Top Curr Chem (2014) 345: 139–180
DOI: 10.1007/128_2013_486

Published online: 28 November 2013

Data Mining Approaches to High-Throughput Crystal Structure and Compound Prediction

Geoffroy Hautier

Abstract Predicting unknown inorganic compounds and their crystal structure is a critical step of high-throughput computational materials design and discovery. One way to achieve efficient compound prediction is to use data mining or machine learning methods. In this chapter we present a few algorithms for data mining compound prediction and their applications to different materials discovery problems. In particular, the patterns or correlations governing phase stability for experimental or computational inorganic compound databases are statistically learned and used to build probabilistic or regression models to identify novel compounds and their crystal structures. The stability of those compound candidates is then assessed using ab initio techniques. Finally, we report a few cases where data mining driven computational predictions were experimentally confirmed through inorganic synthesis.

Keywords Ab initio computations · Crystal structure prediction · Data mining · High-throughput computing

Contents

G. Hautier (✉)
Université Catholique de Louvain, Institute of Condensed Matter and Nanosciences (IMCN), Chemin des étoiles 8, bte L7.03.01, 1348 Louvain-la-Neuve, Belgium
e-mail: geoffroy.hautier@uclouvain.be

1 Introduction

First principles or ab initio computations aim at computing materials properties (e.g., thermodynamic stability, conductivity, light absorbance) from the fundamental laws of quantum physics. Following the emergence of ab initio techniques and especially of density functional theory (DFT) [1], the field has seen a combination of theoretical developments, standard codes developments (e.g., [2–4]), and increase in computational power. Materials science is even moving to a new paradigm where computations are not only used to explain experimental observations but also to *predict* new materials and their properties [5].

One emerging route towards computational discovery of materials is to use high-throughput computing. High-throughput computing consists of evaluating material properties on thousands of different materials to identify the best performing compounds and to understand trends from large datasets [5, 6]. This approach has already been used in various fields such as catalysis [7], Li-ion batteries [8, 9], scintillators [10], photocatalytic water splitters [11–13], thermoelectric materials [14, 15], mercury sorbents [16], organic photovoltaics [17], and topological insulators [18]. High-throughput infrastructures have reached such a maturity that large sets of computations are nowadays stored in computational databases such as

the materials project [19, 20] and others [21, 22] that can be accessed through web interfaces. With the data repository and analysis tools they provide, materials scientists now have access to an unprecedented amount of data [23].

Many high-throughput studies have concentrated on evaluating properties on known compounds extracted from databases such as the Inorganic Crystal Structure Database (ICSD) [24] or on a limited structural framework (e.g., perovskites [11]). While those studies are of great value, they face some limitations. Databases are often not up to date, i.e., they do not have the latest reported structures in the literature. Also, many inorganic compounds are known to exist at a given stoichiometry but their crystal structure has not been determined from powder diffraction data. Finally, compounds of greatest interest for a specific application might not have been synthesized yet. This is especially the case for multicomponent systems (e.g., ternaries and quaternaries) or less common chemistries.

Finding new compounds and determining their crystal structure before synthesis is called crystal structure prediction. Since 1988, when Nature's editor John Maddox called our inability to properly perform crystal structure prediction one of "continuous scandal in physical science," the field has greatly evolved [25–27]. Among the different approaches to structure prediction, data mining has been developed in parallel with high-throughput computational searching. Indeed, in contrast to other approaches, data mining typically compromises on the exhaustivity of the search in favor of less computational time and an access to much larger chemical spaces to explore. The idea behind data mined compound prediction is very simple and has been driving solid state chemistry for centuries: nature is not random and there are patterns that one could learn from observing phase stability. The novelty lies in the use of quantitative mathematical approaches from the fields of machine learning or statistical learning.

In this chapter we will start by presenting how thermodynamical phase stability can be evaluated from DFT computations (Sect. 2). The different techniques and accuracy of approximations will be outlined. The general idea behind data mining driven structure prediction will be presented in Sect. 3 and specific examples of methods and algorithms will be explained in detail in Sects. 4, 5, and 6. Finally, we will present in Sect. 7 a few selected examples of successful data mining compound predictions where the computational suggestion was followed by successful experimental verification.

2 Phase Stability Evaluation Through Ab Initio Computing

An important factor determining the existence of inorganic compounds is their thermodynamical phase stability. To evaluate whether a compound is thermodynamically stable, one needs to compare its (free) energy with the (free) energy of other competing phases. This step is essential for the compound prediction problem and DFT computations are routinely used to perform such an analysis. In this section we will overview the standard thermodynamic constructions along with the different approximations involved and assess their accuracy.

2.1 Low Temperature Stability: The Convex Hull Construction

Assessing thermodynamical phase stability in a chemical system requires the comparison of the free energy of the different phases present [28, 29]. An isothermal, isobaric and closed system requires the use of the Gibbs free energy as thermodynamic potential. For a binary component system with N_A atoms of A and N_B atoms of B, at temperature T and pressure p, the Gibbs free energy G is expressed as

$$G(N_A, N_B, T, p) = E(N_A, N_B, T, p) + pV(N_A, N_B, T, p) - TS(N_A, N_B, T, p), \quad (1)$$

where V is the volume, S the entropy, and E the energy.

The first approximation we will make is to assume that the pV term is small. This approximation is valid when only solid phases are involved in the phase equilibrium. In addition, we will work at zero temperature. No entropic effects need to be taken into account then. Entropic effects can be modeled but this would require a more important computational budget as all relevant excitations (vibrational, configurational, and electronic) would need to be considered [30–32].

Under these approximations, the relevant thermodynamic potential is the energy. The energy normalized by the total number of particles in the system ($N = N_A + N_B$) : $\overline{E}(x_A, x_B)$ and fractions instead of amounts: $x_A = \frac{N_A}{N}$ and $x_B = \frac{N_B}{N}$ will be used. The normalized energy is usually expressed in meV/atom.

Solving the Kohn–Sham equation in the DFT framework can directly provide an approximation to this energy. Ab initio computations can therefore associate an energy to any compound present in a given chemical system. In the specific case of zero temperature and negligible volume effects, phase stability can then be directly computed from a simple set of DFT ionic relaxations on all the phases of interest. Let us illustrate this with the example of a simple binary A-B chemical system. In this system, computations have been performed for compounds at a composition A_2B, AB_2, and AB in different crystal structures designated respectively by α_1, α_2, β_1, β_2, β_3, and γ. The elemental phases have also been computed and, as a convention, all energies will be expressed as formation energies from the elements. Figure 1 plots the formation energy for the different phases computed in function of the fraction of B. From this plot, a very simple construction called the *convex hull* can be performed. The construction consists of finding a convex envelope containing all the points in the plot. This envelop called the convex hull (or hull) is plotted in green in Fig. 1. The phases present on this convex hull are the most stable phases or *ground states* for the system studied. For instance, α_2 is thermodynamically unstable and will decompose to form α_1. The phase γ will decompose into two phases: α_1 and β_2 (as γ is above the tie line formed by α_1 and β_2).

This construction can be performed in any dimension and thus on multi-component systems such as ternaries, quaternaries, etc.

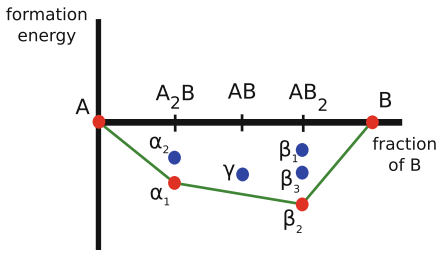

Fig. 1 Convex hull construction for an A-B system. The *points* represent different phases. The *line* is the convex hull. The *points on the line* are the most stable phases or ground states and *points above the line* are unstable phases according to the construction

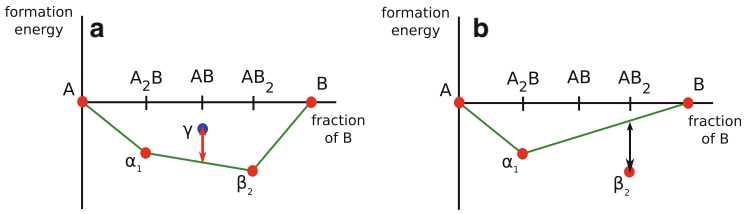

Fig. 2 Illustration of different measure of stability from the convex hull construction. The energy above the hull is illustrated for the unstable phase γ by the *double arrow* in (**a**). The inverse distance to the hull is represented for the stable phase β_2 by the *double arrow* in (**b**). Reprinted with permission from [33]. Copyright 2012 American Chemical Society

Different measures of (in)stability can be defined using this convex hull construction:

- *Energy above the hull* (*or distance to the hull*)
  For an unstable phase, the energy above the hull consists in the energy separating the phase from its decomposition tie-line (see red double arrow in Fig. 2a). It is equivalent to the opposite of the energy associated with the decomposition reaction from the phase to the stable products. It is a positive number and usually expressed in meV/atom. Stable phases have by definition an energy above the hull equals to zero.
- *Inverse energy above the hull* (*or inverse distance to the hull*)
  This quantity is defined only for stable phases. It is computed by removing the phase of interest from the convex hull and constructing a new convex hull. The distance to the new convex hull for the phase is then computed and called the inverse energy above the hull. It is equivalent to the opposite of the energy of formation of the phase of interest from the phases that would be stable if it did not exist. It is a positive number and expressed in meV/atom. A large inverse distance to hull represents a high stability of the predicted structure. The inverse energy above the hull is represented for the phase β_2 in Fig. 2b.

Convex hull constructions and the analysis of computed phase diagrams can be performed using the pymatgen package [34].

2.2 Stability for Open Systems

Oxides are very important compounds technologically and are better studied with an open instead of close thermodynamical system approach. A ternary system composed of particles of A, B, and oxygen will be used here as an example. In the previous section we assumed that the relevant thermodynamic variables are the amount of constituents (N_A, N_B, and N_O), the temperature T, and the pressure p. In reality, very often during oxide synthesis, the amount of oxygen present in the system is not directly controlled and the system is an open system to oxygen. In this case, the relevant thermodynamic potential is the Legendre transform of the Gibbs free energy with respect to the oxygen amount: the oxygen grand potential φ:

$$\varphi(N_A, N_B, \mu_O, T, p) = G - \mu_O N_O. \tag{2}$$

Normalizing the grand canonical potential by $N = N_A + N_B$ and using factions of A, B, x_A, and x_B, we get

$$\overline{\varphi}(N_A, N_B, \mu_O, T, p) = \frac{G - \mu_O N_O}{N}. \tag{3}$$

This is a situation very similar to that in the previous section except that the Gibbs free energy is replaced by the oxygen grand potential. Here, the effect of volume and temperature can be approximated by assuming that the dominant volume and entropy factors come from the gaseous oxygen and that the entropy and volume factors from the solid phase can be neglected. This approximation has been successfully used by Ong et al. for the study of the Li-Fe-P-O phase diagram [35]. The normalized grand canonical potential is then

$$\overline{\varphi}(N_A, N_B, \mu_O, T, p) = \frac{E - \mu_O N_O}{N}. \tag{4}$$

Only the μ_O term has a pressure and temperature dependence. Practically, a convex hull construction using the normalized grand canonical potential at a fixed μ_O can be performed to obtain the stable phases in specific conditions. The oxygen chemical potential can be linked to the oxidizing or reducing nature of the environment. Ways to increase the oxygen chemical potential (i.e., to be more oxidizing) are to decrease the temperature or increase the oxygen partial pressure. In contrast, the oxygen chemical potential can be decreased (i.e., be more reducing) by increasing the temperature or lowering the oxygen partial pressure.

It follows from this analysis that any oxide compound exists in an oxygen potential window with a maximal and minimal oxygen chemical potential. Any environment setting a chemical potential lower than the minimal oxygen chemical potential would be too reducing for the compound to form while any environment setting a higher chemical potential than the maximal oxygen chemical potential would be too oxidizing.

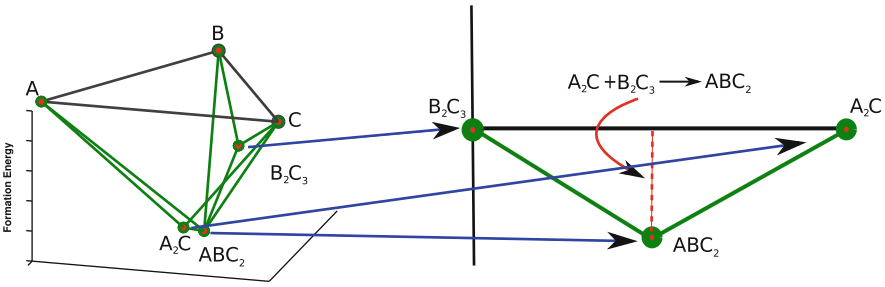

Fig. 3 Convex hull of a typical A-B-C system. The stability of the ternary ABC_2 phase will depend directly on the reaction energies from the binaries, not from the elements

2.3 Accuracy of DFT(+U) in Determining Phase Stability

Curtarolo et al. performed one of the first large scale studies of the performance of DFT on phase stability [36]. The authors focused on binary metals. They computed a large number of competing crystal structure prototypes in 80 binary metal systems and they studied how often the experimentally observed ground state was in agreement with the computed one. DFT successfully found the actual ground state in at least 90% of the cases.

For oxides and other insulators or semiconductors, the typical errors from standard DFT in oxides on the elemental formation energies can be quite large and up to hundreds of meV/atom [37]. However, for multicomponent compounds, phase stability will not depend directly on the elemental formation energy but more often on the reaction energies between multicomponent phases. Figure 3 illustrates this by presenting the convex hull of an A-B-C system. The stability of the ABC_2 phase does not depend directly on the $A + B + 2C \rightarrow ABC_2$ reaction (i.e., the formation energy from the elements) but will depend on the $A_2C + B_2C_3 \rightarrow ABC_2$ reaction (dashed red line). For instance, determining whether a ternary oxide is stable or not will depend on its reaction energy from the binary oxides. A recent study showed that those reaction energies are significantly better described by DFT than by elemental reaction energies due to cancellation of errors when comparing chemically similar phases [38]. Comparing computed to experimental reaction energies, an error distribution centered on 0 and with a standard deviation around 25 meV/atom was found. When analyzing compound prediction results, this error bar should be kept in mind.

For metal oxides with partially occupied *d* orbitals (i.e., FeO, Mn_3O_4, etc.), DFT is known to perform poorly because of a self-interaction error present in the typical functionals used in DFT. The DFT+U method is one way of circumventing this issue by effectively localizing *d* electrons and providing a more physically accurate picture of the bonding in oxides [39, 40]. On the other hand, in metals the electron delocalization induced by pure DFT is actually close to the real metallic bonding state and applying a U correction would only cause the model to deviate from the reality. We stand therefore in a situation in which, for transition metals, DFT reproduces

sufficiently well the energy in metallic systems but in oxides, only DFT+U does. As computations with two different Hamiltonians (DFT and DFT+U) cannot be directly compared, it is impossible to compute energies and then evaluate phase stability when compounds of different natures are involved, such as oxides and metals. To treat this situation, Jain et al. developed an approach relying on an energy shift of the DFT energies [41]. This shift is based on a calibration on experimental binary oxides formation energies from the metal. After applying this shift to DFT computed phases, all computed data can be compared and used to assess phase stability. A similar approach has been proposed by Stevanovic et al. [42].

3 Data Mining Compound and Crystal Structure Prediction

Section 2 showed how the phase stability of compounds is assessed using DFT. However, the most challenging part of the compound prediction problem lies in the efficient selection of compound candidates to test for stability. Nowadays this selection is typically performed following one of two approaches: optimization or data mining-based.

3.1 Optimization Approaches

Optimization-based methods consider that finding the most stable crystal structure (at a given composition) can be mapped to the mathematical problem of finding the values of the structural degrees of freedom (i.e., lattice parameters and atomic positions) minimizing the (free) energy. The search for a global minimum on the energy landscape is, however, far from simple as the energy function (or landscape) is very large, complex, and presents many local minima [43].

One popular way of simplifying this problem has been to reduce the number of degrees of freedom by working on a fixed crystal lattice, only allowing different decorations of the underlying crystal structure framework. For instance, we can study any ordering on a face-centered cubic lattice at a composition AB and find possibly a rock salt ground state. This approach is usually coupled with the use of a simplified Hamiltonian fitted on a limited set of computations performed on selected orderings through the cluster expansion technique [44–46]. Identifying new phases on a fixed lattice has been especially useful in alloy theory [47–49], but close-packed oxides have also been studied through cluster expansion [50].

However, when the underlying lattice is not known, researchers must rely on advanced optimization techniques such as simulated annealing or genetic algorithms to explore the rugged energy landscape. Simulated annealing (and the related basin hopping) [51, 52] rely on applying perturbations to a starting configuration. Those perturbation are accepted or not depending on how the energy is

lowered, offering a way to scan the energy landscape efficiently in search of a global minimum. Genetic algorithms, on the other hand, are inspired by the biological process of evolution and the idea of survival of the fittest [53–57].

Optimization methods have been used to study many different chemistries, often with empirical potentials. However, a growing number of studies are now being performed purely on first-principles computations (e.g., the Na-N [58], W-N [59], Fe-B [60] chemical systems). New phases proposed by optimization approaches include new high-pressure phases of boron [59], $CaCO_3$ [55, 61], and FeB_4 [62] as well as a new metastable polymorph of LiBr [63]. The optimization approach to structure prediction is very appealing but suffers from very extensive requirements in terms of computational budget, especially when multicomponent systems are explored. For instance, finding the ground state of $MgSiO_3$ by a genetic algorithm required around 1,000 energy evaluations [56].

3.2 Data Mining Approaches

The optimization approach assumes no previous knowledge (except for the energy model). On the other hand, solid state chemists have been for long using empirical or heuristic rules to rationalize and sometimes predict crystal structures. A very well known example of such a set of rules is the Pauling rules relating stability to atomic factors (such as ionic size, charge) and structural factors (such as the number of edges or facets shared by cation-anion polyhedra) [64].

Another common heuristic approach consists of building structure maps [65–67]. Structure maps rely on the existence of common *crystal structure prototypes*. Different compounds can form similar arrangement of atoms called prototypes. Traditionally, these structure prototypes are named after the formula and/or name of the mineral from one of the compounds forming this structure. For example, the "NaCl" or "rocksalt" structure prototype is formed not only by NaCl but also by CoO, AgBr. etc. (see Fig. 4).

Structure maps are constructed by plotting for what values of atomic factors certain crystal structure prototypes form. These atomic factors can, for instance, be ionic radii or chemical scales such as the Mendeleev number in Pettifor maps. If the factors are relevant, the structure types will cluster in different regions of the structure map.

Empirical rules such as the Pauling rules are not really predictive and are mainly used to rationalize the existence of already characterized crystals. While structure maps can be used as a predictive tool as shown by Morgan et al. [68], they present limitations due to their focus on specific factors such as size or electronegativity and tend to be available only for very well populated stoichiometries.

Inspired by the success of empirical rules, researchers have been developing data mining or machine learning techniques that learn from previous computations or experiments and make informed guesses about likely crystal structure candidates [69]. The approach relies greatly on the recent developments in data mining,

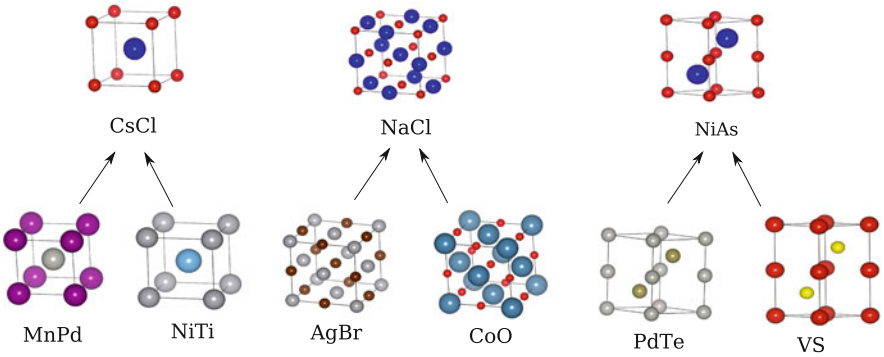

Fig. 4 Some examples of compounds and their crystal structure prototypes

machine learning, and statistical learning [70]. While we will focus on inorganic compounds in this chapter, data mining approaches are also used more and more in the fields of organic chemistry (see for instance [71, 72]).

Sections 4, 5, and 6 will present in more detail some data mining approaches to compound prediction. They all rely on the use of a database of experimental or computed data that is used to fit a probabilistic or regression model. This data mined model can propose likely compound and crystal structure candidates that are tested for stability with DFT.

4 Linear Regression Based Approaches to Data Mining Crystal Structure Prediction

The work from Curtarolo et al. pioneered the use of data mining approaches in combination with ab initio computations [73]. The authors focused on the correlations existing between the energy of crystal structure prototypes in a binary system.

4.1 The Principal Component Analysis Model

Curtarolo et al. built a database of 114 crystal structure prototypes in 55 binary metallic systems. They computed the energy of each of those compounds using DFT.

The information included in this database can be expressed as a series of 55 vectors E_i (1 for each binary system) with 114 dimensions:

$$E_i = (E_{i1}, E_{i2}, \ldots, E_{in}) \tag{5}$$

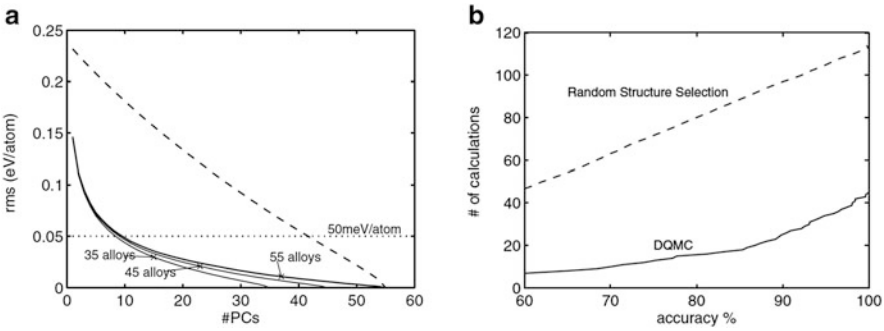

Fig. 5 Root mean squared error in function of the dimension reduction (**a**) and (**b**) number of computations as a function of the number of ground states accurately predicted. The *dashed line* indicates picking the structures randomly and the *plain lines* indicate the data mining driven approach. Reprinted figures with permission from [73]. Copyright 2003 by the American Physical Society

If the energies are not distributed randomly in the 114 dimension (i.e., if there are correlations between energies in different alloys and crystal structures), we can represent the energy vectors in a subspace of lower dimension than the full 114 dimensions. This dimension reduction can be formally performed with the commonly used principal component analysis (PCA).

PCA starts by expressing the vector E_i as an expansion on a subspace of smaller dimension:

$$E_i = \sum_{j=1}^{d} \alpha_{ij} e_i + \varepsilon_i(d), \tag{6}$$

where ε_i is the error on the alloy i. PCA then finds the basis set $\{e_i\}$ minimizing the squared sum of errors $\sum_i \varepsilon_i^T \varepsilon_i$. This new basis set consists of a new set of axes in the 114 dimension space that are adequate to represent our set of alloy energies in reduced dimensions.

Reducing the dimension naturally induces an error compared to the full database in the 114 dimensions. The smaller the dimension reduction (the larger d), the smaller the error induced by dimensional reduction. This is illustrated in Fig. 5a which shows the root mean squared error depending on the number of dimensions. Only nine dimensions (nine alloys) are necessary to obtain the energy of an alloy in a specific crystal structure within an error of 50 meV/atom.

4.2 Prediction Procedure

The correlations indicated by the PCA can be used to accelerate the prediction of new phases. Using these correlations, the amount of ab initio computations to perform can be reduced dramatically. A data mining driven structure prediction

procedure consists of three stages: *prediction*, *suggestion*, *calculations*. Given a previously computed library of crystal structure prototypes in different alloys, we can use the PCA to predict the energies of crystal structures not computed yet in a given alloy. Using these data mined predicted energies we can identify the structures that are the farthest below the convex hull or the closest to the hull. This limited set of candidates are then computed by DFT. The new DFT results are added to the database and a new series of prediction, suggestion, calculations is performed until a convergence to a stable solution is reached.

Figure 5b compares the number of calculations required to reach a certain percentage of ground states accurately predicted in both the random selection (dashed line) and data mining driven case (plain line). The data mining approach performs significantly better.

This technique has been used to perform searches of new borides [74, 75] or rhodium alloys [76].

5 Data Mining Approach Based on Correlations Between Crystal Structure Prototypes

The approach based on PCA presented in Sect. 4 is of great interest but requires a database of computed energies for known (often stable) compounds and for hypothetical compounds (often unstable) and their crystal structures. Such a database is unfortunately not available for most areas of chemistry. On the other hand, experimental crystal structure databases such as the ICSD are widely available, giving access to observed inorganic compounds. In 2006, Fischer et al. proposed an approach based on correlations between observed crystal structures that do not require any previous computational data [77]. Instead of a regression problem (i.e., predicting continuous quantities such as energies), a classification problem is tackled: predicting whether a given crystal structure is likely to be stable or not (without modeling how stable it will be). We will present here the algorithm in detail and its application on a high-throughput large scale search for ternary oxides [78].

5.1 General Principle of the Algorithm

Crystalline inorganic compounds have a limited set of crystal structure prototypes (see Fig. 4). The basic idea behind the algorithm is to consider that the presence of a given crystal structure prototype in a chemical system can be correlated to factors such as the elements in this chemical system and the crystal structures co-existing at other compositions. For instance, the crystal structure prototype of $LaMn_2O_5$ forms very often with Mn. A strong correlation exists between the presence of this crystal

structure prototype in a chemical system and manganese. Likewise, the $FeSb_2O_6$ and Sb_2O_5 crystal structure prototypes are also strongly correlated. From this observation one can think about using partial information about a chemical system (e.g., the presence of Mn or of the Sb_2O_5 prototype) to infer the crystal structures likely to form. In the following sections we will discuss how this basic idea is implemented mathematically. The data abstraction and variables will be introduced along with the probabilistic model rigorously integrating all those correlations.

5.2 Data Abstraction

We will assume that a prototype label has been assigned to all the compounds in the database. This prototyping step can be fully automated by using, for instance, the algorithm proposed by Hundt et al. [79]. After transformation of the raw database to a prototyped database, the data are in the form of a composition-crystal structure prototype pair for each compound.

For the sake of simplicity we will use discrete composition variables in our model. Compositions are continuous variables and, to project this continuous problem to a discrete one, we will consider any composition to be present in a composition bin. For instance, the composition bins could be AB, A_2B, AC_2, etc. for the binaries and ABC, ABC_2, etc. for the ternaries. Each of these composition bins c_i is associated with a variable x_{c_i} indicating what crystal structure is present at this composition. For example, if c_i represents the composition AB_2C_4 then x_{c_i} may have values such as *spinel*, *olivine*, etc. The condition $x_{c_i} = \text{no structure}$ value indicates the absence of a compound at the given composition. In addition, variables representing the system's constituents (e.g., E_i=Ag, Cu, Na, etc.) are defined. With these definitions, any chemical system of C constituents and n compositions can be represented by a vector $X = (x_{c_1}, x_{c_2}, \ldots, x_{c_n}, x_{E_1}, x_{E_2}, \ldots, x_{E_c})$ where the composition space is discretized by using n composition bins.

In this formalism, any information from the database on a chemical system can be represented by an instance of the vector $\mathbf{X}$ (see Fig. 6). Any prototyped crystal structure database D can then be represented as a collection of N $\mathbf{X}_i$ instances, $\mathsf{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$.

5.3 Probabilistic Function and New Compound Discovery Procedure

The probability density $p(\mathbf{X})$ provides information as to what crystal structures tend to coexist in a chemical system. Based on the available information at known compositions in a system, this probability density can be used to assess if another composition (c_j) is likely to be compound-forming. Mathematically, this is evaluated by computing the probability of forming a compound:

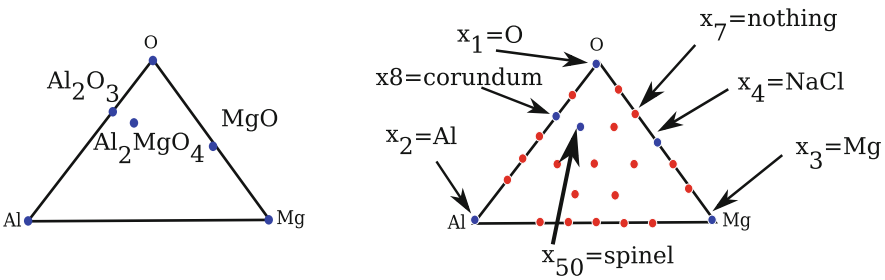

Fig. 6 An example of how the information on the Al-Mg-O chemical system is projected onto the composition variables. All *dots* indicate composition bins. *Red dots* are composition bins without any known compound and *blue dots* are composition bins with a known compound crystallizing in a specific prototype marked by an *arrow*

$$p_{\text{compound}}(c_j) = 1 - p\left(x_{c_j} = \text{nostructure} \middle| x_{c_1}, x_{c_2}, \ldots, x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, \ldots, x_{E_1}, x_{E_2}, \ldots, x_{E_c}\right). \tag{7}$$

In addition, when a composition c_j of interest is targeted, the probability density can be used to suggest the most likely crystal structures by evaluating the following:

$$p\left(x_{c_j} \middle| x_{c_1}, x_{c_2}, \ldots, x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, \ldots, x_{E_1}, x_{E_2}, \ldots, x_{E_c}\right). \tag{8}$$

For the different values of x_{c_j} (i.e., for the different crystal structure prototypes known at this composition), a list of the l most likely crystal structure candidates can be established. These candidate crystal structures can then be tested for stability by an accurate energy model such as DFT. The procedure for compound discovery is summarized in Fig. 7.

We should stress that, in contrast to most optimization techniques, this approach can not only suggest likely crystal structures for a given composition but also suggest which compositions are likely to form stable compounds. This is very important, especially for multi-component systems (ternaries or quaternaries), as the compositional space is larger than for binary compounds.

5.4 Approximated Probabilistic Function

While very useful for structure prediction, this probability function is extremely complex. In the case of ternary oxides, our model requires 183 variables. With roughly 100 crystal structure prototypes possible per variable, this probability function is defined on a domain of around 10^{366} values!

For all practical purpose this probability function needs to be approximated. The way the approximation is made here is to use an approach known in statistical

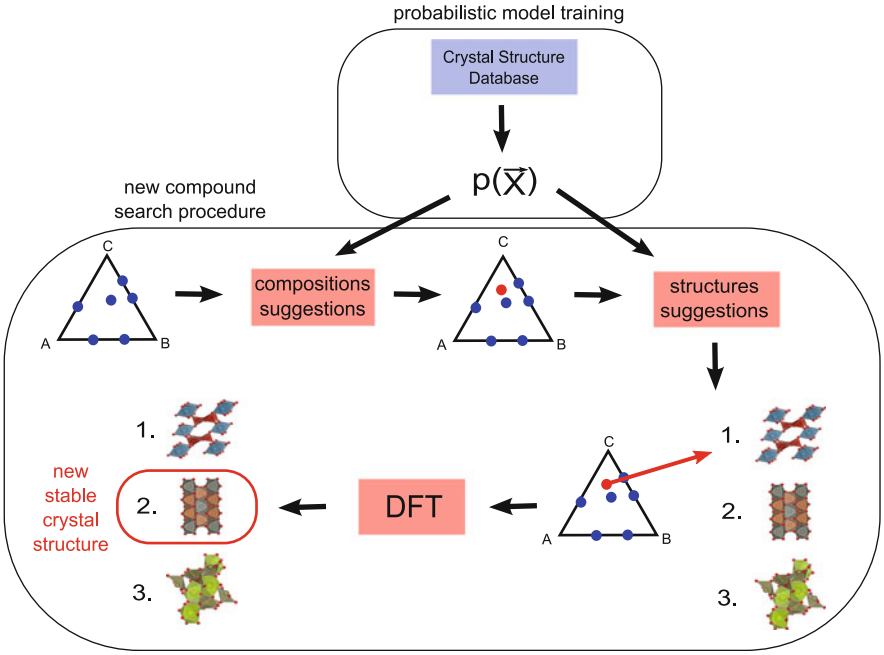

Fig. 7 Data-mining driven compound discovery procedure. A probabilistic model is built from a crystal structure database. In any system A-B-C, this model is used to identify the new compositions (*red dots*) most likely to form a compound. For those compositions, the most likely crystal structures are proposed using the same probabilistic model. These structure candidates are then tested for stability by an accurate energy model as DFT

mechanics as the cumulant expansion [80]. The cumulant expansion can be presented starting with the identity

$$p(\mathbf{X}) = \prod_i g_i(x_{c_i}) \prod_{j<k} g_{jk}\left(x_{c_j}, x_{c_k}\right) \prod_{l<m<n} g_{lmn}(x_{c_l}, x_{c_m}, x_{c_n}) \ldots \tag{9}$$

Following this expression, $p(\mathbf{X})$ can be seen as a product of independent variables with corrections from pair, triplet, etc., correlations. The cumulant terms can be defined recursively. Starting with a one variable probability function, we trivially have

$$g_i(x_{c_i}) = p(x_{c_i}); \tag{10}$$

with a two variables probability function we have

$$p\left(x_{c_i}, x_{c_j}\right) = p(x_{c_i}) p\left(x_{c_j}\right) g_{ij}\left(x_{c_i}, x_{c_j}\right), \tag{11}$$

which implies that

$$g_{ij}\left(x_{c_i}, x_{c_j}\right) = \frac{p\left(x_{c_i}, x_{c_j}\right)}{p(x_{c_i})p\left(x_{c_j}\right)}. \tag{12}$$

The general form for a cumulant over the variable X_α is

$$g_\alpha(x_\alpha) = \frac{p(x_\alpha)}{\Pi_{\beta\subset\alpha} g_\beta\left(x_\beta\right)}, \tag{13}$$

for which the products at the denominator extends over all subsets of α.

So far, no approximation has been introduced. The approximation will consist of truncating the cumulant expansion, considering that all the cumulants beyond pairs (triplets, quadruplets etc. . .) are equal to 1 so that

$$p(\mathbf{X}) = \frac{1}{Z} \prod_i p(x_{c_i}) \prod_{j<k} \frac{p\left(x_{c_i}, x_{c_j}\right)}{p(x_{c_i})p\left(x_{c_j}\right)}, \tag{14}$$

where Z is a normalization constant or partition function.

5.5 *Estimating the Probabilistic Function from Available Data*

Having decided on the form of an approximated probability function (14), we still need to estimate the values of these function parameters. Using a database D, we will search for the values $p\left(x_{c_i}, x_{c_j}\middle|\mathsf{D}\right)$ and $p\left(x_{c_i}\middle|\mathsf{D}\right)$ in best agreement with the data. One can see this process – called parameter estimation – as a fit of the model to the available data.

We will present two common ways of estimating the parameters of a probabilistic model from the data: the maximum likelihood and the Bayesian approach. For pedagogical purposes we will first present derivations for the single variable case and will generalize later on the multi-variable case [81].

5.5.1 Single Variable Multinomial Parameter Estimation by Maximum Likelihood

Let us assume a random variable X that can take on n possible values $x \in \{v_1, v_2, \ldots, v_q\}$. Assuming we have a database D of N observed values for $\mathsf{D} = \{x_1, x_2, \ldots, x_N\}$, we would like to infer the probability function $p\left(x\middle|\mathsf{D}\right)$. For each of the possible q values of X we assign a parameter with the value of the probability function. We then have q parameters θ_{v_i} with $p(x = v_i) = \theta_{v_i}$. All these parameters can be for notation purpose regrouped in one vector $\boldsymbol{\theta}$.

It is very common to approach the parameter estimation using the maximum likelihood approach [82]. The best estimate for the parameter is the one maximizing the (log)-likelihood of the data l:

$$l(\mathsf{D},\boldsymbol{\theta}) = \log p\left(\mathsf{D}\,|\boldsymbol{\theta}\right) = \log p\left(x_1, x_2, \ldots, x_N|\boldsymbol{\theta}\right) = \sum_{t=1}^{N} \log p\left(x_t|\boldsymbol{\theta}\right)$$
$$= \sum_{x} n(x) \log \theta_x \tag{15}$$

This derivation has been performed assuming that all the x_i observations are independent. $n(x)$ indicates the number of times the value x is observed in the data D. Maximizing the likelihood function in (15), under the constraint that $\sum_x \theta_x = 1$, leads to

$$\theta_x^{\mathrm{ML}} = \frac{n(x)}{\sum_{x'} n(x')}. \tag{16}$$

The maximum likelihood estimate of the probability for a given value to be drawn is therefore the frequency at which this value appeared in the data set.

5.5.2 Single Variable Multinomial Parameter Bayesian Estimation

In the simple maximum likelihood approach presented in the previous section, there is one set of values for the $\boldsymbol{\theta}$ parameters. Another approach, called Bayesian estimation, considers that assigning a *unique* value for a parameter is too rigid and argues that one should be interested in discovering instead the probability distribution of the parameter $p\left(\boldsymbol{\theta}|\mathsf{D}\right)$. As an illustration, if one is observing a coin toss leading to 1,001 heads and 999 tails, a maximum likelihood approach would find out that the probability for heads should be 0.5005. A Bayesian approach, in contrast, will argue that from this information one cannot rule out the possibility that the value of the parameter is 0.5 for example. From this information the Bayesian approach would rather propose a $p\left(\boldsymbol{\theta}|\mathsf{D}\right)$ peaked on 0.5005 but allowing some spread and non-zero values for values close to 0.5005. A very complete presentation of the Bayesian approach to probability can be found in Jaynes [83].

In the Bayesian approach, the probability for a value x to be observed is now computed by integrating on all possible values of $\boldsymbol{\theta}$ weighted by their probability:

$$p\left(x|\mathsf{D}\right) = \int p\left(x|\boldsymbol{\theta},\mathsf{D}\right) p(\boldsymbol{\theta},\mathsf{D}) d\boldsymbol{\theta}. \tag{17}$$

The parameters θ_x are now defined as

$$\theta_x = p\left(x\middle|\boldsymbol{\theta}, \mathsf{D}\right). \tag{18}$$

The parameter estimation process consists in finding $p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right)$. Using Bayes' rule of probability, we can show that

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = p\left(\mathsf{D}\middle|\boldsymbol{\theta}\right)\frac{p(\boldsymbol{\theta})}{p(\mathsf{D})} \tag{19}$$

$$= p\left(x_1, x_2, \ldots, x_N\middle|\boldsymbol{\theta}\right)\frac{p(\boldsymbol{\theta})}{p(x_1, x_2, \ldots, x_N)} \tag{20}$$

$$= \lambda \prod_x \theta_x^{n(x)} p(\boldsymbol{\theta}) \tag{21}$$

With $\lambda = \dfrac{1}{p(x_1, x_2, \ldots, x_N)}$.

A new quantity appeared during this derivation: $p(\boldsymbol{\theta})$. This is called the *prior* on the parameters. This represents the a priori belief the observer had before any observation was actually done. In the multinomial case, a common prior used for convenience is the Dirichlet distribution:

$$p(\boldsymbol{\theta}) = \beta(\boldsymbol{\alpha}) \prod_x \theta_x^{\alpha_x - 1}, \tag{22}$$

where $\beta(\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_x \alpha_x\right)}{\prod_x \Gamma(\alpha_x)}$ and Γ is the Gamma function. Plugging the Dirichlet prior (22) in the expression of the posterior (20), we get

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = \lambda\beta(\boldsymbol{\alpha}) \prod_x \theta_x^{n(x)+\alpha_x - 1}. \tag{23}$$

As we can see, using the Dirichlet prior with a multinomial distribution leads to a multinomial distribution as posterior. This very convenient behavior makes the Dirichlet distribution the so-called conjugate prior of a multinomial distribution.

The last piece of our problem not yet solved is the value of λ. We can use the normalization condition $\int p(\boldsymbol{\theta}|\mathsf{D})d\boldsymbol{\theta} = 1$. Applying this constraint, it can be shown that

$$p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right) = \Gamma\left(\sum_{x'} n\left(x'\right) + \alpha_{x'}\right) \prod_x \frac{\theta_x^{n(x)+\alpha_x - 1}}{\Gamma(n(x) + \alpha_x)} \tag{24}$$

$$= C(n, \boldsymbol{\alpha}) \prod_{x'} \theta_{x'}^{n\left(x'\right)+\alpha_{x'}}, \tag{25}$$

where the part of the expression involving the Gamma function has been regrouped for clarity in $C(n, \boldsymbol{\alpha})$. Now that we have found the expression for $p\left(\boldsymbol{\theta}\middle|\mathsf{D}\right)$, we can evaluate the probability to observe a value v_i for the variable X:

$$p(x = v_i) = \int \theta_{v_i} p(\boldsymbol{\theta}, \mathsf{D}) d\theta \tag{26}$$

$$= C(n, \boldsymbol{\alpha}) \int \theta_{v_i} \prod_{x'} \theta_{x'}^{n(x')+\alpha_{x'}} d\theta \tag{27}$$

$$= \frac{n(v_i) + \alpha_{v_i}}{\sum_{x'} n(x') + \alpha_{x'}}. \tag{28}$$

This final expression can be compared to that obtained using the maximum likelihood (16). The way the prior influences the result is by adding extra counts α_x to the evaluation of the probability. We can see that if there is an important amount of data available the probability will be driven mainly by the frequency of counts. On the other hand, if there are very few data points, the prior will drive the probability.

While we have chosen the Dirichlet prior, we still have to choose what parameters $\boldsymbol{\alpha}$ to use. There is no unique answer to that question. This choice would depend on the prior belief we have in the outcome. In the case of no prior information being available [84, 85], there is a common choice of prior called the minimum information uniform Dirichlet prior, where $\boldsymbol{\alpha}$ is chosen as

$$\alpha_x = \frac{1}{q} \tag{29}$$

where q represents the number of possible values for X.

5.5.3 Generalization to Multiple Variables

The results presented in the two previous sections can be generalized for multiple variables. Let us say that we have two variables X and Y and we want to estimate $p(x, y|\mathsf{D})$. D refers to a set of N observations $\mathsf{D} = \{(x, y)_1, (x, y)_2, \ldots, (x, y)_N\}$. If there are q possible values for X and r values possible for Y, then there are qr possible values for the pair (X, Y). Results from the single variable case can then be directly used with a multinomial defined on qr values. Then the maximum likelihood is

$$\theta_{x,y}^{\mathrm{ML}} = \frac{n(x, y)}{N}; \tag{30}$$

the Bayesian estimate is

$$p(x = v_i, y = w_j|\mathsf{D}) = \frac{n(v_i, w_j) + \alpha_{v_i, w_j}}{N + \sum_{x,y} \alpha_{x,y}}; \tag{31}$$

and the minimum information Dirichlet prior is

$$\alpha_x = \frac{1}{qr}. \tag{32}$$

5.6 *Searching for Unknown Ternary Oxides Using Data Mining Compound Prediction*

Ternary oxides are important for many technologies. The model presented here has been used to search for new ternary oxides. We estimated a cumulant expansion probabilistic model (14) using the oxide experimental data available in the ICSD [24] and the Bayesian estimation procedure presented in Sect. 5.5. The 2006 version of the ICSD was searched for duplicate compounds. After this analysis, 616 unique binary and 4,747 ternary oxides compounds were identified. These compounds were grouped by crystal structure prototype. Both duplicate checks and prototyping were performed using Hundt et al.'s algorithm [79]. Composition bins were binned into the 30 most common binary compositions and the 120 most common ternary compositions. Any compound not fitting perfectly in one of these bins was binned in the closest composition bin. Adding the 3 element variables, 183 variables were used in total in the probability model.

5.6.1 New Ternary Oxides Predictions

We then searched for new compounds in 2,211 A-B-O systems with A and B taken from H, Li, Be, B, C, N, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Y, Zr, Nb, Mo, Ag, Cd, In, Sn, Sb, Te, I, Cs, Ba, La, Hf, Ta, W, Pt, Hg, Tl, Pb, Bi, Ce, Pr, Nd, Sm, Eu, Gd, Dy, Ho, Er, Tm, Yb, and Lu. In these systems we used the procedure described in Fig. 7 and searched for compositions where no ternary oxide is given in the ICSD but for which the probability for forming a compound (7) is higher than a certain threshold. This threshold represents a compromise between the computational budget required and the rate of discovery expected. The value of the threshold we chose suggested 1,261 possible compositions and exhibited a 45% true positive rate during cross-validation. At these selected compositions, the most likely crystal structures were determined from the data mined probability density using (8). The number of suggested crystal structures at each composition corresponds to the list length that gave 95% accuracy in cross-validation. This corresponds to a total of 5,546 crystal structures whose energy needed to be calculated with ab initio DFT. All existing binary, ternary, and element structures in the ICSD were also calculated so that relative phase stability can be assessed (using the thermodynamical convex hull construction presented in Sect. 2). Hence, a new structure is stable when its energy is lower than any combination of energies of compounds in the system weighted to the same composition.

From the 1,261 compositions suggested by the model, the ab initio computations confirmed 355 to be stable against every compound known in the ICSD.

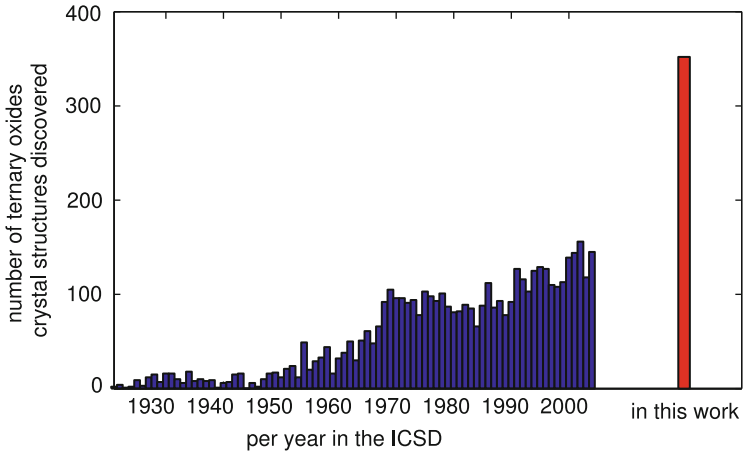

Fig. 8 New ternary oxide discovery per year according to the ICSD. The *bars* from 1930 to 2005 indicate the number of new ternary oxides discovered per year. They are compared to the number of new compounds discovered in this work

This represents 1 new stable compound predicted per 16 DFT computations. A fully exhaustive search (i.e., computing all possible structure prototypes in any composition bin) in the 2,211 A-B-O systems of interest would be prohibitive and require 5,428,287 computations. Even restricting such an exhaustive search to the crystal structure prototypes present in the selected 1,261 compositions bins would need substantially more computations (183,007) than the 5,546 needed while using the machine learned model.

To put this number of 355 new compounds predicted in perspective, we compared it to the number of experimentally discovered and characterized ternary oxides. We identified the earliest date of publication for any ternary oxide compound present in the ICSD. We did not take into account multiple reports of the same compound and compounds with partial occupancies. Figure 8 indicates in blue how many new ternary oxide compounds were discovered each year according to the ICSD from 1930 to 2005. The red bar shows how many new compounds have been discovered in this work. The experimental discovery rate for ternary oxides is around 100 per year since the 1970s. The 355 new compounds suggested were obtained with about 55 days of computing on 400 Intel Xeon 5140 2.33-Ghz cores. Those numbers show the potential for accelerating new compound discovery through combining data mining with DFT computations.

Details and discussion on the results are available in Hautier et al. [78] and details of all the new compounds are available on a web site [86].

6 Data Mined Ionic Substitution Model

In Sect. 6 we present a compound prediction algorithm based on correlations between crystal structures co-existing in a same chemical system. This algorithm was used in combination with high-throughput DFT computations to discover new ternary oxides.

While, in theory, this algorithm can be used to make predictions in chemical systems with any number of components, there are practical limitations to its application, for instance, to the prediction of quaternary compounds. Indeed, the data available for quaternaries is sparser than for ternaries, making the extraction of informative correlations more difficult. More specifically, as the model presented in the previous section is based on correlations between crystal structure prototypes, it shows predictive limits for the crystal structure prototypes appearing only once in the database. Those unique crystal structure prototypes do not have enough occurrences for the model to capture useful correlations. The problem associated with unique prototypes is already present in ternary compounds but tends to be even more critical in the quaternary space. In the ICSD, 20% of the ternary crystal structure prototypes are unique but up to 50% are unique in the case of quaternary prototypes.

In the coming section we will show how a different data mining approach can be used to make predictions in sparser regions. A probabilistic model can be built to assess the likelihood for ionic species to substitute for each other while retaining the crystal structure [87]. We describe the mathematical model and its training on an experimental crystal structures database. The model predictive power is then evaluated by cross-validation and the emerging chemical substitution rules are analyzed.

6.1 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 [88].

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 [89]. While those rules have been widely used to rationalize a posteriori experimental observations, they lack a real quantitative predictive power.

The data mining ionic substitution approach follows this substitution idea but proposes 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.

This method follows an approach used in the field of machine translation [90]. 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.

6.2 The 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 an n component vector:

$$\mathbf{X} = (X_1, X_2, \ldots X_n). \tag{33}$$

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

$$\Omega = \left\{Fe^{2+}, Fe^{3+}, Ni^{2+}, La^{3+}, \ldots\right\}. \tag{34}$$

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

$$p_n\left(\mathbf{X}, \mathbf{X}'\right) = p_n\left(X_1, X_2, \ldots, X_n, X'_1, X'_2, \ldots, X'_n\right). \tag{35}$$

Knowing such a probability function allows one 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*.[91] 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 a result. They can be as complex as required by the problem. For an ionic substitution model, one could choose, for example, as a feature function:

$$f\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & \text{if } Ca^{2+} \text{ substitutes for } Ba^{2+} \text{ in the presence of } O^{2-} \\ 0 & \text{else} \end{cases} \tag{36}$$

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:

$$p_n\left(\mathbf{X},\mathbf{X}'\right) \approx \frac{e^{\sum_i \lambda_i f_i^{(n)}\left(\mathbf{X},\mathbf{X}'\right)}}{Z}. \tag{37}$$

Here λ_i indicates 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 (37) follows a commonly used convention in the machine learning community [92].

The model presented is extremely general and can be adjusted by using whatever feature function is considered relevant. A first assumption made 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, etc.).

Therefore we will omit any reference to n in the probability and feature functions. Equation (37) then becomes

$$p_n\left(\mathbf{X},\mathbf{X}'\right) \approx \frac{\mathrm{e}^{\sum_i \lambda_i f_i\left(\mathbf{X},\mathbf{X}'\right)}}{Z}. \tag{38}$$

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

$$f_k^{a,b}\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & X_k = a \quad \text{and} \quad X'_k = b \\ 0 & \text{else} \end{cases} \tag{39}$$

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 functions will be related to the Ca^{2+} to Ba^{2+} substitution:

$$f_k^{\mathrm{Ca}^{2+},\mathrm{Ba}^{2+}}\left(\mathbf{X},\mathbf{X}'\right) = \begin{cases} 1 & X_k = \mathrm{Ca}^{2+} \quad \text{and} \quad X'_k = \mathrm{Ba}^{2+} \\ 0 & \text{else} \end{cases} \tag{40}$$

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

Finally, the features weights should satisfy certain constraints so that any permutations of the components do not change the result of the probability evaluation. Those symmetry conditions are

$$\lambda_k^{a,b} = \lambda_k^{b,a}, \tag{41}$$

and

$$\lambda_k^{a,b} = \lambda_l^{a,b}. \tag{42}$$

6.3 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 [79, 93]. 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 $t = 1, \ldots, m$

$$\mathsf{D}=\left\{\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^1,\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^2,\ldots,\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^{m-1},\left(\mathbf{X},\mathbf{X}'\right)=\left(\mathbf{x},\mathbf{x}'\right)^m\right\}. \tag{43}$$

Coming back to our analogy to machine translation, probabilistic translation models are estimated from databases of texts with their corresponding translation. The analogue 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 a database [82]. The weights maximizing the likelihood to observe the training data are considered as the best estimates to use in the model. For notation purposes 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 as

$$l(\mathsf{D},\boldsymbol{\lambda}) = \sum_{t=1}^{m} \log p\left(\left(\boldsymbol{x},\boldsymbol{x}'\right)^t|\boldsymbol{\lambda}\right) \tag{44}$$

$$= \sum_{t=1}^{m}\left[\sum_i \lambda_i f_i\left(\left(\boldsymbol{x},\boldsymbol{x}'\right)^t\right) - \log Z(\boldsymbol{\lambda})\right] \tag{45}$$

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

$$\boldsymbol{\lambda}_{\mathbf{ML}} = \arg\max_{\boldsymbol{\lambda}} l(\mathsf{D},\boldsymbol{\lambda}). \tag{46}$$

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.

6.4 Compound Prediction Process

When the substitution probabilistic model in (37) 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

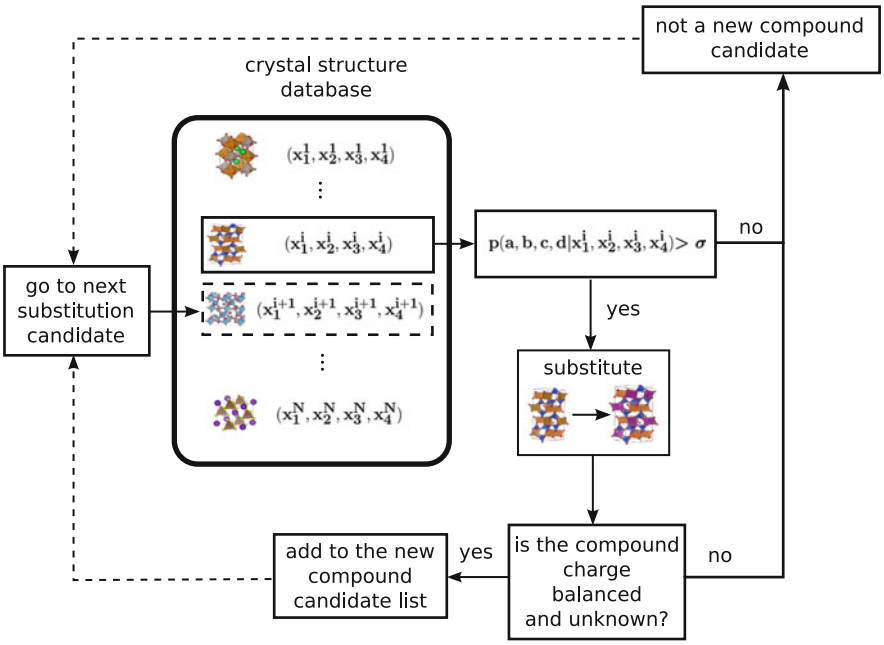

Fig. 9 Procedure to predict new compounds formed by the *a*, *b*, *c*, and *d* species using the substitutional probabilistic model. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

in Fig. 9. 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 compound 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 vs all already known compounds by accurate ab initio techniques such as DFT (see Sect. 2).

6.5 *Analysis of the Model*

A binary feature model based on the ternary and quaternary ionic compounds present in the inorganic crystal structure database (ICSD, [24]) 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 using Hundt et al.'s algorithm [79] and used to obtain the database D of m assignments ((43) necessary to train the model. A binary feature model was fitted on this data set using a maximum likelihood procedure.

6.5.1 Cross-Validation on Quaternary ICSD Compounds

The procedure to discover new compounds using the probabilistic model was presented in Sect. 6.4. Using this procedure, we evaluated the predictive power of this approach by performing a cross-validation test [70]. 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 three equal-sized groups. We performed three cross-validation tests using all compounds in one of the groups as test set and the remaining quaternary and ternary compounds as training set. This extensive cross-validation tested 2,967 compounds in total. The cross-validation tests excluded compounds forming in prototypes unique to one compound, as our substitution strategy by definition cannot predict compounds in such unique prototypes. We also only considered substitution leading to charge balanced compounds.

Figure 10 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):

$$TP_{rate}(\sigma) = \frac{TP(\sigma)}{P}, \tag{47}$$

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):

$$FP_{rate}(\sigma) = \frac{FP(\sigma)}{N}, \tag{48}$$

where P is the number of compounds of proposed compounds non-existing in the ICSD but considered during cross-validation and $TP(\sigma)$ 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 give more true hits at the expense of

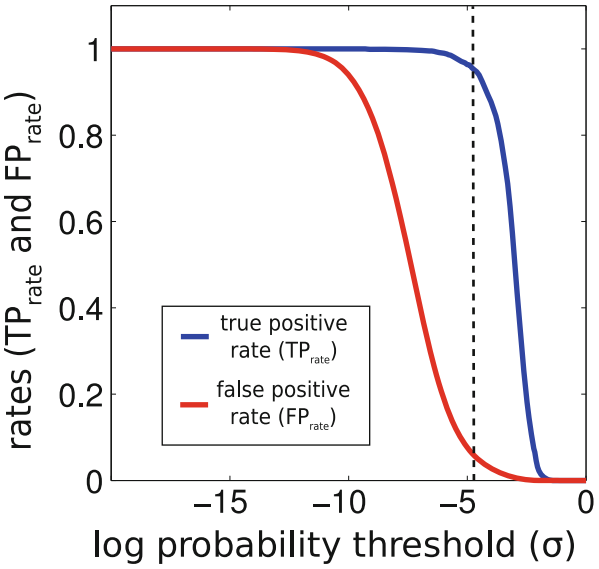

Fig. 10 True positive rate (TP_{rate}, *blue line*) and false positive rate (FP_{rate}, *red line*) in function of the probability threshold (σ) logarithm during cross-validation. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

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 Fig. 10 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, Fig. 10 can be used to estimate a value 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 in Fig. 10 by a dashed line.

6.5.2 Ionic Pair Substitution Analysis

The tendency for a pair of ions to substitute for each other can be estimated by computing the pair correlation:

$$g_{ab} = \frac{p\left(X_1 = a, X'_1 = b\right)}{p(X_1 = a)p(X_1 = b)} \tag{49}$$

$$= \frac{p\left(X_1 = a, X'_1 = b\right)}{\sum_j p\left(X_1 = a, X'_1 = x'_j\right) \sum_j p\left(X_1 = b, X'_1 = x'_j\right)} \tag{50}$$

$$= \frac{\frac{1}{Z} e^{\lambda_1^{a,b}}}{\frac{1}{Z} \sum_j e^{\lambda_1^{a,x'_j}} \frac{1}{Z} \sum_j e^{\lambda_1^{b,x'_j}}} \tag{51}$$

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, i.e., a sum over all possible ionic species.

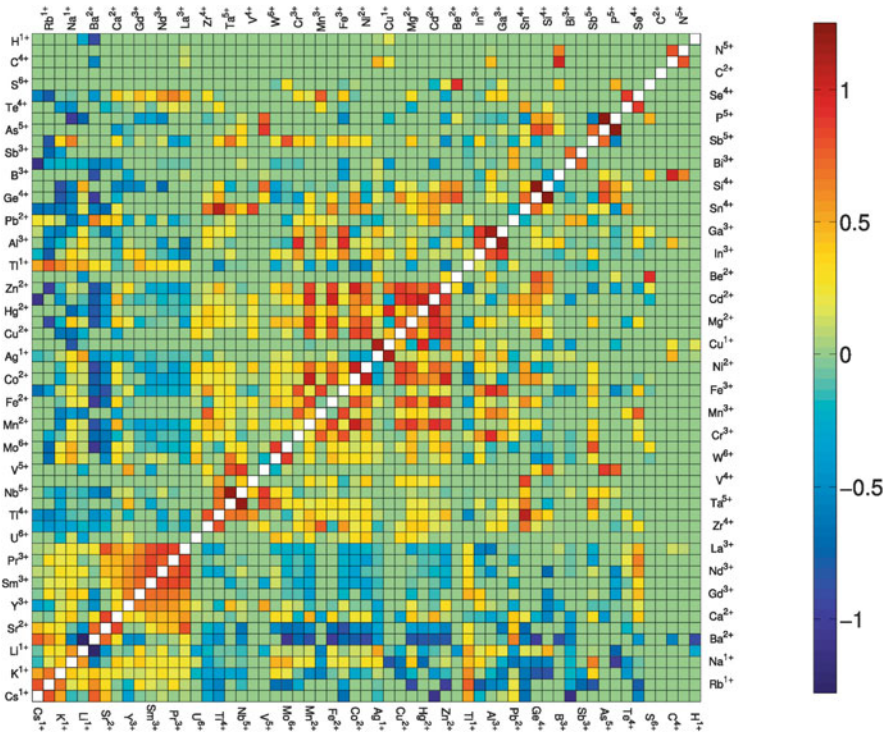

Fig. 11 Logarithm (base 10) of the pair correlation g_{ab} for each ion couple a, b. Equation (49) was used to evaluate the pair correlation g_{ab}. The ions are sorted according to their element's Mendeleev number. Only the 60 most common ions in the ICSD are presented in this graph. These correlation coefficients were obtained by training our probabilistic model on the ICSD. Positive values indicate a tendency to substitute while negative values, in contrast, show a tendency not to substitute. The symmetry of the pair correlation ($g_{ab} = g_{ba}$) is reflected in the symmetry of the matrix. Reprinted with permission from [87]. Copyright 2011 American Chemical Society

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 1 ($g_{ab} > 1$) 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 11 plots the logarithm (base10) of this pair correlation for the 60 most common cations in the ICSD (the pair correlation for all the ionic pairs is presented in supplementary information). Positive values indicate a tendency to substitute while negative values show a tendency not to substitute. The ions are sorted by their element Mendeleev number [65]. 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 also form 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 into 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 [94].

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 earth metals. 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 $MgSeO_3$, $CoSeO_3$, $ZnSeO_3$, $CrLaO_3$, $InLaO_3$, $MnPrO_3$, etc...

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 large +3 ion substitutes mainly with Pb^{2+} and Bi^{3+}, while the smaller +5 ion substitute preferentially with transition metals Mo^{6+}, Cr^{3+}, Fe^{3+}, etc.

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.

6.6 Limits and Strengths of the Model

The substitution model makes several simplifying assumptions. The absence of dependence on 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. In reality, 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. [95]). In some ways 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 this context to be taken 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 vs more powerful models such as that presented in Sect. 5.

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 [92].

The ionic substitution model has been used to search with high-throughput computing for novel multicomponent oxides and polyanionic systems (e.g., phosphates) in the field of Li-ion batteries [8, 38, 96, 97]. The technique has also been used recently to explore the field of oxynitrides for water splitting. The lack of knowledge of oxynitride chemistry justified relying heavily on data mining driven compound prediction [13].

7 From Computer to Synthesis: Examples of Successful Compound Prediction Through Data Mining

The ultimate success of a compound prediction technique is to lead to an experimental synthesis of the predicted phase. The theoretical approaches presented in this review chapter have already led to several successful syntheses of compounds suggested through computation. We will outline briefly (and not exhaustively) some of those successful predictions and describe their context.

7.1 Assigning a Structure to a Powder Diffraction Pattern

There are a significant number of compounds present in powder diffraction databases (e.g., the PDF4+ database [98]) that do not have any crystal structure assigned. This is an important issue, especially for computational materials science, as ab initio computations need a material's crystal structure to evaluate any property. Structure assignment from powder diffraction data, for instance by Rietveld refinement, needs a structural guess of the crystal structure that data mining crystal structure prediction algorithms can provide. In the large scale search for ternary oxides presented in Sect. 5, 355 compounds not present in the ICSD were suggested [78]. Of those 355 compounds, 64 compositions are present in a powder diffraction database but without any structural data associated with the ICSD. Figure 12 compares the simulated vs the experimental powder diffraction spectrum present in the PDF database for two predicted compounds: $MgMnO_3$ and $CoRb_2O_3$ (00-024-0736 [99] and 00-027-0515 [100]). Not only did the algorithm identify successfully the stoichiometries absent from the ICSD 2006 database (without data from the PDF database) but the computed and experimental patterns are in good agreement (if one takes into account the overestimation of the lattice constant by a few percent present with DFT computations in the generalized gradient approximation). Only one peak in the 50° region does not match the powder diffraction pattern for $MgMnO_3$.

These two examples show that a purely data mining driven approach based on no human intervention can successfully assign crystal structure to powder diffraction patterns.

7.2 $SnTiO_3$

Among the compounds without any data available (even powder diffraction data), the large scale data mined ternary oxide search presented in Sect. 5 found $SnTiO_3$ to be a stable stoichiometry with an ilmenite structure being the most stable phase. This $SnTiO_3$ ilmenite prediction is of technological interest as $SnTiO_3$ perovskite has been predicted through ab initio computation to be a good candidate Pb-free

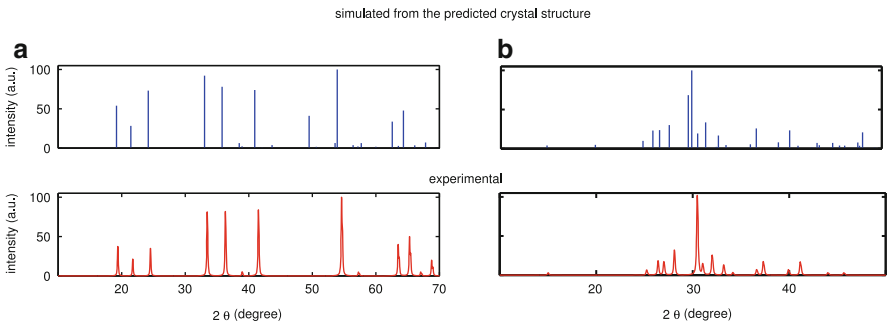

Fig. 12 Comparison between the predicted (*above*) and the experimental (from PDF4+ database, *below*) powder diffraction patterns for $MgMnO_3$ (**a**) and $CoRb_2O_3$ (**b**)

ferroelectric material [101]. Unfortunately, the interesting piezoelectric properties are only present for the perovskite structure. The synthesis of $SnTiO_3$ had been unsuccessful at the time of publication of the paper on ternary oxides but was reported very shortly after by Fix et al. [102]. The experimental results very clearly confirm the computed prediction of an ilmenite phase. Not only is this example a success of computational prediction but it illustrates how important it is to study the stability of the phases that are used to make materials properties prediction in the ab initio literature.

7.3 $Li_9V_3(P_2O_7)_3(PO_4)_2$

Finding novel cathodes for Li-ion batteries is of great importance for energy storage [103–105]. Using the possibility to predict important battery properties by ab initio computations (voltage, Li-ion diffusion, stability when charged) [106, 107], a high-throughput computational search for new cathode materials has been performed by Ceder et al. This project made extensive use of some of the data mining based compound prediction approaches that have been previously described.

During this high-throughput study, an entirely novel phase – $Li_9V_3(P_2O_7)_3(PO_4)_2$ – was predicted by the ionic substitution approach suggesting that a substitution of Fe^{3+} to V^{3+} in $Li_9Fe_3(P_2O_7)_3(PO_4)_2$ leads to a compound lying low in energy [8, 108]. This example shows how unusual structures, beyond the common spinels, rock salt, ilmenite etc., can also be suggested by data mining approaches and lead to technologically relevant materials.

We should note that an independent report on this phase by Kuang et al. [109] had appeared in the literature. However, the patent anteriority date from the Ceder team (before Kuang et al.'s publication) clearly confirms the true predictive nature of the result.

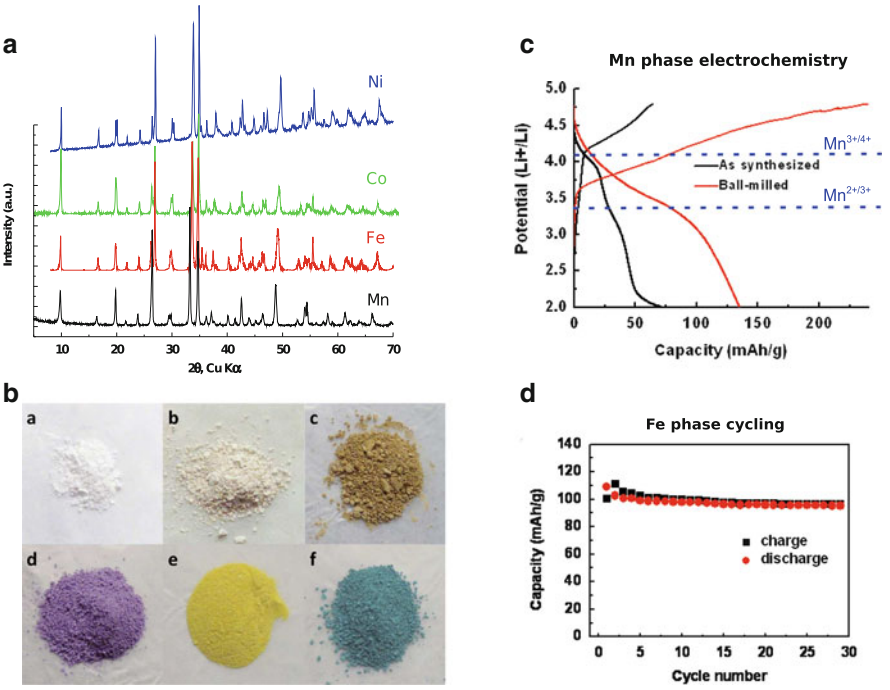

Fig. 13 XRD patterns (**a**) and powders (**b**) of first-time synthesized $Na_3M(CO_3)(PO_4)$ with M = Mn, Ni, Fe, Co, etc. The electrochemical activity (voltage vs capacity) of the Mn-based Li version $Li_3Mn(CO_3)(PO_4)$ (**c**) and the cyclability of the $Li_3Fe(CO_3)(PO_4)$ phase (**d**). Adapted with permission from [33] and [110]. Copyright 2012 American Chemical Society

7.4 *Sidorenkite*

The high-throughput cathode project also led to the identification of an even more exotic class of materials: the sidorenkite carbonophosphates [33, 38, 110]. Carbonophosphates had only been known as rare minerals but were identified by high-throughput computations to form very promising lithium-ion battery cathodes. The predicted compounds were then synthesized by hydrothermal reaction followed by ion exchange as suggested by computational phase stability analysis. Some carbonophosphates have shown electrochemical activity and very good cyclability as Li-ion battery cathode (see Fig. 13c, d).

7.5 *$LiCoPO_4$*

Compound prediction can also push for the reinvestigation of chemical systems that were believed to be very well known. In their high-throughput phosphate analysis, Hautier et al. made the surprising observation that data mining and DFT suggested

a polymorph of the well studied $LiCoPO_4$ olivine structure [8]. While $LiCoPO_4$ olivine incorporates Co coordinated by octahedra of oxygen, the new predicted polymorph shows the structure of $LiZnPO_4$ based on tetrahedral Co. The prediction was confirmed by Jähne et al. when they reported on the first synthesis of tetrahedral $LiCoPO_4$ in the structure that was suggested computationally [111].

8 Conclusion and Future Avenues

Materials science is moving more and more towards computationally oriented materials design. Compound and crystal structure prediction is a critical step in this new paradigm. Current DFT techniques are mature enough to model the phase stability reasonably well and different approaches to compound predictions have been developed. Among them, data mining offers high-throughput-friendly, efficient methods that have already been used in several fields from Li-ion batteries to oxynitrides for water splitting. We not only presented these methods in details but also reported on several successes where computational predictions were confirmed by experimental synthesis.

In the future, the development of large databases of freely available computed data such as the Materials Project will surely help in providing large data sets to be used for fitting more efficient data mining crystal structure prediction models. We can expect an improvement in the predictive power of data mining based techniques as the models are refined and the data sets become larger.

However, the main limitation of data mining techniques is their inability to predict (in contrast to optimization techniques such as genetic algorithms) crystal structures that have never been observed before. Combination of optimization and data mining approaches could offer a solution to this problem, aiming at keeping the low computational budget of knowledge-based methods while approaching the exhaustivity of the optimization approaches.

We hope the many compound prediction techniques available and the current understanding of the accuracy of phase stability prediction will in the future make phase stability a more central part of the computational materials design process. Too often new phases with exceptional computed properties are proposed without assessing their phase stability.

Finally, while computations can be truly predictive to determine the existence of an inorganic phase, the step between computational compound prediction and finding the most appropriate synthesis route is still very empirical. A better fundamental understanding of the different synthesis approaches (solid state reaction, hydrothermal, etc.) needing a joint effort from experimentalists and theorists would be of great value here.

References

1. Kohn W, Sham L (1965) Self-consistent equations including exchange and correlation effects. Phys Rev 140(4A):1131–1138
2. ABINIT (2004). http://www.abinit.org/. Accessed 1 July 2013
3. Vienna ab initio simulation package (VASP). http://www.vasp.at/. Accessed 1 July 2013
4. Quantum Espresso (2012). http://www.quantum-espresso.org/. Accessed 1 July 2013
5. Hautier G, Jain A, Ong SP (2012) From the computer to the laboratory: materials discovery and design using first-principles calculations. J Mater Sci 47(21):7317–7340
6. Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nat Mater 12(3):191–201
7. Greeley J, Jaramillo TF, Bonde J, Nørskov JK, Chorkendorff IB (2006) Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. Nat Mater 5(11):909–913
8. Hautier G, Jain A, Ong SP, Kang B, Moore C, Doe R, Ceder G (2011) Phosphates as lithium-ion battery cathodes: an evaluation based on high-throughput ab initio calculations. Chem Mater 23:3495–3508
9. Mueller T, Hautier G, Jain A, Ceder G (2011) Evaluation of tavorite-structured cathode materials for lithium-ion batteries using high-throughput computing. Chem Mater 23:3854–3862
10. Setyawan W, Gaume RM, Lam S, Feigelson RS, Curtarolo S (2011) High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. ACS Comb Sci 13(4):382–390
11. Castelli IE, Olsen T, Datta S, Landis DD, Dahl S, Thygesen KS, Jacobsen KW (2012) Computational screening of perovskite metal oxides for optimal solar light capture. Energy Environ Sci 5(2):5814
12. Jain A, Castelli IE, Hautier G, Bailey DH, Jacobsen KW (2013) Performance of genetic algorithms in search for water splitting perovskites. J Mater Sci 48:6519–6534
13. Wu Y, Lazic P, Hautier G, Persson K, Ceder G (2013) First principles high throughput screening of oxynitrides for water-splitting photocatalysts. Energy Environ Sci 6:157–168
14. Madsen GKH (2006) Automated search for new thermoelectric materials: the case of LiZnSb. J Am Chem Soc 128(37):12140–12146
15. Wang S, Wang Z, Setyawan W, Mingo N, Curtarolo S (2011) Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. Phys Rev X 1(2):021012
16. Jain A, Seyed-Reihani SA, Fischer CC, Couling DJ, Ceder G, Green WH (2010) Ab initio screening of metal sorbents for elemental mercury capture in syngas streams. Chem Eng Sci 65(10):3025–3033
17. Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sánchez-Carrera RS, Vogt L, Aspuru-Guzik A (2011) Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy Environ Sci 4:4849–4861
18. Yang K, Setyawan W, Wang S, Buongiorno Nardelli M, Curtarolo S (2012) A search model for topological insulators with high-throughput robustness descriptors. Nat Mater 11(7):614–619
19. Materials project. http://www.materialsproject.org. Accessed 1 July 2013
20. Jain A, Hautier G, Moore CJ, Ping Ong S, Fischer CC, Mueller T, Persson KA, Ceder G (2011) A high-throughput infrastructure for density functional theory calculations. Comp Mater Sci 50:2295–2310
21. AFLOWLIB: http://www.aflowlib.org. Accessed 1 July 2013
22. "The Electronic Structure Project", http://gurka.fysik.uu.se/ESP/. Accessed 1 July 2013
23. Service RF (2012) Materials scientists look to a data-intensive future. Science 335:1434–1435

24. Inorganic Crystal Structure Database (ICSD), http://www.fiz-karlsruhe.de/icsd.html, Accessed 1 July 2013
25. Maddox J (1988) Crystals from first principles. Nature 335:201
26. O'Keeffe M (2010) Aspects of crystal structure prediction: some successes and some difficulties. Phys. Chem. Chem. Phys. 12:10–15
27. Woodley SM, Catlow R (2008) Crystal structure prediction from first principles. Nat Mater 7(12):937–946
28. Callen HB (1985) Thermodynamics and an introduction to thermostatistics. Wiley, New York
29. Chandler D (1987) Introduction to modern statistical mechanics. Oxford University Press, Oxford
30. Ceder G, Ven A, Marianetti C, Morgan D (2000) First-principles alloy theory in oxides. Modelling Simul. Mater. Sci. Eng. 8:311–321
31. Van De Walle A, Ceder G (2000) First-principles computation of the vibrational entropy of ordered and disordered Pd_3V. Phys Rev B 61(9):5972–5978
32. Zhou F, Maxisch T, Ceder G (2006) Configurational electronic entropy and the phase diagram of mixed-valence oxides: the case of Li_xFePO_4. Phys Rev Lett 97:155704
33. Chen H, Hautier G, Ceder G (2012) Synthesis, computed stability and crystal structure of a new family of inorganic compounds: carbonophosphates. J Am Chem Soc 134(48):19619–19627
34. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comp Mater Sci 68:314–319
35. Ong SP, Wang L, Kang B, Ceder G (2008) Li-Fe-P-O_2 phase diagram from first principles calculations. Chem Mater 20(5):1798–1807
36. Curtarolo S, Morgan D, Ceder G (2005) Accuracy of methods in predicting the crystal structures of metals: a review of 80 binary alloys. CALPHAD 29(3):163–211
37. Lany S (2008) Semiconductor thermochemistry in density functional calculations. Phys Rev B 78(24):245207
38. Hautier G, Ong SP, Jain A, Moore CJ, Ceder G (2012) Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. Phys Rev B 85:155208
39. Dudarev SL, Savrasov SY, Humphreys CJ, Sutton AP (1998) Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. Phys Rev B 57(3):1505–1509
40. Zhou F, Cococcioni M, Marianetti CA, Morgan D, Ceder G (2004) First-principles prediction of redox potentials in transition-metal compounds with LDA+U. Phys Rev B 70:235121
41. Jain A, Hautier G, Ong SP, Moore CJ, Fischer CC, Persson KA, Ceder G (2011) Formation enthalpies by mixing GGA and GGA+*U* calculations. Phys Rev B 84:045115
42. Stevanović V, Lany S, Zhang X, Zunger A (2012) Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. Phys Rev B 85:115104
43. Oganov AR, Valle M (2009) How to quantify energy landscapes of solids. J Chem Phys 130(10):104504
44. Ceder G (1993) A derivation of the Ising model for the computation of phase diagrams. Comp Mater Sci 1(2):144–150
45. Ducastelle F (1991) Order and phase stability in alloys, volume 3 (cohesion and structure). North Holland, Amsterdam
46. Sanchez JM, Ducastelle F, Gratias D (1984) Generalized cluster description of multicomponent systems. Physica A 128:334–350
47. Blum V, Zunger A (2004) Structural complexity in binary bcc ground states: the case of bcc Mo-Ta. Phys Rev B 69(2):20103
48. Hart GLW (2009) Verifying predictions of the $L1_3$ crystal structure in Cd-Pt and Pd-Pt by exhaustive enumeration. Phys Rev B 80(1):014106

49. Sanati M, Wang L, Zunger A (2003) Adaptive crystal structures: CuAu and NiPt. Phys Rev Lett 90(4):045502
50. Van Der Ven A, Aydinol MK, Ceder G (1998) First-principles evidence for stage ordering in Li_xCoO_2. J Electrochem Soc 145(6):2149
51. Wales DJ, Doye JPK (1997) Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A 101(28):5111–5116
52. Wales DJ, Scheraga HA (1999) Global optimization of clusters, crystals, and biomolecules. Science 285(5432):1368–1372
53. Abraham NL, Probert MIJ (2006) A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. Phys Rev B 73(22):224104
54. Bush TS, Catlow CRA, Battle PD (1995) Evolutionary programming techniques for predicting inorganic crystal structures. J Mater Chem 5(8):1269–1272
55. Oganov AR, Glass CW (2006) Crystal structure prediction using ab initio evolutionary techniques: principles and applications. J Chem Phys 124(24):244704
56. Oganov AR, Glass CW (2008) Evolutionary crystal structure prediction as a tool in materials design. J Phys Condens Matter 20(6):064210
57. Trimarchi G, Zunger A (2007) Global space-group optimization problem: finding the stablest crystal structure without constraints. Phys Rev B 75(10):104113
58. Zhang X, Zunger A, Trimarchi G (2010) Structure prediction and targeted synthesis: a new $Na_{(n)}N_2$ diazenide crystalline structure. J Chem Phys 133(19):194504
59. Oganov AR, Chen J, Gatti C, Ma Y, Ma Y, Glass CW, Liu Z, Yu T, Kurakevych OO, Solozhenko VL (2009) Ionic high-pressure form of elemental boron. Nature 457 (February):863–868
60. Kolmogorov A, Shah S, Margine E, Bialon A, Hammerschmidt T, Drautz R (2010) New superconducting and semiconducting Fe-B compounds predicted with an ab initio evolutionary search. Phys Rev Lett 105(21):217003
61. Ono S, Kikegawa T, Ohishi Y (2007) High-pressure transition of $CaCO_3$. Am Mineral 92(7):1246–1249
62. Gou H, Dubrovinskaia N, Bykova E, Tsirlin AA, Kasinathan D, Richter A, Merlini M, Hanfland M, Abakumov AM, Batuk D, Van Tendeloo G, Nakajima Y, Kolmogorov AN, Dubrovinsky L (2013) Discovery of a superhard iron tetraboride superconductor. Phys Rev Lett 111:157002
63. Liebold-Ribeiro Y, Fischer D, Jansen M (2008) Experimental substantiation of the "energy landscape concept" for solids: synthesis of a new modification of LiBr. Angew Chem Int Edit 47(23):4428–4431
64. Pauling L (1929) The principles determining the structure of complex ionic crystals. J Am Chem Soc 51:1010–1026
65. Pettifor DG (1990) Structure maps in alloy design. J Chem Soc Faraday Trans 86 (8):1209–1213
66. Pettifor DG (2003) Structure maps revisited. J Phys Condens Matter 15:13–16
67. Villars P (1983) A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. J Less Common Met 92(2):215–238
68. Morgan D, Rodgers J, Ceder G (2003) Automatic construction, implementation and assessment of Pettifor maps. J Phys Condens Matter 15:4361–4369
69. Ceder G, Morgan D, Fischer C, Tibbetts K, Curtarolo S (2006) Data-mining-driven quantum mechanics for the prediction of structure. MRS Bull 31:981–985
70. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. (Springer Series in Statistics), Springer, chap 4, pp 80–113
71. von Lilienfeld OA (2013) First principles view on chemical compound space: gaining rigorous atomistic control of molecular properties. Int J Quantum Chem 113(12):1676–1689
72. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108:058301

73. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G (2003) Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 91(13):135503
74. Kolmogorov AN, Curtarolo S (2006) Prediction of different crystal structure phases in metal borides: a lithium monoboride analog to MgB_2. Phys Rev B 73(18):180501
75. Kolmogorov AN, Curtarolo S (2006) Theoretical study of metal borides stability. Phys Rev B 74(22):224507
76. Levy O, Chepulskii RV, Hart GLW, Curtarolo S (2009) The new face of rhodium alloys: revealing ordered structures from first principles. J Am Chem Soc 132(2):833–837
77. Fischer CC, Tibbetts KJ, Morgan D, Ceder G (2006) Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 5(8):641–646
78. Hautier G, Fischer CC, Jain A, Mueller T, Ceder G (2010) Finding nature's missing ternary oxide compounds using machine learning and density functional theory. Chem Mater 22(12):3762–3767
79. Hundt R, Schön JC, Jansen M (2006) CPMZ-an algorithm for the efficient comparison of periodic structures. J Appl Crystallogr 39:6–16
80. Morita T (1957) Cluster variation method of cooperative phenomena and its generalization I. J Phys Soc Jpn 12(7):753–755
81. Fischer CC (2007) A machine learning approach to crystal structure prediction. PhD thesis, Massachusetts Institute of Technology
82. Eliason SR (1993) Maximum likelihood estimation: logic and practice. Sage Publications, Inc, Newberry Park
83. Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge
84. Buntine W (1991) Theory refinement on Bayesian networks. In: Proceedings of the seventh conference on uncertainty in artificial intelligence, Citeseer 91:52–60
85. Lynch RSJ, Willett PK (2003) Adaptive Bayesian classification using noninformative Dirichlet priors. IEEE Trans Syst Man Cybern 33(3):2812–2815
86. Ternary oxides predictions. http://ceder.mit.edu/ternaryoxides, accessed: 01 July 2013
87. Hautier G, Fischer C, Ehrlacher V, Jain A, Ceder G (2011) Data mined ionic substitutions for the discovery of new compounds. Inorg Chem 50:656–663
88. Johrendt D, Pöttgen R (2008) Pnictide oxides: a new class of high-T_C superconductors. Angew Chem Int Edit 47(26):4782–4784
89. Goldschmidt V (1926) Die gesetze der krystallochemie. Naturwissenschaften 14:477–485
90. Brown PF, Della Pietra SA, Della Pietra VJ, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19:263–312
91. Berger A, Della Pietra VJ, Della Pietra SA (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–72
92. Della Pietra SA, Della Pietra VJ, Lafferty J (1997) Inducing features of random fields. IEEE Trans Pattern Anal Mach Intell 19(4):1–13
93. Parthé E, Gelato L (1984) The standardization of inorganic crystal-structure data. Acta Crystallogr A 40:169–183
94. Gaudin E, Boucher F, Evain M (2001) Some factors governing Ag^+ and Cu^+ Low coordination in chalcogenide environments. J Solid State Chem 160(1):212–221
95. Zhang H, Li N, Li K, Xue D (2007) Structural stability and formability of ABO3-type perovskite compounds. Acta Crystallogr Sec B 63:812–818
96. Jain A, Hautier G, Moore CJ, Kang B, Lee J, Chen H, Twu N, Ceder G (2012) A computational investigation of $Li_9M_3(P_2O_7)_2(PO_4)_2$ (M=V, Mo) as cathodes for Li ion batteries. J Electrochem Soc 159(5):A622–A633
97. Ma X, Hautier G, Jain A, Doe R, Ceder G (2013) Improved capacity retention for $LiVO_2$ by Cr substitution. J Electrochem Soc 160(2):A279–A284
98. International centre for diffraction data. PDF4+ database. http://www.icdd.com/products/pdf4.htm. Accessed 1 July 2013
99. Chamberland B, Sleight AW, Weiher JF (1970) Preparation and characterization of $MgMnO_3$ and $ZnMnO_3$. J Solid State Chem 1(3–4):512–514

100. Jansen M, Hoppe R (1974) Neue oxocobaltate (IV):$Cs_2[CoO_3]$, $Rb_2[CoO_3]$ und $K_2[CoO_3]$. Z Anorg Allg Chem 408:75–82
101. Matar S, Baraille I, Subramanian M (2009) First principles studies of $SnTiO_3$ perovskite as potential environmentally benign ferroelectric material. Chem Phys 355(1):43–49
102. Fix T, Sahonta SL, Garcia V, MacManus-Driscoll JL, Blamire MG (2011) Structural and dielectric properties of $SnTiO_3$, a putative ferroelectric. Crystal Growth Des 11:1422–1426
103. Ellis BL, Lee KT, Nazar LF (2010) Positive electrode materials for Li-ion and Li-batteries. Chem Mater 22(3):691–714
104. Goodenough JB, Kim Y (2010) Challenges for rechargeable Li batteries. Chem Mater 22(3):587–603
105. Whittingham MS (2004) Lithium batteries and cathode materials. Chem Rev 104(10):4271–4302
106. Ceder G, Hautier G, Jain A, Ong SP (2011) Recharging lithium battery research with first-principles methods. MRS Bull 36(3):185–191
107. Meng YS, Arroyo-de Dompablo ME (2013) Recent Advances in First Principles Computational Research of Cathode Materials for Lithium-Ion Batteries, Acc Chem Res, 46 (5):1171–1180
108. Ceder G, Jain A, Hautier G, Kim JC, Kang B, Daniel R (2013) Mixed phosphate-diphosphate electrode materials and methods of manufacturing same US8399130 B2
109. Kuang Q, Xu J, Zhao Y, Chen X, Chen L (2011) Layered monodiphosphate $Li_9V_3(P_2O_7)_3(PO_4)_2$: a novel cathode material for lithium-ion batteries. Electrochim Acta 56(5):2201–2205
110. Chen H, Hautier G, Jain A, Moore C, Kang B, Doe R, Wu L, Zhu Y, Tang Y, Ceder G (2012) Carbonophosphates: a new family of cathode materials for Li-ion batteries identified computationally. Chem Mater 24(11):2009–2016
111. Jähne C, Neef C, Koo C, Meyer HP, Klingeler R (2013) A new $LiCoPO_4$ polymorph via low temperature synthesis. J Mater Chem A 1(8):2856