by varying the predictive threshold p. As the predictive threshold rises, more sites, both Li and non-Li, will be labeled as Li sites. The ratio of true positives to false positives changes as we vary the distance threshold below which a given Voronoi point is classified as a Li site.

A perfect classifier should correctly identify all of the true positives before returning a false positive, and the ROC curve of a perfect classifier would go straight up from (0, 0) to (0, 1) before going right to (1, 1). The area under the ROC curve (AUC) is a commonly used figure of merit to assess the quality of a binary classifier. A perfect classifier has AUC = 1.

Examining figure 4, the benefits of Li site prediction by substructural similarity becomes clear. Substructural similarity achieves a higher area under the curve by correctly identifying Li sites earlier, before mis-identifying non-Li sites.

Finally, figure 5 depicts the performance of Li site classification broken down by compound complexity. Here, we define the compound complexity as the number of symmetrically distinct ion sites in the crystal structure; this number appears to be linearly correlated with the number of symmetrically distinct Voronoi points in the crystal structure (shown in black). Again, the number of guesses required to find all the Li sites by substructural similarity (shown in blue) is consistently lower than the number of guesses required to find all the Li sites by radius (shown in red). Interestingly, the number of guesses required to find all the Li sites does not appear to be correlated with the complexity of the compound. This implies that Li sites are well-separated from non-Li sites by the local substructural similarity. Finally, there are a number of outlying poor performers distributed across several compound complexities which require 100 or more guesses to identify all the Li sites. This may be because these poor performers contain Li environments that are

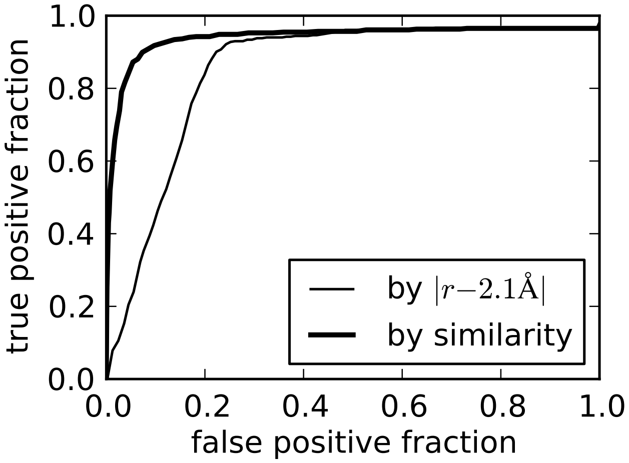

FIG. 4. The receiver operating characteristic for Li site classification. The x axis depicts the false positive fraction, or fraction of non-Li-site Voronoi points that were mistakenly identified as Li sites. The y axis depicts the true positive fraction, or the fraction of Li sites that were correctly identified as Li sites. The black line represents classification by substructural similarity; the dotted line represents classification by site radius.

not well represented in the training sets, and could be found with more comprehensive data.

IV. DISCUSSION

We have developed a definition of substructure that is mathematically precise, continuous against small variations in position, and calculable for every atomic position in every crystal structure. We built upon that definition of substructure a similarity function that takes into account geometric and chemical similarity and produces a number between 0 and 1 that is higher if the two substructures are more similar. This substructural similarity function allows for the quantitative study of sites, void spaces, and substructural packings within a crystal structure. We believe this function can be used for many purposes, including the prediction of substructures for structure prediction, the quantitative screening both sites and diffusion paths through a structure, to organize the space of crystal substructures and to screen for and analyze materials properties that are related to local substructures.

We validated the definition of substructure and substructural similarity functions via the prediction Li sites in oxide compounds, finding all the Li sites within a crystal structure with 9 incorrect guesses 90% of the time. This application has the potential to greatly reduce computational time in the search for Li insertion sites, as it is not uncommon for there to be hundreds of symmetrically distinct Voronoi points per oxide crystal structure. One of the strengths of the proposed data mining algorithm for interstitial insertion sites is that it is both general and flexible. While the presented work considers only the insertion of Li ions, there is no reason this work could not be extended to predict the insertion sites for Na or Mg ions; indeed, this framework can be used to identify, analyze, and predict the site and void space preferences of any ion. Ranking Voronoi points by substructural similarity finds the site preferences of an ion first, recovering the stable sites, but progressing to more unstable sites has the potential to discover the Voronoi points on the diffusion path of an ion through a material.

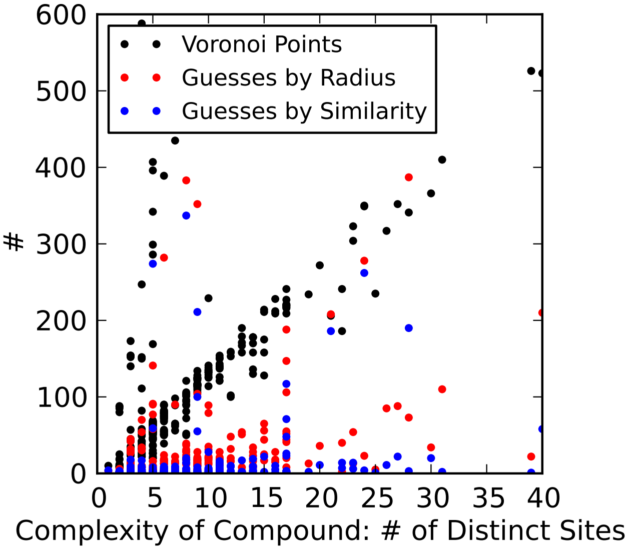

FIG. 5. Performance of Li site prediction by compound complexity. The number of incorrect guesses required to find all the Li sites is plotted versus the complexity of the oxide host. The complexity of the host is given by the number of symmetrically distinct sites in the host structure. Ranking by substructural similarity performs equally well on simple oxide structures as it does on more complex oxides.

Furthermore, the proposed data mining algorithm was used to validate the substructure and substructural similarity constructions. We present a brief analysis of the failure modes of this data mining algorithm. Of 312 lithiated oxides, there were 8 structures for which the Li site finding algorithm required more than 100 guesses to identify all the Voronoi points that were within 0.5Å of a Li site. Table I gives summary information for the those 8 lithiated structures. The table shows the number of symmetrically distinct Li sties, the number of guesses to find all the Li sites, and the number of Voronoi Points in the crystal structure. Finally, the table also gives two other quantities of interest. The minimum distance between any Voronoi point in the crystal structure and any Li site in the training set is sometimes pertinent; the average distance between any voronoi point and any Li site is approximately 0.6. Therefore if the *minimum* distance between any Voronoi point in the structure and any Li site is higher than 0.9, this crystal structure would be a statistical outlier and should be further examined. Secondly, the distance at which the last Li site was identified is pertinent as a measure of how unusual the most unusual Li site in the crystal structure is. The greater the distance, the more unusual the site. Table I gives all of the quantities described above.

Looking through Table I , we notice in the last two columns two structures with unusually high Voronoi distances. $Li_1Nd_9Mo_{16}O_{35}$ and $Li_8Rb_8B_{32}O_{56}$ were both reported in the ICSD with no charge state information. When ions are reported without charge state information, the algorithm searches for similar ions with a charge state of 0, yielding very low similarities to known substructures. The minimum distances between *any* Voronoi point and all known Li sites for these two structures are 0.92 and 0.95. Error due to unreported charge states can be easily addressed in future work by assigning reasonable charge states. Additionally, approximately 5% of Li sites are never found because 5% of the Li sites are not within 0.5 Åof a Voronoi point; this is not a flaw in the ranking of sites by substructural similarity, but rather a flaw in the construction of the set of possible sites.

The other 6 structures illustrate a weakness of any data mining algorithm, in that the data mined predictions are only as good as the data set at hand. The distances between the last found Li site and the closest Li site in the training set for each of these structures is greater 0.44, whereas the average distance between an Li site and the closest Li site in the training set is 0.31. In contrast, the average distance between a randomly drawn Voronoi point and the closest Li site in the training set is 0.57. The distribution of Li site distances and the number of guesses necessary to find an Li site is given in figure 6. The red line and the right hand axis plot the distance of a site to the closest Li site in the training set versus the average number of guesses necessary to find it; the blue line and the left hand axis plot the distance of a site to the closest Li site in the training set versus the percentage of Li sites in the test set that are found at that distance. A good prediction algorithm increases the lag between the blue curve and the red curve, identifying all of the Li sites before increasing the number of guesses. The other 6 structures represent the tail end of the red curve; the unusually high distances of the last Li sites to be found indicate that there are no highly similar Li sites in the training set. This is either because of an usual chemistry as demonstrated by $Li^{+}_{10}As^{5+}_{22}U^{6+}_{26}O^{2-}_{138}$, or an unusual geometry.

While the substructural data mining methods developed in this paper are general and could theoretically be applied to any number of ionic species, the limitation of a data mining method lies in the quality of the data set at hand. For example, the quality of the Li site predictions depends strongly on the number of Li sites in the database; in this case, the oxide database provided 1631 occurrences of Li in 312 crystal structures across 458 unique substructures. However, if we were to repeat this prediction procedure for Mg site predictions, the same oxide database would provide only 839 occurrences of Mg in 140 crystal structures across 176 unique substructures. For this reason, we expect data mining predictions to fare less well when applied to less common chemistries. In such a scenario, it would be reasonable to use ionic similarity to gather data from across similar chemistries. We could include data from ions with high ionic substitutional similarity (like Ca and Ba), weighted by ionic substitutional similarity, when making predictions for Mg sites.

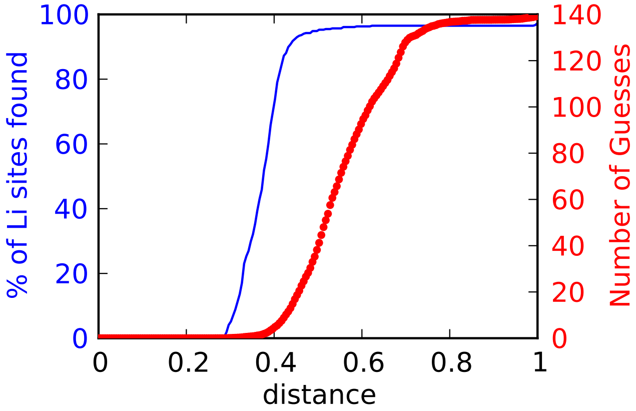

FIG. 6. Distance distribution of Li sites and number of guesses. The x axis depicts the distance as calculated by the similarity metric between an Li site in the test set and the closest Li site in the training set. The blue y axis on the left hand side depicts the distribution of Li sites versus distance; the red y axis on the right hand side depicts the average number of guesses at a given distance. An ideal ranking system would increase the lag between the blue and the red lines, finding 100% of the Li sites before requiring more than 1 guess.

There are several reasonable extensions of the current algorithm. We extracted a list of known Li sites and compared potential sites directly to the known sites, ranking potential sites by similarity to a known site. It would be reasonable to take into account other factors - for instance, the radius of the potential site, the ratio of anions to cations in the host structure, or even the composition of the host structure. One could easily extend the algorithm by conditioning the ranking of the potential site upon not only substructural similarity to an Li site, but also substructural similarity to a known Na site. One could also condition the ranking upon the composition of the host structure being similar to the composition of a structure that is known to host Li. Another algorithm would search through the database for structures with similar compositions and extract Li sites from those compositions only. There are many possibilities for reasonable site prediction algorithms, each with its own tradeoffs in terms of computational time, dependency on the robustness of the data set at hand, and quality of potential results.

Examining the ROC curve shown in figure 4 more

TABLE I. Failure Modes for Li Site Prediction

Composition	Li Sites	Guesses	Voronoi Points	Minimum Distance	Maximum Li Distance
$Li_3^+V_2^{3+}(P^{5+}O_4^{2-})_3$	2	117	221	0.36	0.51
$Li_1Nd_9Mo_{16}O_{35}$	1	186	206	0.92	0.99
$Li_{56}^+Si_{14}^{4+}O_{56}^{2-}$	19	190	909	0.33	0.44
$Li_{18}^+Cr_6^{3+}P_{16}^{5+}O_{58}^{2-}$	3	211	1049	0.31	0.44
$Li_{10}^+As_{22}^{5+}U_{26}^{6+}O_{138}^{2-}$	5	236	1220	0.33	0.44
$Li_8Rb_8B_{32}O_{56}$	2	262	349	0.95	0.99
$Li_{12}^+W_6^{6+}O_{24}^{2-}$	2	274	396	0.36	0.54
$Li_{10}^+Cl_2^-B_{14}^{3+}O_{25}^{2-}$	2	357	615	0.41	0.60

closely, we find that all curves reach their maximum true positive rate at around 95%. Approximately 5% of the Li sites are never found because 5% of the Li sites are not within 0.5Å of a Voronoi point. A reasonable extension of this work would include a more thorough search for potential Li sites; it is not clear that the Voronoi points are an optimal set. From figure 5, we infer that the number of incorrect guesses required by the substructural similarity algorithm does not grow with the number of Voronoi points, so the performance cost of adding more potential Li sites is minimal. The cost of computing the substructural similarity is low and easily parallelized. One could consider discretizing the space within a given crystal structure into 0.5Å cubes and reducing the resultant points by symmetry.

It is worth mentioning that while the framework presented in this paper captures only local, first-neighbor interactions, there are a number of potentially meaningful extensions to explore. It would not be difficult to extend the substructures in this work to include second-neighbor interactions by representing each substructure as a graph that includes second-neighbor connections. Alternatively, one can view each crystal structure as an overlapping tiling of substructures; each ion participates not only in the substructure to which it is central, but also in all of it's first neighbor substructures. Using this framework, one can mine for patterns in the interconnectivity of substructures. What substructures tend to overlap the most? What combination of substructures allows for Li-ion diffusion?

We have outlined a mechanism for the prediction of Li sites in the oxides. For the purpose of validation, we began with lithiated oxide crystal structures, removed the Li ions, and then predicted the Li sites in the artificially delithiated structures. This construction allowed us to recover an experimentally verified set of Li sites. However, when applying this algorithm to the identification of Li sites in delithiated structures, we expect to obtain less accurate results for two reasons. Firstly, structures relax when Li ions are added or removed. The artificially delithiated structures we ran our predictions on were not relaxed; in essence, they were artificially frozen in a structure with Li-ion vacancies, making it easier to identify Li-ion sites. Secondly, this algorithm does not take into account the effect of Li concentration on the Li sites predicted. In effect, this algorithm is a mechanism for the prediction of Li sites in the dilute limit. It would be theoretically possible to insert Li ion-by-ion into a structure, re-running the prediction algorithm between each insertion to find the next Li site. As this algorithm only takes into account local effects, we expect the ability of this algorithm to predict Li orderings to be limited Another factor to consider is that in a given lithiated, test set structure, it is possible that there were more Li sites than reported experimentally. Perhaps the experimentally reported structure was not fully lithiated. In this case, our algorithm would have found a number of Li sites that were counted as incorrect guesses, penalizing not the accuracy of our model but the performance of our model under the given test.

We have used a definition of substructural similarity that can be tuned for greater or lesser geometric and chemical sensitivities. We set the tuning parameters σ, μ, and c to maximize the area under the ROC curve for the application of predicting Li sites in the oxides. However, differing applications of the substructural similarity function will call for different tunings. For example, the current algorithm ranks all potential Li sites across a host of candidate oxide compounds. Another potential application of substructural similarity would be to rank potential Li sites within a single oxide compound to predict diffusion pathways. The substructural similarity function can be used to determine which void spaces are more likely to hold Li than others. In this second application, we would expect the optimal tuning of the substructural similarity to be more sensitive to geometric differences and less sensitive to chemistry, as finding a diffusion path requires the ability to distinguish between small geometric differences.

This paper has presented a definition of substructure that is mathematically rigorous, continuous with respect to small displacements in ion position, and dependent upon both chemistry and geometry. We have further defined a similarity function between any two substructures that is 1 if the two substructures are identical, and decays towards 0 with growing differences in geometry and chemistry. This definition of substructure and substructural similarity allow for the decomposition and analysis of crystal structure databases. We have validated substructural analysis via the reproduction of Li sites in the oxides.

V. ACKNOWLEDGMENTS

We gratefully acknowledge Anthony Gamst, Randall Notestine, Shyue-Ping Ong and Bo Xu for helpful conversation. Additionally, this work was supported in part by the SI2-SSI Collaborative Research program of the National Science Foundation under Award No. OCI-1147503. We acknowledge the use of data and code from the Materials Project, which was supported in part by Department of Energys Basic Energy Sciences program under Grant No. EDCBEE.

VI. APPENDIX: DATA SET CLEANING

The data set that we used to validate substructure similarity was very similar to the data set used to calculate composition similarity[14]. The oxide data set constructed for the work in this section differs from the one used in the composition similarity paper in that the compounds were additionally screened for charge-balanced structures only; the ICSD database was updated to include all data from year 2014, and the affine mapping algorithm used to prototype the database was updated to pymatgen version 2.1.2[25].

All of the compounds in the Inorganic Crystal Structure Database[4] (ICSD) 2012 were searched for compounds that satisfied the following criteria:

- Compounds must be oxides, as indicated by at least 20% oxygen content by ion count.
- Compounds must not be peroxides or superoxides, as indicated by O-O bond lengths $L < 1.50$ Å.
- Compounds must not be marked high pressure, HP, high temperature, or HT.
- Compounds must not have improbably short ($<$ 1Å) bond lengths.
- Compounds must not have a mismatch between the reported composition and the ions given in the crystal structure.
- Compounds must not contain hydrogen. The reported crystal structures of compounds containing hydrogen are often unreliable.
- Compounds must be charge balanced, as indicated by the total charge of all the species reported summing to an absolute value $< |0.001|$.

[*] gceder@mit.edu

[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, Computational Materials Science **50**, 2295 (2011).

[3] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Materials **1**, 011002 (2013).

[4] G. Bergerhoff, I. Brown, F. Allen, G. Bergerhoff, and R. Sievers, Chester: International Union of Crystallography (1987).

[5] A. R. Oganov and M. Valle, The Journal of chemical physics **130**, 104504 (2009).

[6] E. Willighagen, R. Wehrens, P. Verwer, R. De Gelder, and L. Buydens, Acta Crystallographica Section B: Structural Science **61**, 29 (2005).

[7] R. de Gelder, IUCr CompComm Newsletter **7**, 59 (2006).

[8] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical review letters **108**, 058301 (2012).

[9] L. Pauling, Journal of the American Chemical Society **51**, 1010 (1929).

[10] J. Daams and P. Villars, Engineering Applications of Artificial Intelligence **13**, 507 (2000).

[11] C. Mellot Draznieks, J. M. Newsam, A. M. Gorman, C. M. Freeman, and G. Férey, Angewandte Chemie International Edition **39**, 2270 (2000).

[12] C. Mellot-Draznieks, S. Girard, G. Férey, J. C. Schön, Z. Cancarevic, and M. Jansen, Chemistry-A European Journal **8**, 4102 (2002).

[13] C. Mellot-Draznieks, J. Dutour, and G. Férey, Angewandte Chemie International Edition **43**, 6290 (2004).

[14] L. Yang and G. Ceder, Phys. Rev. B **88**, 224107 (2013).

[15] M. O'Keeffe, Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography **35**, 772 (1979).

[16] G. Brunner and D. Schwarzenbach, Zeitschrift fur Kristallographie **133**, 127 (1971).

[17] G. Voronoi, Journal für die reine und angewandte Mathematik **1908**, 97 (1908).

[18] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, ACM Transactions on Mathematical Software (TOMS) **22**, 469 (1996).

[19] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorganic Chemistry **50**, 656 (2011).

[20] H. Burzlaff and Y. Malinovsky, Acta Crystallographica Section A **53**, 217 (1997).

[21] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*, Vol. 2 (Springer, 2009).

[22] A. Togo, "spglib: A c library for finding and handling crystal symmetries," http://spglib.sourceforge.net/ (2009).

[23] R. D. Shannon, Acta Crystallographica Section A **32**, 751 (1976).

[24] I. Brown, Acta Crystallographica Section B: Structural Science **44**, 545 (1988).

[25] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).