Inorganic Chemistry
Article

Inorg. Chem. **XXXX**, *XXX*, 000–000 **A**
DOI: 10.1021/ic102031h

Data Mined Ionic Substitutions for the Discovery of New Compounds

Geoffroy Hautier, Chris Fischer, Virginie Ehrlacher, Anubhav Jain, and Gerbrand Ceder*

Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

Received October 6, 2010

The existence of new compounds is often postulated by solid state chemists by replacing an ion in the crystal structure of a known compound by a chemically similar ion. In this work, we present how this new compound discovery process through ionic substitutions can be formulated in a mathematical framework. We propose a probabilistic model assessing the likelihood for ionic species to substitute for each other while retaining the crystal structure. This model is trained on an experimental database of crystal structures, and can be used to quantitatively suggest novel compounds and their structures. The predictive power of the model is demonstrated using cross-validation on quaternary ionic compounds. The different substitution rules embedded in the model are analyzed and compared to some of the traditional rules used by solid state chemists to propose new compounds (e.g., ionic size).

1. Introduction

The discovery of new inorganic compounds is critical to the development of many technologically relevant fields (e.g., high temperature superconductivity, catalysis, or energy storage).[1] Unfortunately, searching for new solid state compounds can be very slow, involving mainly a combination of chemical intuition and serendipity.[2]

Solid phase stability can be accurately and efficiently predicted through *ab initio* computations in the density functional theory (DFT) framework.[3−5] This offers the opportunity to accelerate the materials discovery process by orienting the experimentalist to computationally predicted compounds of potential interest. The main direction pursued nowadays to address this compound and crystal structure prediction problem consists in approaching it as an optimization problem.[6] An optimization algorithm (e.g., simulated annealing[7,8] or genetic algorithm[9−11]) is used to find the crystal structure parameters (lattice constants, angles, and atomic coordinates) minimizing the energy obtained by a model such as DFT. While appealing as a quite exhaustive search, and successful in some cases reported in the literature (see for instance Oganov et al.[12]), this optimization approach requires very significant computational resources especially when used in conjunction with an *ab initio* energy model.[13] For example, around a thousand energy evaluations were needed to find the high-pressure ground state for $MgSiO_3$ using a genetic algorithm.[10]

On the other hand, even before *ab initio* computations were broadly available, new compounds and their crystal structures have been suggested by heuristic models such as Hume−Rothery rules[14] or structure maps.[15−17] Those models generally correlate the formation of specific structures with atomic factors such as size, electronegativity, number of electrons, or position in the periodic table, and have been shown to have reasonable predictive capability across a limited range of crystal structures.[18]

*To whom correspondence should be addressed. E-mail: gceder@mit.edu.

(1) Cheetham, A. *Science* **1994**, *264*, 794–795.
(2) DiSalvo, F. J. *Pure Appl. Chem.* **2000**, *72*, 1799–1807.
(3) Curtarolo, S.; Morgan, D.; Ceder, G. *CALPHAD: Comput. Coupling Phase Diagrams Thermochem.* **2005**, *29*, 163–211.
(4) Ong, S. P.; Wang, L.; Kang, B.; Ceder, G. *Chem. Mater.* **2008**, *20*, 1798–1807.
(5) Akbarzadeh, A. R.; Wolverton, C.; Ozolins, V. *Phys. Rev. B* **2009**, *79*, 1–10.
(6) Woodley, S. M.; Catlow, R. *Nat. Mater.* **2008**, *7*, 937–946.
(7) Pannetier, J.; Bassas-Alsina, J.; Rodriguez-Carvajal, J.; Caignaert, V. *Nature* **1990**, *346*, 343–345.
(8) Schön, J. C.; Jansen, M. *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 1286–1304.
(9) Abraham, N. L.; Probert, M. I. J. *Phys. Rev. B* **2006**, *73*, 1–6.
(10) Oganov, A. R.; Glass, C. W. *J. Phys.: Condens. Matter* **2008**, *20*, 064210.
(11) Woodley, S. M.; Battle, P. D.; Gale, J. D.; Richard, A.; Catlow, C. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2535–2542.
(12) Oganov, A. R.; Chen, J.; Gatti, C.; Ma, Y.; Ma, Y.; Glass, C. W.; Liu, Z.; Yu, T.; Kurakevych, O. O.; Solozhenko, V. L. *Nature* **2009**, *457*, 863–868.
(13) The computational time required for a total energy computation with ionic relaxation can vary greatly depending on the system, parameters, and precision required. Our own computational cost average estimate is around 100 CPU hours for one standard generalized-gradient approximation (GGA) computation on a Intel Xeon 5140 2.33 GHz CPU.
(14) Hume-Rothery, W.; Raynor, G. V. *The Structure of Metals and Alloys*; Institute of Metals: London, U.K., 1962.
(15) Pettifor, D. G. *J. Chem. Soc., Faraday Trans.* **1990**, *86*, 1209–1213.
(16) Villars, P. *J. Less Common. Met.* **1983**, *92*, 215–238.
(17) Muller, O.; Roy, R. *The major ternary structural families*; Springer-Verlag: New York, 1974.
(18) Morgan, D.; Rodgers, J.; Ceder, G. *J. Phys.: Condens. Matter* **2003**, *15*, 4361–4369.

Finding inspiration in this tradition of empirical models, some recent work has shown that statistical knowledge models trained by extracting the "chemical rules" present in a crystal structure database can be very efficient in the discovery of new compounds when combined with the accuracy of DFT. For instance, a model based on correlation existing between crystal structure prototypes at different composition has recently been used to predict around 100 new ternary oxides with a limited computational budget.[19,20]

In this work, we present how a probabilistic model can be built to assess the likelihood for ionic species to substitute for each other while retaining the crystal structure. We describe the mathematical model and its training on an experimental crystal structures database. The model's power in predicting compounds is then evaluated by cross-validation. Finally, the chemical rules this model captures are discussed and compared to more traditional approaches based on ionic size or position in the periodic table.

2. Method

2.1. Ionic Substitution Approach to New Compound Discovery. Chemical knowledge often drives researchers to postulate new compounds based on substitution of elements or ions from another compound. For instance, when the first superconducting pnictide oxide $LaFeAsO_{1-x}F_x$ was discovered, crystal chemists started to synthesize many other isostructural new compounds by substituting lanthanum with other rare-earth elements such as samarium.[21]

A formalization of this substitution approach exists in the Goldschmidt rules of substitution stating that the ions closest in radius and charge are the easiest to substitute for each other.[22] While those rules have been widely used to rationalize a posteriori experimental observations, they lack a real quantitative predictive power.

Our approach follows this substitution idea but develops a mathematical and quantitative framework around it. The basic principle is to learn from an experimental database how likely the substitution of certain ions in a compound will lead to another compound with the same crystal structure. Mathematically, the substitution knowledge is embedded in a substitution probability function. This probability function can be evaluated to assess quantitatively if a given substitution from a known compound is likely to lead to another stable compound. For instance in the simple case of the $LaFeAsO_{1-x}F_x$ compound, we expect the probability function to indicate a high likelihood of substitution between La^{3+} and Sm^{3+} and thus a high likelihood of existence for the $SmFeAsO_{1-x}F_x$ compound in the same crystal structure as $LaFeAsO_{1-x}F_x$ but with Sm on the La sites.

Our method follows an approach used in the field of machine translation.[23] The aim of machine translation, is to develop models able to translate texts from one language to another. Therefore, one approach is to build probabilistic models that evaluate the probability for a word in one language to correspond to another word in another language. In the case of our ionic substitution model, the approach is similar, but it is a correspondence between ionic species instead of words that is sought.

2.2. Probabilistic Model. We present here the different variables and the mathematical form of the substitution probabilistic model.

Let us represent a compound formed by n different ions by a n component vector:

$$\mathbf{X} = (X_1, X_2, ..., X_n) \quad (1)$$

Each of the X_j variables are defined on the domain Ω of existing ionic species

$$\Omega = \{Fe^{2+}, Fe^{3+}, Ni^{2+}, La^{3+}, ...\} \quad (2)$$

The quantity of interest to assess the likelihood of an ionic substitution is the probability p_n for two n-component compounds to exist in nature in the same crystal structure. If X_j and X_j' respectively indicate the ions present at the position j in the crystal structure common to two compounds, then one needs to determine:

$$p_n(\mathbf{X}, \mathbf{X}') = p_n(X_1, X_2, ..., X_n, X'_1, X'_2, ..., X'_n) \quad (3)$$

Knowing such a probability function allows to assess how likely any ionic substitution is. For example, by computing $p_4(Ni^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$, one can evaluate how likely Fe^{2+} in a lithium transition metal phosphate is to be substituted by Ni^{2+}. In this specific example, this value is expected to be high as Ni^{2+} and Fe^{2+} are both transition metals with similar charge and size. Actually, $LiNiPO_4$ and $LiFePO_4$ both form in the same olivine-like structure. On the other hand, the substitution of Fe^{2+} by Sr^{2+} would be less likely and $p_4(Sr^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$ should have a low value. We must point out that the probability function does not have any crystal structure dependence. The fact that the compound targeted for substitution forms an olivine structure does not influence the result of the evaluated probability. This is an approximation in our approach.

The probability function $p_n(\mathbf{X},\mathbf{X}')$ is a multivariate function defined in a high-dimensional space and cannot be estimated directly. For all practical purposes, this function needs to be approximated. We follow here an approach successfully used in other fields such as machine translation, and based on the use of binary indicators f, so-called *feature functions*.[24] These feature functions are mathematical representations of important aspects of the problem. The only mathematical requirement for a feature function is to be defined on the domain of the probability function $(\mathbf{X},\mathbf{X}')$ and return 1 or 0 as result. They can be as complex as required by the problem.

(19) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. *Nat. Mater.* **2006**, *5*, 641–646.
(20) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. *Chem. Mater.* **2010**, *22*, 3762–3767.
(21) Johrendt, D.; Pöttgen, R. *Angew. Chem., Int. Ed.* **2008**, *47*, 4782–4784.
(22) Goldschmidt, V. *Naturwissenschaften* **1926**, *14*, 477–485.
(23) Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; Mercer, R. L. *Comput. Linguistics* **1993**, *19*, 263–312.
(24) Berger, A.; Della Pietra, V. J.; Della Pietra, S. A. *Comput. Linguistics* **1996**, *22*, 39–72.

For an ionic substitution model, one could choose for example as a feature function:

$$f(\mathbf{X},\mathbf{X}') = \begin{cases} 1 & \text{if } Ca^{2+} \text{substitutes for } Ba^{2+} \text{in the presence of } O^{2-} \\ 0 & \text{else} \end{cases} \tag{4}$$

The relevant feature functions are commonly defined by experts from prior knowledge. If our chosen set of feature functions are informative enough, we expect to be able to approximate the probability function by a weighted sum of those feature functions:

$$p_n(\mathbf{X},\mathbf{X}') \approx \frac{e^{\sum_i \lambda_i f_i^{(n)}(\mathbf{X},\mathbf{X}')}}{Z} \tag{5}$$

The λ_i indicate the weight given to the feature $f_i^{(n)}(\mathbf{X},\mathbf{X}')$ in the probabilistic model. Z is a partition function ensuring the normalization of the probability function. The exponential form chosen in eq 5 follows a commonly used convention in the machine learning community.[25]

2.3. Binary Feature Model. A first assumption we make is to consider that the feature functions do not depend on the number n of ions in the compound. Simply put, we assume that the ionic substitution rules are independent of the compound's number of components (binary, ternary, quaternary, ...).

Therefore, we will omit any reference to n in the probability and feature functions. Equation 5 becomes

$$p(\mathbf{X},\mathbf{X}') \approx \frac{e^{\sum_i \lambda_i f_i(\mathbf{X},\mathbf{X}')}}{Z} \tag{6}$$

While the feature functions could be more complex, only simple binary substitutions are considered in this paper. This means that the likelihood for two ions to substitute to each other is independent of the nature of the other ionic species present in the compound. Mathematically, this translates in assuming that the relevant feature functions are simple binary features of the form:

$$f_k^{a,b}(\mathbf{X},\mathbf{X}') = \begin{cases} 1 & X_k = a \text{ and } X_k' = b \\ 0 & \text{else} \end{cases} \tag{7}$$

Each pair of ions a and b present in the domain Ω is assigned a set of feature functions with corresponding weights $\lambda_k^{a,b}$ indicating how likely the ions a and b can substitute in position k. For instance, one of the feature function will be related to the Ca^{2+} to Ba^{2+} substitution.

$$f_k^{Ca^{2+},Ba^{2+}}(\mathbf{X},\mathbf{X}') = \begin{cases} 1 & X_k = Ca^{2+} \text{ and } X_k' = Ba^{2+} \\ 0 & \text{else} \end{cases} \tag{8}$$

The magnitude of the weight $\lambda_k^{Ca^{2+},Ba^{2+}}$ associated with this feature function indicates how likely this binary substitution is to happen.

Finally, the features weights should satisfy certain constraints for any permutation of the components to not change the result of the probability evaluation. Those symmetry conditions are

$$\lambda_k^{a,b} = \lambda_k^{b,a} \tag{9}$$

and

$$\lambda_k^{a,b} = \lambda_l^{a,b} \tag{10}$$

2.4. Training of the Probability Function. While the mathematical form for our probabilistic model is now well established, the model parameters (the weights $\lambda_k^{a,b}$) still need to be evaluated. Those weights are estimated from the information present in an experimental crystal structure database.

From any experimental crystal structure database, structural similarities can be obtained using structure comparison algorithms.[26,27] For instance, $CaTiO_3$ and $BaTiO_3$ both form cubic perovskite structures with Ca and Ba on equivalent sites. This translates in our mathematical framework as a specific assignment for the variables vector $(\mathbf{X},\mathbf{X}') = (Ca^{2+}, Ti^{4+}, O^{2-}, Ba^{2+}, Ti^{4+}, O^{2-})$. We will follow the convention in probability theory designing specific values of the random variable vector $(\mathbf{X},\mathbf{X}')$ by lower case letters $(\mathbf{x}, \mathbf{x}')$. An entire crystal structure database D will lead to m assignments: $(\mathbf{X},\mathbf{X}') = (\mathbf{x}, \mathbf{x}')^t$ with $t = 1, ..., m$

$$D = \{(\mathbf{X},\mathbf{X}') = (\mathbf{x},\mathbf{x}')^1, (\mathbf{X},\mathbf{X}') = (\mathbf{x},\mathbf{x}')^2, ..., (\mathbf{X},\mathbf{X}') = (\mathbf{x},\mathbf{x}')^{m-1}, (\mathbf{X},\mathbf{X}') = (\mathbf{x},\mathbf{x}')^m\} \tag{11}$$

Coming back to our analogy to machine translation, probabilistic translation models are estimated from databases of texts with their corresponding translation. The analog to the translated texts database in our substitution model is the crystal structure database.

Using these assignments obtained from the database, we follow the commonly used maximum-likelihood approach to find the adequate weights from the database available.[28] The weights maximizing the likelihood to observe the training data are considered as the best estimates to use in the model. For notational purpose we will represent the set of weights by a weight vector $\boldsymbol{\lambda}$.

From those m assignments, the log-likelihood l of the observed data D can be computed:

$$l(D,\boldsymbol{\lambda}) = \sum_{t=1}^{m} \log p((\mathbf{x},\mathbf{x}')^t|\boldsymbol{\lambda}) \tag{12}$$

$$= \sum_{t=1}^{m} [\sum_i \lambda_i f_i((\mathbf{x},\mathbf{x}')^t) - \log Z(\boldsymbol{\lambda})] \tag{13}$$

(25) Della Pietra, S. A.; Della Pietra, V. J.; Lafferty, J. *IEEE Trans. Pattern Anal. Machine Intell.* **1997**, *19*, 1–13.

(26) Parthé, E.; Gelato, L. *Acta Crystallogr., Sect. A* **1984**, *40*, 169–183.

(27) Hundt, R.; Schön, J. C.; Jansen, M. *J. Appl. Crystallogr.* **2006**, *39*, 6–16.

(28) Eliason, S. R. *Maximum Likelihood Estimation: Logic and Practice*; Sage Publications, Inc: Thousand Oaks, CA, 1993.

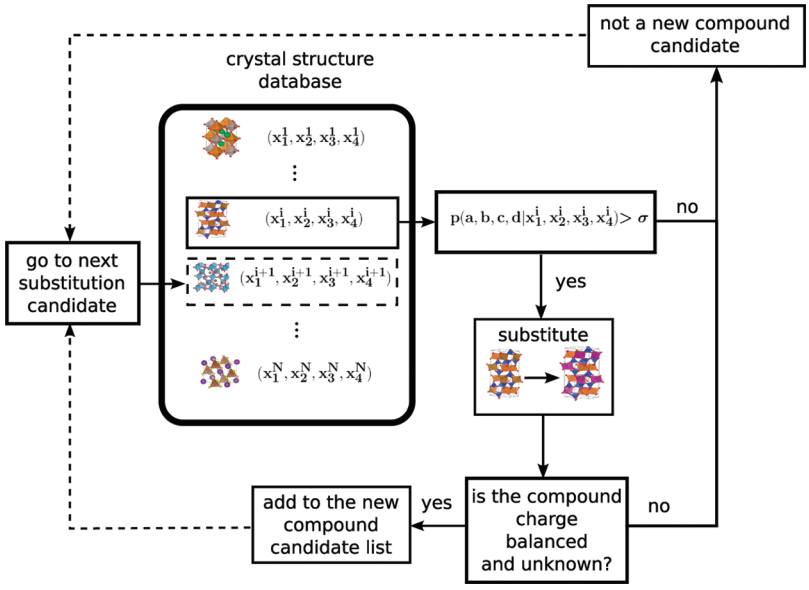

Figure 1. Procedure to predict new compounds formed by the *a*, *b*, *c*, and *d* species using the substitutional probabilistic model.

The feature weights maximizing the log-likelihood of observing the data D ($\boldsymbol{\lambda}_{ML}$) are obtained by solving:

$$\boldsymbol{\lambda}_{ML} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ l(D, \boldsymbol{\lambda}) \tag{14}$$

There is a last caveat in the training of this probability function. Any ionic pair never observed in the data set could theoretically have any weight value. All those unobserved ionic pair weights will be set to a common value α. As these ionic pairs should be unlikely, a low value of α (for instance $\alpha = 10^{-5}$ in the rest of this work) will be used. A rational way to set this α value is to use cross-validation to find its optimal value in terms of predictive power. Multiple cross-validations could be run for different values of α. The quality of the prediction could be then compared for each of those cross-validations. From this comparison, an optimal α maximizing the predictive power of the model could be chosen.

2.5. Compound Prediction Process. When the substitution probabilistic model in eq 6 has been trained, it can be used to predict new compounds and their structures from a database of existing compounds. The procedure to predict a compound formed by species *a*, *b*, *c*, and *d* is presented in Figure 1. For each compound containing (x_i^1, x_i^2, x_i^3, x_i^4) as ionic species, the probability to form a new compound by substitution of *a*, *b*, *c*, and *d* for x_i^1, x_i^2, x_i^3, and x_i^4 is evaluated by computing $p(a, b, c, d|x_i^1, x_i^2, x_i^3, x_i^4)$. If this probability is higher than a given threshold σ, the substituted structure is considered. If this new compound candidate is charge balanced and previously unknown, it can be added to our list of new compounds candidates. If not, the algorithm goes to the next $i + 1$ compound in the crystal structure database. The substitutions proposed by the model do not have to be isovalent. However, all suggested compounds have to be charge balanced.

At the end of the new compound prediction process, a list of new compounds candidates in the *a*, *b*, *c*, *d* chemistry is available. This list should be tested in a second step for stability versus all already known compounds by accurate *ab initio* techniques such as DFT.

3. Results

A binary feature model based on the ternary and quaternary ionic compounds present in the Inorganic Crystal Structure Database (ICSD)[29] has been built. In this work, we consider a compound to be ionic if it contains one of the following anions: O^{2-}, N^{3-}, S^{2-}, Se^{2-}, Cl^-, Br^-, I^-, F^-. Only ordered compounds (i.e., compounds without partially occupied sites) are considered. Crystal structure similarity was found by an affine mapping technique and used to obtain the database D of m assignments (eq 11) necessary to train the model.[27] A binary feature model was fitted on this data set using a maximum likelihood procedure as presented in the methods section.

3.1. Cross-Validation on Quaternary ICSD Compounds. The procedure to discover new compounds using the probabilistic model was presented in the methods section. Using this procedure, we evaluated the predictive power of this approach by performing a cross-validation test.[30] Cross-validation consists in removing part of the data available (the test set) and training the model on the remaining data set (the training set). The model built in this way is then used to predict back the test set and evaluate its performance. We divided the quaternary ordered and ionic chemical systems from the ICSD in 3 equal-sized groups. We performed 3 cross-validation tests using all compounds in one of the group as test set and the remaining quaternary and ternary compounds as training set. This extensive cross-validation tested 2967 compounds in total. The cross-validation tests excluded compounds forming in prototypes unique to one compound, as our substitution strategy by definition cannot predict

(29) ICSD, Inorganic Crystal Structure Database, 2006; http://icsd.fiz-karlsruhe.de/icsd/.

(30) Hastie, T.; Tibshirani, R.; Friedman, J. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2009; Chapter 4, pp 80−113.

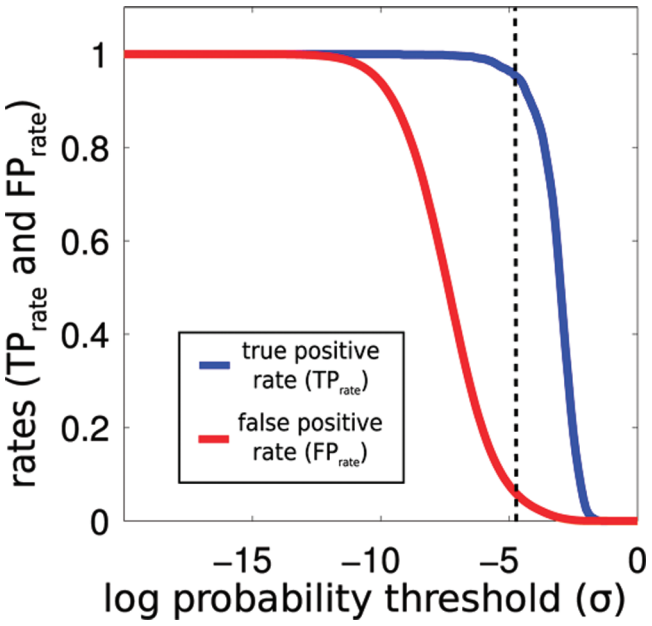

Figure 2. True positive rate (TP_{rate}, blue line) and false positive rate (FP_{rate}, red line) as a function of the probability threshold (σ) logarithm during cross-validation.

compounds in such unique prototypes. We also only considered substitution leading to charge balanced compounds.

Figure 2 indicates the false positive and true positive rates for a given threshold σ. The true positive rate (TP_{rate}) indicates the fraction of existing ICSD compound that are indeed found back by the model (i.e., true hits):

$$TP_{rate}(\sigma) = \frac{TP(\sigma)}{P} \tag{15}$$

Where P is the number of existing compounds considered during our cross-validation test and $TP(\sigma)$ is the number of those existing compounds found by our model with a given threshold σ (i.e., the number of true positives). The false positive rate (FP_{rate}) indicates the fraction of compounds not existing in the ICSD and suggested by the model (i.e., false alarms):

$$FP_{rate}(\sigma) = \frac{FP(\sigma)}{N} \tag{16}$$

Where N is the number of compounds of proposed compounds non-existing in the ICSD but considered during cross-validation and $FP(\sigma)$ is the number of those non-existing compounds proposed by our model with a given threshold σ (i.e., the number of false positives).

High threshold values will lead to fewer false alarms but will imply fewer true hits. On the other hand lower threshold values gives more true hits but at the expense of generating more false alarms. In practice, an adequate threshold is found by compromising between these two situations.

The clear separation between the two curves in Figure 2 shows that the model is indeed predictive and can effectively distinguish between the substitutions leading to an existing compound and those leading to non-existing ones. Moreover, Figure 2 can be used to estimate a value

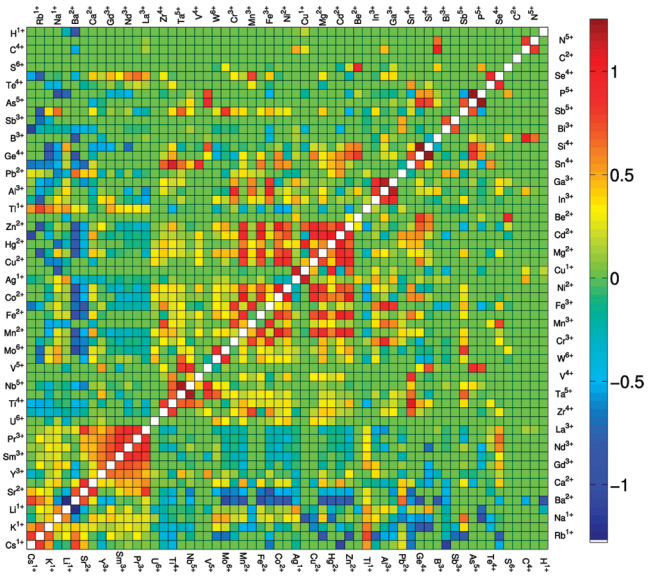

Figure 3. Logarithm (base 10) of the pair correlation g_{ab} for each ion couple a,b. Equation 17 was used to evaluate the pair correlation g_{ab}. The ions are sorted according to their element's Mendeleev number. Only the 60 most common ions in the ICSD are presented in this graph. These correlation coefficients were obtained by training our probabilistic model on the ICSD. Positive values indicate a tendency to substitute while negative values on the contrary show a tendency to not substitute. The symmetry of the pair correlation ($g_{ab} = g_{BA}$) is reflected in the symmetry of the matrix.

of probability threshold for a given true positive rate. For instance, the threshold required to find back 95% of the existing compounds during cross-validation is indicated on the figure by a dashed line.

These cross-validation results can also be used to compare our knowledge based method to a brute force approach in which all charge balanced substitutions from known compounds would be attempted. The brute force approach would require the testing of 884037 compound candidates to recover the full set of known compounds during cross-validation. Using our model, recovering 95% of those known compounds would require testing only 53251 candidates (i.e., only 6% of the number of brute force candidates).

3.2. Ionic Pair Substitution Analysis. The tendency for a pair of ions to substitute for each other can be estimated by computing the pair correlation:

$$g_{ab} = \frac{p(X_1 = a,\ X_1' = b)}{p(X_1 = a)p(X_1 = b)} \tag{17}$$

$$= \frac{p(X_1 = a,\ X_1' = b)}{\sum_j p(X_1 = a,\ X_1' = x_j')\sum_j p(X_1 = b,\ X_1' = x_j')} \tag{18}$$

$$= \frac{\frac{1}{Z}e^{\lambda_1^{a,b}}}{\frac{1}{Z}\sum_j e^{\lambda_1^{a,x_j'}}\frac{1}{Z}\sum_j e^{\lambda_1^{b,x_j'}}} \tag{19}$$

Where a and b are two different ions and the sum represent a summation on all the possible values x_j' of the variable X_1', that is, a sum over all possible ionic species.

This pair correlation measures the increased probability to observe two ions at equivalent positions in a particular crystal structure over the probability to observe each of these ions in nature. Two ions which

substitute well for each other will have a pair correlation higher than one ($g_{ab} > 1$) while ions which rarely substitute will have a pair correlation lower than one ($g_{ab} < 1$). The pair correlation is therefore a useful quantitative measure of the tendency for two ions to substitute for each other.

Figure 3 plots the logarithm (base 10) of this pair correlation for the 60 most common cations in the ICSD (the pair correlation for all the ionic pairs is presented in the Supporting Information). Positive values indicate a tendency to substitute while negative values on the contrary show a tendency not to substitute. The ions are sorted by their element Mendeleev number.[15] This ordering relates to their position in the periodic table. Therefore, the different ions are automatically clustered by chemical classes (alkali, alkali-earth, rare-earth, transition metals and main group elements).

Different "blocks" of strong substitutional tendency are observed. For instance, the rare-earth elements tend to substitute easily to each other. The similar charges (usually +3) and ionic size for those rare-earth elements explain this strong substitution tendency.

The alkali elements form also a strongly substituting group. Only the ions with the largest size difference (Cs with Na or Li) do not substitute easily.

While transition metals in general tend to substitute easily for each other, two subgroups of strong pair correlation can be observed: the early transition metals (Zr^{4+}, Ti^{4+}, Ta^{5+}, Nb^{5+}, V^{4+}, V^{5+}, W^{6+}, Mo^{6+}) and late transition metals (Cr^{3+}, Mn^{2+}, Mn^{3+}, Fe^{2+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Hg^{2+}, Cd^{2+}, Zn^{2+}). This separation in two groups could be explained by a charge effect. The early transition metals have higher common oxidation states (+4 to +6) than the late ones (+2 to +3). Two notable exceptions to the general strong substitution tendency between transition metals are Ag^{1+} and Cu^{1+}. While substituting strongly for each other, those two ions do not substitute for any other transition metal. Indeed, electronic structure factors drive both ions to form very unusual linear environments.[31]

On the other hand, the main group elements do not have a homogeneous strong substitution tendency across the entire chemical class. Only smaller subgroups such as Ga^{3+}, Al^{3+}, and In^{3+} or Si^{4+}, Ge^{4+}, and Sn^{4+} can be observed.

Regions of unfavorable substitutions are also present. Transition metals do not likely substitute for alkali or alkali-earths. Only the smallest ions: Li^{1+}, Na^{1+}, and Ca^{2+} exhibit mild substitution tendencies for some transition metals. In addition, transition metals are very difficult to substitute for rare-earths. Only Y^{3+} (and Sc^{3+} not shown in the figure) can substitute moderately with both rare-earth and transition metals indicating their ambivalent nature at the edge of these two very different chemistries.

Rare-earth compounds do not substitute with main group elements with the surprising exception of Se^{4+}. Se^{4+} can occupy the high coordination sites that rare-earth elements take in the very common *Pnma* perovskite structure formed by $MgSeO_3$, $CoSeO_3$, $ZnSeO_3$, $CrLaO_3$, $InLaO_3$, $MnPrO_3$, and so forth.

(31) Gaudin, E.; Boucher, F.; Evain, M. *J. Solid State Chem.* **2001**, *160*, 212–221.

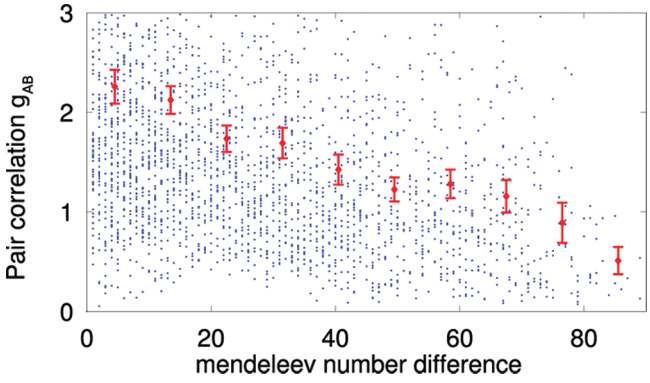

Figure 4. Pair correlation g_{ab} in function of the difference in Mendeleev number between the two ions a and b. Equation 17 was used to evaluate the pair correlation g_{ab}. The blue points are the raw data obtained from fitting the model on the ionic compounds in the ICSD. To distinguish the general trend from the scatter, the data has also been binned in 10 equally sized bins along the Mendeleev number difference axis. Each red point indicates the pair correlation mean for each bin with a 95% confidence interval as error bar. The pair correlation tends to decrease as the Mendeleev number difference increase.

The oxidation state of an element can have a significant impact on whether an element will substitute for others. The two main oxidation states for antimony Sb^{3+} and Sb^{5+} behave very differently. The rather big +3 ion substitutes mainly with $Pb^{2+}$and Bi^{3+}, while the smaller +5 ion substitutes preferentially with transition metals Mo^{6+}, Cr^{3+}, Fe^{3+}, and so forth.

Some ions tend to form very specific structures and local environments. Those ions will substitute only with very few others. For instance, C^{4+} almost only substitutes with B^{3+}. Both ions share a very uncommon tendency to form planar polyanions such as CO_3^{2-} and BO_3^{3-}. Hydrogen is an even more extreme example with no favorable substitution from H^{1+} (with the exception of a mild substitution with Cu^{1+}) to any other ion, in agreement with its very unique nature.

3.3. Chemical and Size Effects over the Substitution Tendencies. The previous analysis shows that strong or weak substitution tendencies can be often rationalized using chemical arguments (i.e., the relative position of the ionic pair in the periodic table). To study this effect, Figure 4 plots the ionic pair correlation defined in eq 17 as function of the difference in Mendeleev number between the two ions. A relation is observed between this difference in Mendeleev number and the pair correlation. Higher pair correlation are associated with smaller differences in Mendeleev numbers. However, this is only true on average and a large spread is observed around the mean values.

Some very interesting outliers can be pointed out. For instance, Cr^{6+} and S^{6+} while significantly distant from each other in the periodic table can easily substitute because of their common tendency to form tetrahedral polyanionic compounds (sulfates and chromates). Ti^{4+} and Sn^{4+} are also two ions with a high pair correlation coefficient despite an important difference in Mendeleev number. Conversely, $Rh^{3+}$and Co^{3+}, while in the same column of the periodic table, do not substitute strongly ($g_{ab} = 0.98$).

In addition to chemical effects, size effects are also very often used to estimate how likely an ionic substitution is.

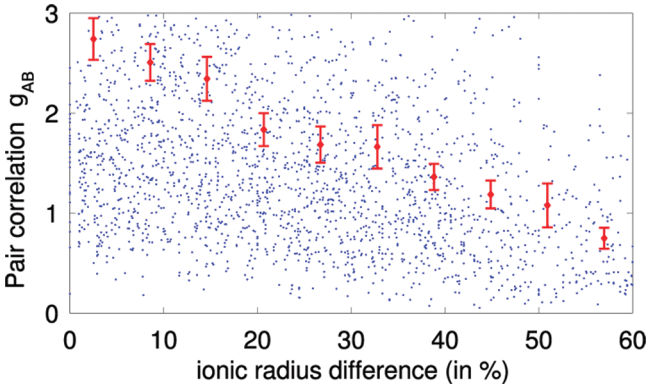

Figure 5. Pair correlation g_{ab} in function of the difference in ionic size between the two ions a and b. Equation 17 was used to evaluate the pair correlation g_{ab}. The ionic size difference is computed as the difference in ionic size divided by the size of the largest of the two ions. This gives relative ionic radius differences. The ionic size for the two ions are obtained from the Shannon radii table and for the coordination 6 have been used.[32] The blue points are the raw data obtained from fitting the model on the ionic compounds in the ICSD. To distinguish the general trend from the scatter, the data has also been binned in 10 equally sized bins along the Mendeleev number difference axis. Each red point indicates the pair correlation mean for each bin with a 95% confidence interval as error bar. The pair correlation tends to decrease as the difference in ionic size increase.

Ions of similar size tend to be considered easier to substitute for each other. In Figure 5, the pair correlation g_{ab} is plotted as a function of the difference in ionic size between the two ions. The ionic size used is the 6-fold coordinated size according to Shannon.[32] A clear relation between the two quantities can be observed. The highest pair correlations tend to be found for smaller differences in ionic size. As for the chemical effects, there is an important spread around the general trend. Again, S^{6+} and Cr^{6+} do not follow the general trend. Those two highly substitutable ions (g_{ab} = 9.7) have a 50% difference in their ionic size.

Au^{1+} and Cu^{1+} while very different in ionic size (1.37 A and 0.77A) show an important correlation number (g_{ab} = 5.0). On the other hand, an ion very close in radius such as Li^{1+} (0.76A) does not substitute easily to Cu^{1+} (g_{ab} = 0.9). The tendency for Au^{1+} and Cu^{1+} to form the peculiar linear environments wins over their significant size difference. Another case in point is the pair Hg^{2+}-Na^{1+}. Those ions have the same size according to the Shannon radii table but do not substitute (g_{ab} = 0.19).

3.4. Online Ionic Substitution Model. The ionic substitution model is available online at http://www.materialsgenome.org/substitutionpredictor. Any user can query the model for four ionic species predictions. An e-mail with the proposed substitutions and the crystal structures of the predicted compounds in the crystallographic information file (cif) format will be sent to the user after computations.

4. Discussion

We presented a machine learned ionic substitution model trained on experimental data. This model can be used with significant predictive power to discover new compounds and their crystal structure.

Our model makes several simplifying assumptions. The absence of dependence with the number of components implies that, for instance, the substitution rules do not change if the compounds are ternaries or quaternaries. If Fe^{2+} is established to substitute easily for Ni^{2+} in ternary compounds, the same substitution should be likely in quaternaries.

In addition, the substitution rules do not depend on structural factors. However, how easy a chemical substitution is will depend somewhat on the specific structure. Some crystal structure sites will accommodate for instance a wider range of ions with different size without major distortion. Perovksites are a good example of structures where the specific size tolerance factor is established (see for instance Zhang et al.[33]). In some sense, our model is "coarse grained" over structures.

The second major assumption is the use of binary features only. This implies that the substitution model only focuses on two substituted ions at a given site and does not take into account the "context" such as the other elements present in the crystal structure. Here again, a more accurate description will require to take this context into account. For instance, two cations might substitute in oxides but not in sulfides.

Those simplifying assumptions are however very useful in the sense that they allow the model to capture rules from data dense regions and use them to make predictions in data sparse regions. The substitution rules learned from ternary chemical systems can be used to predict compounds in the much less populated quaternary space. Likewise, substitution rules learned from very common crystal structure prototypes can be learned and used to make predictions in uncommon crystal structures. It is this capacity for this simpler model to make predictions in sparser data regions which constitutes its main advantage versus more powerful models such as the one presented in Fischer et al.[19]

Of course, our model could be refined in many ways. The most straightforward way to add structural factors would be to introduce a dependence on the ion local environment. The features could also be extended to go beyond binary features. Interesting work in feature selection has shown that complex features can be built iteratively from the data by combining very simple basic features.[25]

A limitation of this model lies in its inability to predict totally new crystal structures. Indeed, any new compound will be proposed by ionic substitutions from a compound with an already known crystal structure. This usual limitation to crystal structure prediction methods based on data mining is, however, compensated by their much smaller computational requirements than a more exhaustive search based on optimization such as with a genetic algorithm.

We must stress that this substitution model does not prejudge any atomic factor such as charge, size, electronegativity, or position in the periodic table to be important in determining crystal structure. While correlation with some of those parameters is definitely reproduced by the model, the purely data-driven formulation of the problem automatically weights those factors without having had to make a priori decisions on their role. Moreover, the model takes into account the potential substitution outliers that do not follow simple rules based on those atomic factors.

Finally, while experts trained in solid state chemistry can readily qualitatively assess the likelihood of ionic substitution, we must stress that our model is able to perform this task

(32) Shannon, R. D. *Acta Crystallogr., Sect. A* **1976**, *32*, 751–767.

(33) Zhang, H.; Li, N.; Li, K.; Xue, D. *Acta Crystallogr., Sect. B* **2007**, *63*, 812–818.

quantitatively and without human supervision. The possibility to perform such an automatic search on large amounts of data is critical to the high-throughput computational search for new materials.[34−36]

5. Conclusion

We proposed a probabilistic model that predicts ionic substitutions which keep the crystal structure of a compound unchanged. We showed how such a model can be used to predict new compounds and their crystal structures. The model's predictive power was demonstrated using cross-validation on the ICSD quaternary compounds.

While the substitution model captures factors (e.g., ionic size and position in the periodic table) already used by solid state chemists for suggesting new compounds, its purely data-driven nature allows to weight all those factors and others (e.g., electronic structure driven factors) in one single quantitative model.

We believe such a tool will be very useful to the large scale computational search of new inorganic compounds.

Acknowledgment. This work was supported by the NSF (under Contract No. DMR-0606276) and by the Department of Energy, Office of Basic Energy Sciences (under Contract No. DE-FG02-96ER4557). Virginie Ehrlacher thanks the Ecole des Ponts Paristech and the CERMICS (Champs-sur-Marne, France) for funding. Anubhav Jain acknowledges funding from the U.S. Department of Energy, Department of Energy Computational Science Graduate Fellowship (DOE CSGF) (under Grant DE-FG02-97ER25308).

Supporting Information Available: Pair correlation for all ionic pairs. This material is available free of charge via the Internet at http://pubs.acs.org.

(34) Levy, O.; Chepulskii, R. V.; Hart, G. L. W.; Curtarolo, S. *J. Am. Chem. Soc.* **2009**, No. 2, 833–837.
(35) Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I. B.; Nørskov, J. K. *Nat. Mater.* **2006**, *5*, 909–913.
(36) Hummelshøj, J. S.; et al. *J. Chem. Phys.* **2009**, *131*, 014101.