High-Throughput Data Mined Prediction of Inorganic Compounds and Computational Discovery of New Lithium-Ion Battery Cathode Materials

by

Geoffroy Hautier

Ingénieur civil en Sciences des Matériaux, Université Libre de Bruxelles (2004)
Ingénieur, Ecole Centrale Paris (2004)

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author . . . .
Department of Materials Science and Engineering
January 19, 2011

Certified by . . . .
Gerbrand Ceder
R.P. Simmons Professor of Materials Science
Thesis Supervisor

Accepted by . . . .
Christopher Schuh
Chairman, Department Committee on Graduate Theses

High-Throughput Data Mined Prediction of Inorganic Compounds and Computational Discovery of New Lithium-Ion Battery Cathode Materials

by

Geoffroy Hautier

Submitted to the Department of Materials Science and Engineering
on January 19, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The ability to computationally predict the properties of new materials, even prior to their synthesis, has been made possible due to the current accuracy of modern *ab initio* techniques. In some cases, high-throughput computations can be used to create large data sets of potential compounds and their computed properties. However, regardless of the field of application, such a computational high-throughput approach faces a major problem: to be relevant, the properties need to be computed on compounds (i.e., stoichiometries and crystal structures) that will be stable enough to be synthesized. In this thesis, we address this compound prediction problem through a combination of data mining and high-throughput Density Functional Theory. We first describe a method based on correlations between crystal structure prototypes that can be used with a limited computational budget to search for new ternary oxides. In addition, for the treatment of sparser data regions such as quaternaries, a new algorithm based on the data mining of ionic substitutions is proposed and analyzed. The second part of this thesis demonstrates the application of this high-throughput *ab initio* computing technique to the lithium-ion battery field. Here, we describe a large-scale computational search for novel cathode materials with specific battery properties, which enables experimentalists to focus on only the most promising chemistries. Finally, to illustrate the potential of new compound computational discovery using this approach, a novel chemical class of cathode materials, the carbonophosphates, is presented along with synthesis and electrochemical results.

Thesis Supervisor: Gerbrand Ceder
Title: R.P. Simmons Professor of Materials Science

Acknowledgments

First of all, I would like to express all my gratitude to my advisor Gerd Ceder. I came to MIT especially to work with Gerd and I must say I never regretted it. Without Gerd's constant trust and support, this thesis would never have been completed. My acknowledgments also go to my thesis committee Chris Schuh and Jeffrey Grossman.

I had the privilege to interact and collaborate with many outstanding scientists during this thesis. A special thanks should be addressed to Chris Fischer for his patience, friendship and great mentoring. Working on the high-throughput battery project with Anubhav Jain, Charles Moore, Shyue Ping Ong, Kristin Persson, Robert Doe, Hailong Chen, Xiaohua Ma, Byoungwoo Kang, Shirley Meng, Tim Mueller, Denis Kramer, Reece Daniel, Vincent Chevrier, Jae Chul Kim, and Qing Hao has been a real pleasure. My acknowledgments also go to the rest of the Ceder group members who made these years so enjoyable: Maria Chan, Sahak Petrosyan, Kisuk Kang, Fei Zou, Byungchan Han, Kevin Tibbetts, Yoyo Hinuma, Rahul Malik, Yabi Wu, ShinYoung Kang, Ruoshi Sun, Lusann Yang, Aziz Abdellahi, Pedrag Lazic, Yifei Mo, Jinhyuk Lee, Sangtae Kim, and Nancy Twu. Thank you also to Will Tipton, Virginie Ehrlacher, and Wenhao Sun for having been such great undergraduate students to mentor. Kathy Simons, our administrative assistant, needs also to be thanked for her constant help making sure no paperwork was preventing me to do science.

I am grateful to the Belgian American Education Foundation (BAEF), Total, the MIT-Energy initiative, Umicore and Bosch for their generous financial support.

Maintaining a somewhat balanced life during graduate school can be challenging. I was lucky enough to meet amazing friends who made sure to keep me out of the lab. Anubhav Jain, Tiffany Ziebell, Deborah Ha, Bryan Ho, Cristina Mc Calla, Jin Suntivich, Sigh Pichamol Jirapinyo, Karen Shu, Bryan Ng, Charles Moore, Zenzile Brooks, David Bangerter, Andrew Sawchuk, Wui Siew Tan, and Melissa Smith provided a constant source of entertainment. They have been also of great support during the difficult moments and I must thank them all for this. My roommates: Phillip Zukin, Nicolas Smith, Catherine Lefebvre, Sarah Vigeland and Nikita Bernstein, need also to be thanked for having made “174 Morrison Avenue” such a nice place to live.

I must thank also my friends from home for their constant encouragement. When back to Brussels or on the phone, Benjamin Nguyen-Huu, Grégory Gosselin, Aurore Pary, Christophe Maufroy, Nicolas Devue, Didier Quoistiaux, Iouri Tougarinoff, my brother Grégory and his wife Cécile always made me feel as if I had never left.

I cannot be more grateful to anyone than my parents. They always provided me the best education and made my personal fulfillment their priority, whatever path I decided to take. I owe them two skills that have been extremely useful during this thesis: a taste of science and rigor from my mother, and an awareness of the importance of good communication from my father.

Last but not least, I would like to thank my wife, Véro. These years of graduate school have sometime been difficult for us but her unconditional love and support made sure we made it through.

Contents

List of Figures

List of Tables

Chapter 1

Total energy computations with Density Functional Theory

> *The fundamental laws necessary for the mathematical treatment of a large part of physics and the whole of chemistry are thus completely known, and the difficulty lies only in the fact that application of these laws leads to equations that are too complex to be solved*
>
> Paul Dirac, 1929

Computing the total energy of a crystalline solid is necessary to the evaluation of many materials properties. In the following chapters, total energy computations will be used to assess the phase stability of new compounds or the voltage of battery materials. Among the different tools available to the materials scientist, *ab initio* computations in the Density Functional Theory (DFT) framework can provide an accurate estimate of this total energy. [6, 7, 8] *Ab initio* or first principles methods consist in evaluating materials properties at the atomistic level using the fundamental laws of quantum mechanics. In this chapter, we present briefly the different approximations at the root of DFT. The specific implementations and parameters chosen in this thesis are also described. More details on the derivations presented can be found in standard solid state physics and electronic structure textbooks (for instance [9, 10, 11, 12]). In addition, as this thesis uses DFT computations on a large scale, running thousands of DFT computations in many different chemical systems without much human intervention, we will present briefly the infrastructure needed for such *high-throughput ab initio* computations.

1.1 The Schrödinger equation and the many-body problem

From quantum mechanics, the ground state energy E_0 of a system of N nuclei (located at coordinates $\{\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_{N-1}, \boldsymbol{R}_N\}$) and n electrons (located at coordinates $\{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n\}$) can be obtained by solving the time-independent Schrödinger equation:

$$H\psi = E\psi \tag{1.1}$$

Where ψ is the wave function for the full system (nuclei and electrons) and H is the Hamiltonian[1]

$$H = -\sum_i \nabla_i^2 - \sum_{i,I} \frac{Z_I}{|\boldsymbol{r_i} - \boldsymbol{R}_i|} + \frac{1}{2}\sum_{i \neq j} \frac{1}{|\boldsymbol{r_i} - \boldsymbol{r}_j|} - \sum_I \frac{1}{2M_I}\nabla_I^2 + \frac{1}{2}\sum_{i \neq j} \frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_j|} \tag{1.2}$$

Where M_I and Z_I are the masses and charges of the different nuclei.

This hamiltonian can be decomposed in five terms

$$H = T_e + V_{ext} + V_{int} + T_N + E_N \tag{1.3}$$

These terms are

- the kinetic energy of the electrons:

$$T_e = -\sum_i \nabla_i^2 \tag{1.4}$$

- the potential acting on the electrons due to the nuclei

$$V_{ext} = \sum_{i,I} \frac{Z_I}{|\boldsymbol{r_i} - \boldsymbol{R}_i|} \tag{1.5}$$

- the electron-electron interaction

[1] We will use in this chapter the Hartree atomic units

$$V_{int} = \frac{1}{2}\sum_{i \neq j} \frac{1}{|\boldsymbol{r_i} - \boldsymbol{r}_j|} \tag{1.6}$$

- the kinetic energy of the nuclei

$$T_N = \sum_I \frac{1}{2M_I} \nabla_I^2 \tag{1.7}$$

- the interaction between nuclei

$$E_N = \frac{1}{2}\sum_{i \neq j} \frac{Z_I Z_J}{|\boldsymbol{R}_I - \boldsymbol{R}_J|} \tag{1.8}$$

The first approximation we follow is to neglect the kinetic energy of the nuclei. The important difference in mass between electrons and nuclei (on the order of 10^3), makes this Born-Oppenheimer approximation accurate for the vast majority of systems.

The Hamiltonian is reduced then to:

$$H = T_e + V_{ext} + V_{int} + E_N \tag{1.9}$$

The nuclei coordinates are only parameters in the Hamiltonian and the wave function depends only on the electronic variables:

$$\psi(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n) \tag{1.10}$$

The problem we are facing now is a quantum many-body problem for the electrons in an Hamiltonian set by the nuclei positions. This many-body problem is extremely difficult to solve directly and needs typically to be approximated by a simpler problem with a solution expected to be close enough to the exact solution.

1.2 The Hartree-Fock approach

1.2.1 The Hartree approximation

One crude approximation to the many-body problem, proposed by Hartree, consists in assuming that the n electrons can be treated as independent particles. The wave function becomes then:

$$\psi(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n) = \phi_1(\boldsymbol{r}_1)\phi_2(\boldsymbol{r}_2)...\phi_{n-1}(\boldsymbol{r}_{n-1})\phi_n(\boldsymbol{r}_n) \tag{1.11}$$

The $\phi_i(\boldsymbol{r}_i)$ are n independent electron wave functions.

A fundamental result in quantum mechanics states that if E_0 is the ground state energy solution of the Schrodinger equation, for any wavefunction φ:

$$\frac{< \varphi|H|\varphi >}{< \varphi|\varphi >} > E_0 \tag{1.12}$$

This is called the *variational principle.* This principle can be used with the hamiltonian in (1.9) and the constraint that the wave function should have the Hartree form (presented in (1.11)) to prove that the solution to the Schrödinger equation in the Hartree approximation is obtained by solving the *Hartree equation*:

$$[-\frac{1}{2}\sum_i \nabla_i^2 - \sum_I \frac{Z_I}{|\boldsymbol{r_i} - \boldsymbol{R_I}|} + \sum_{j\neq i} \int \phi_j^*(\boldsymbol{r_j}) \frac{1}{|\boldsymbol{r_i} - \boldsymbol{r_j}|} \phi_j(\boldsymbol{r}_j)]\phi_i(\boldsymbol{r_i}) = \epsilon_i \phi_i(\boldsymbol{r_i}) \tag{1.13}$$

Each electron i is treated independently but in an effective potential determined by an integration over the wave functions of the other electrons. In this regard, the Hartree approximation is a mean-field approximation replacing the complicated many-body problem by n simpler problems in a mean-field potential.

When solving the equation for the i^{th} wave function the effective potential depends on all the other wave functions. This equation therefore needs to be solved by *self-consistency.* Self-consistency is a procedure in which the wave function for the step k are found through solving the equation (1.13) with the effective potential determined by the wave function in step $k-1$. The procedure is repeated until all the wave functions converge to a solution.

1.2.2 The Hartree-Fock approximation

Electrons being fermions, the exact wavefunction needs to be antisymmetric by exchange of electrons:

$$\psi(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_j, ..., \boldsymbol{r}_k, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n) = -\psi(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_k, ..., \boldsymbol{r}_j, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n) \tag{1.14}$$

This constraint can be added to the independent electron Hartree approach by using a Slater determinant as wavefunction instead of (1.11):

$$\Psi(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \phi_2(\mathbf{r}_1) & \cdots & \phi_n(\mathbf{r}_1) \\ \phi_1(\mathbf{r}_2) & \phi_2(\mathbf{r}_2) & \cdots & \phi_n(\mathbf{r}_2) \\ \vdots & \vdots & & \vdots \\ \phi_1(\mathbf{r}_n) & \phi_2(\mathbf{r}_n) & \cdots & \phi_n(\mathbf{r}_n) \end{vmatrix} \tag{1.15}$$

Using the variational principle on this Slater determinant, it can be proven that the best solution is obtained by solving the *Hartree-Fock* equation:

$$\begin{aligned}[-\sum \nabla_i^2 - \sum_I \frac{Z_I}{|\boldsymbol{r_i} - \boldsymbol{R}_I|} + \sum_j \int \phi_j^*(\boldsymbol{r_j}) \frac{1}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \phi_j(\boldsymbol{r_j})]\phi_i(\boldsymbol{r}_i) \\ - \sum_j [\int \phi_j^*(\boldsymbol{r}_j) \frac{1}{|\boldsymbol{r_i} - \boldsymbol{r}_j|} \phi_i(\boldsymbol{r}_j)]\phi_j(\boldsymbol{r_i}) = \epsilon_i \phi_i(\boldsymbol{r}_i)\end{aligned} \tag{1.16}$$

The effect of the new constraint is to add a term, called the exchange potential, to the Hartree equation (1.13).

1.2.3 The correlation energy

The Hartree-Fock approach assuming independent electrons in an effective potential is an approximation to the true many-body problem. The energy missing is defined as the *correlation* energy. Many methods exist to introduce this correlation energy very accurately for instance post-Hartree-Fock methods,[13] or quantum monte-carlo methods.[14] However, these methods are computationally very expensive and only the smallest systems can be currently computed. On the other hand, the Density Functional Theory (DFT) approach, by approximating this correlation energy, provides a good compromise between computational resources and accuracy.

1.3 Density functional theory

1.3.1 The Hohenberg-Kohn theorem

The foundation of density functional theory (DFT) relies on the Hohenberg-Kohn theorem.[15] The theorem states that

> A universal functional for the energy $E[n]$, in terms of the charge density $n(\boldsymbol{r})$, can be defined, valid for any external potential $V_{ext}(\boldsymbol{r})$. For any particular potential $V_{ext}(\boldsymbol{r})$, the exact ground state energy of the system is the

global minimum value of this functional, and the density that minimizes the functional is the exact ground state density $n_0(\boldsymbol{r})$.

The total energy of the system can then be expressed as:

$$E_{HK}[n] = T[n] + E_{int}[n] + \int V_{ext}(\boldsymbol{r})n(\boldsymbol{r})d^3\boldsymbol{r} + E_{II} \tag{1.17}$$

Where $E_{HK}[n]$ is the total energy functional, $T[n]$ its kinetic energy part and $E_{int}[n]$ the part coming from the electrons interaction. E_{II} does not depend on the density and is due to the nuclei-nuclei interaction.

This theory implies that, in theory, one does not need to work directly with the complicated many-body wavefunction $\psi(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_{n-1}, \boldsymbol{r}_n)$. Instead, the much simpler charge density of the system $n(\boldsymbol{r})$ can be used. However, even if a functional linking the exact energy of the system to the charge density exists, this functional is unknown and needs to be somehow approximated.

1.3.2 The Kohn-Sham equation

Knowing the Hohenberg-Kohn theorem, it would make sense to deal directly with the electronic density and not refer to wave functions. However, this approach turns out to be not very accurate. The kinetic energy part of the Hohenberg-Kohn functional is especially difficult to approximate directly from the density.

DFT became a theory useful in practice only after Kohn and Sham proposed their ansatz.[16] The Kohn-Sham approach consists in replacing the complicated many-body system with a different auxiliary system of non-interacting electrons having the same charge density than the real system but in a different potential. Even though different than the real interacting system, this non-interacting system sharing the same density will have the same energy according to the Hohenberg-Kohn theorem. The Hohenberg-Kohn expression for the total energy (equation (1.17)) is then rewritten as:

$$E_{KS}[n] = T_S[n] + \int V_{ext}(\boldsymbol{r})n(\boldsymbol{r})d^3\boldsymbol{r} + E_{Hartree}[n] + E_{II} + E_{xc}[n] \tag{1.18}$$

T_S is the independent-particle kinetic energy given by

$$T_S = -\frac{1}{2}\sum_i < \phi_i|\nabla^2|\phi_i >= \frac{1}{2}\sum_i \int |\nabla\phi_i|^2 d^3\boldsymbol{r} \tag{1.19}$$

$E_{Hartree}$ can be related to the Hartree equation (equation (1.13))

$$E_{Hartree}[n] = \int \frac{n(\boldsymbol{r})n(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d^3\boldsymbol{r} d^3\boldsymbol{r'} \tag{1.20}$$

The new term appearing in the Kohn-Sham approach is the *exchange and correlation energy* $E_{xc}[n]$. All the many-body effects lie in this term.

Minimizing the Kohn-Sham functional (equation (1.18)) leads to the *Kohn-Sham equation*:

$$[-\frac{1}{2}\nabla^2 + V_{ext}(\boldsymbol{r}) + V_{Hartree}(\boldsymbol{r}) + V_{xc}(\boldsymbol{r})]\phi_i(\boldsymbol{r}) = \epsilon_i \phi_i(\boldsymbol{r}) \tag{1.21}$$

With:

$$V_{Hartree}(\boldsymbol{r}) = \int \frac{n(\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|} d^3\boldsymbol{r'} \tag{1.22}$$

The equation has the form of the independent electron Hartree equation (equation (1.13)) with an added exchange and correlation $V_{xc}(\boldsymbol{r})$ term. The Kohn-Sham approach is exact in theory. It would lead to the exact many-body solution (exact charge density and ground state energy), if the exact exchange-correlation functional was known. Practically, however, the exchange-correlation functional is approximated.

The Hartree potential and exchange-correlation potential depend on the full charge density. Therefore, the Kohn-Sham equation needs, similarly to the Hartree and Hartree-Fock equations, to be solved self-consistently.

1.3.3 The exchange-correlation approximations

The Kohn-Sham equation separates the kinetic energy and the long-range Hartree term. This facilitates the approximation of the exchange-correlation functional by a local or nearly local functional of the density.

$$E_{xc}(\boldsymbol{r}) = \int n(\boldsymbol{r})\varepsilon_{xc}(n(\boldsymbol{r}))d^3\boldsymbol{r} \tag{1.23}$$

$\varepsilon_{xc}(n(\boldsymbol{r}))$ is the exchange-correlation energy per electron at point $\boldsymbol{r}$. $\varepsilon_{xc}(n(\boldsymbol{r}))$ is local and depends only on the density at the point $\boldsymbol{r}$.

For the homogeneous gas, the $\varepsilon_{xc}(n(\boldsymbol{r}))$ function can be computed analytically for the exchange part and with great accuracy numerically using Monte Carlo methods. The local density approximation (LDA) consists in using the exchange and correlation energy of the homogeneous gas in any problem (atoms, solid, molecules, ...).

A straightforward way to refine the LDA is to make ε_{xc} depend not only on the density at the given point but also on the gradient of this density $|\nabla n(\boldsymbol{r})|$. This lead to the generalized-gradient approximation (GGA). In this thesis, we will use more specifically the GGA parametrization by Perdew, Burke and Enzhenorf (GGA-PBE).[17]

For pedagogical purpose, we have not introduced so far any spin component in our treatment of the electron many-body problem. Practically, spin-polarized computations in DFT are run using two spin charge density: one for the spin up electrons ($n^{\uparrow}(\boldsymbol{r})$) and one for the spin down electrons ($n^{\downarrow}(\boldsymbol{r})$). The Hartree term does not take into account the spin polarization of the electron but the exchange-correlation do. All computations run in this thesis use spin polarized GGA.

1.4 Solving the Kohn-Sham equation for crystals

1.4.1 $\boldsymbol{k}$-points integration

In this thesis, all computations will be performed on crystalline solids. This implies that any system exhibits a translational symmetry. The Bloch theorem applies in this situation and each of the $\phi_i(\boldsymbol{r})$ functions solution of the Kohn-Sham equation will have the form:

$$\phi_{i,k}(\boldsymbol{r}) = e^{i\boldsymbol{k}\boldsymbol{r}} u_{i,\boldsymbol{k}}(\boldsymbol{r}) \tag{1.24}$$

Where $u_{i,\boldsymbol{k}}(\boldsymbol{r})$ is a function with the translational periodicity of the crystal unit cell and $\boldsymbol{k}$ is a wave vector.

The total energy is obtained in this periodic systems by integration in k-space over the Brillouin Zone (a primitive cell in the reciprocal space). Practically, this integration is performed by selecting a particular set of k-points across the Brillouin zone, solving the Kohn-Sham equation for each of those k-points and summing over the results.

Different schemes exist to sample appropriately the Brillouin zone with a set of k-points that will give the best estimate of the full integral. In this work, we will follow the Monkhorst-Pack scheme.[18] The origin will be shifted to the Γ point (center of the Brillouin zone) in the case of hexagonal cells.

1.4.2 Plane-wave expansion

With the $u_{i,\boldsymbol{k}}(\boldsymbol{r})$ functions having the periodicity of the crystal, it is natural to expand them on a plane-wave basis:

$$u_{i,\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{\Omega_{cell}}} \sum_{\boldsymbol{G}} c_{i,\boldsymbol{G}}(\boldsymbol{k}) e^{i\boldsymbol{G}\boldsymbol{r}} \tag{1.25}$$

Where $\boldsymbol{G}$ are the reciprocal lattice vectors, $c_{i,\boldsymbol{G}}$ are expansion coefficients and Ω_{cell} is the volume of the cell.

Practically, the expansion cannot be performed on an infinite number of plane waves and the summation needs to be truncated. The limit to the summation is fixed by a parameter called the energy cut-off E_{cut}. Only plane-waves satisfying the above conditions are considered in the computation:

$$\frac{\hbar^2}{2m_e} |\boldsymbol{k} + \boldsymbol{G}|^2 < E_{cut} \tag{1.26}$$

The plane-wave based Vienna *Ab initio* Package (VASP) (version 4.6) was used to perform all computations presented in this thesis.[19]

1.4.3 Pseudopotentials

Running computations considering all electrons present in a solid can be very expensive and inefficient. Some of the core electrons do not participate in bonding. The idea behind the pseudopotential method is to consider that only valence electrons should be important to capture the essential physics of the system. The true coulombic potential is then replaced by a smoother pseudopotential taking into account the effect of the core electrons. Pseudopotentials are usually built on atomic systems and transferred to larger systems (molecules and solids).

In this work, we will use the Projected-Augmented-Wave (PAW) pseudopotentials.[20] The library of pseudopotentials is provided with VASP.

1.4.4 Ionic relaxation

So far, the positions of the nuclei have been assumed to be known and fixed. Very often, the exact nuclei positions are unknown and one wants to find those nuclei positions by minimizing the total energy. Searching for the nuclei parameters with the lowest energy is called *ionic relaxation* and is performed by standard optimization algorithm such as the conjugate gradient.[21] The procedure consists in starting with initial ionic positions and initial wave functions. An electronic solution to the Kohn-Sham equation is then searched. When a self-consistent solution to the electronic problem is found, the ionic positions are updated and a new electronic solution is sought with the updated nuclei. This procedure is repeated until the nuclei positions minimizing the energy are found.

1.5 Correcting the self-interaction error through GGA+U

One of the major problem with local functionals such as LDA and GGA is the presence of a non-physical interaction of an electron with itself in the Hartree term (see equation (1.20)). This self-interaction is canceled out in approaches that treat exactly the exchange energy such as Hartree-Fock (see equation (1.16)) but remains in approaches approximating the exchange part such as LDA and GGA. This self-interaction error tends to be very problematic for elements containing localized electrons such as d electrons in transition metal oxides.

One popular and computationally cheap way to correct for this self-interaction error is to add a U interaction (as in the Hubbard model to localized orbitals) to the DFT Hamiltonian.[22, 23] This leads to the LDA+U and GGA+U approaches. Following the rotationally invariant implementation proposed by Dudarev,[24] the GGA+U energy is:

$$E^{GGA+U} = E^{GGA} + \frac{U-J}{2}\sum_{\sigma}[(\sum_{m} n^{\sigma}_{n,m}) - (\sum_{m,m'} n^{\sigma}_{m,m'} n^{\sigma}_{m',m})] \tag{1.27}$$

Where E^{GGA} is the energy from a standard GGA functional. The matrix $n^{\sigma}_{n,m}$ is the occupation matrix for a d orbital. The U and J parameters are respectively the on-site Coulomb and exchange parameters. Those two parameters are practically combined in one effective U parameter ($U_{eff} = U - J$). The self-interaction error is here corrected by pushing the system to integer occupations number for the d orbitals which are more physical than the partially occupied results obtained by pure GGA and LDA.

This approach requires to determine this U parameter. U parameters have been determined *ab initio,* [25, 26] but, in this thesis, are adjusted to reproduce experimental data such as enthalpies of formation. [4]

1.6 High-Throughput *ab initio* computations

In the previous section, we have shown how a total energy can be attributed to any crystalline compound using a pseudopotential based plane-wave DFT code. The idea behind high-throughput *ab initio* computing is to evaluate the energy of several thousands of known or predicted compounds using DFT computations.[27, 28, 29, 30, 31, 32] As we will see in the next chapters, this is of interest if one wants, for example, to explore certain chemical spaces by computing new compounds candidates and testing them for stability versus a database of previously known compounds.

While DFT codes are getting more and more stable and user-friendly, they still require a certain amount of human supervision. This occurs even before starting the computations by choosing the appropriate pseudopotentials, U values, and other parameters such as the plane-wave energy cut-off. After the computations is finished, the results still need to be checked for errors or warnings from the DFT code and for sufficient ionic and electronic convergence. Similarly to high-throughput experiments, the challenges lie in the automatization of the humanly performed task. We will now present the approach followed to produce all the high-throughput computations present in this thesis. This work is the result of a collaborative effort and will be published with more details by Jain et al. [33]

1.6.1 Parameters selection

All of our computations have been run in GGA+U with a plane-wave basis code. This type of computations requires to determine a set of parameters.

A U parameter is often required for reproducing valid energetics in transition metal oxides. By evaluating the formation energy from the metal to the oxide for all transition metals, we identified Sc, Y, Ti, Zr, Hf, Zn, Cd and Hg as reproducing formation energies already well enough with GGA to require a U parameter.

For the other transition metals, a procedure to fit the U parameter on experimental enthalpy of formations has been proposed by Wang et al.[4] This procedure consists in identifying a reaction involving a change in oxidation state and varying the U parameter until the computed data agrees with the experiment. For instance, the oxidation of FeO to Fe_2O_3 can be used to find the U parameter for iron. This procedure was used to fit a U parameter for V, Cr, Mn, Fe, Co, Ni, Cu, Nb, Mo, Ag, Ta, and W using the Kubaschewski thermochemical tables. [5] It is assumed that one U parameter per transition metal can be chosen for all computations (i.e. that the U parameter will not vary much between crystal structures and oxidation states).

element	U value (eV)
Ag	1.5
Co	3.4
Cr	3.5
Cu	4.0
Fe	4.0
Mn	3.9
Mo	3.5
Nb	1.5
Ni	6.0
Ta	2.0
V	3.1
W	4.0

Table 1.1: set of U parameters used in this thesis for the high-throughput computations. These U values have been determined following Wang's method by fitting experimental binary oxide formation enthalpy from the Kubaschewski tables.[4, 5]

Table 1.1 provides the U value for each of those transition metals.

The remaining transition metals: Ru, Rh, Pd, Os, Ir, Pt, Au do not produce accurate oxides formation energies by pure GGA but do not have enough reliable thermochemical data for a U to be fitted. Those elements will be not considered in this thesis. We should add that we excluded the radioactive Tc. Finally, Re was not considered as no U could reproduce the experimental oxide formation energies with the PAW pseudopotential provided by VASP.

The plane-wave energy cut-off we used is 30% higher than the maximal energy cut-off specified by the pseudopotential. For oxides, oxygen has the maximal specified energy cut-off of 400 eV, constraining all oxide runs to be performed with 520 eV cut-off energy.

A k-point density of at least 500/(number of atoms in unit cell) k-points was used for all the Brillouin integrations. The Monkhorst-Pack method was used to obtain k-points distributed as much as possible uniformly. [18] A Γ-centered grid was used for hexagonal cells.

While in theory the initialization of the magnetic moments should not matter and the solution with lowest energy should be found, it is common in practice for a system to not reach its global minimum magnetic state but to instead be trapped in a local minimum. Two issues have to be considered in terms of magnetic moment initialization: the spin ordering (ferromagnetic, anti-ferromagnetic, ...) and the magnetic moments magnitude: low-spin or high-spin. For the sake of limiting the computational budget, all computations have been run using ferromagnetic states. Concerning the moment magnitude, high spin initialization (chosen to be 5 Bohr magnetons) is used for the given elements: Sc, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, Y, Zr, Nb, Mo, Ag, Cd, La, Hf, Ta, W, Pt, Hg, Ce, Pr, Nd, Sm, Eu, Gd, Dy, Ho, Er, Tm, Yb, Lu and low spin for all others (chosen to be 0.6 Bohr magneton).

As cobalt can easily be low and high spin, two computations are run for any compound containing this element, one initialized high spin and the other initialized low spin.

1.6.2 Job control scripting infrastructure

To limit human intervention, the aflow software was used as a wrapper around the VASP runs.[34, 27] Aflow catches common VASP errors and restart automatically the runs after performing changes that will tentatively fix the problem. In addition, home-made scripts are used to monitor the convergence of any job. If a job reaches the maximum number of electronic self-consistency iterations the job is aborted and tagged for future restart. This makes sure that non-converging jobs are not wasting computational time. At the end of each run, the results are analyzed automatically for energy and charge convergence and tagged for restart if not converged. The restart process consists in changing the matrix diagonalization scheme used from the fast RMM method [35] to the more robust but slower Davidson algorithm [36] and to restart the job. Self-consistency mixing parameters are also changed during restart.

1.6.3 Database of *ab initio* computations

Due to the amount of data generated during high-throughput computational projects, a careful storage of the data in a database is necessary. In this work, we used a postgresql relational database with an interface through a Java codebase.

This database stores basic information for each computed compound: initial and relaxed crystal structures, chemical formula, total energy computed and DFT parameters. In addition, data requiring some processing of the basic data can be stored such as assigned oxidation states, stability data, and applications specific data such as voltages in the case of lithium ion batteries.

Chapter 2

Ab initio thermodynamical phase stability

> *Thermodynamics is a funny subject. The first time you go through it, you don't understand it at all. The second time you go through it, you think you understand it, except for one or two small points. The third time you go through it, you know you don't understand it, but by that time you are so used to it, it doesn't bother you anymore.*
>
> Arnold Sommerfeld

In this thesis, the phase stability of known and new compounds will be evaluated through *ab initio* computations. This is how we will, for instance, assess if a new compound will be stable with respect to all known other phases. Chapter 1 presented how the energy of crystalline solids can be computed through DFT computations. We will here overview how these *ab initio* energies can be used to compute thermodynamical phase stability. The thermodynamical constructions for closed and open systems will be described along with the different approximations involved.

2.1 Low temperature stability: the convex hull construction

Assessing thermodynamical phase stability in a chemical system requires to compare the free energy of the different phases present.[37, 38] 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:

$$G(N_A, N_B, T, p) = E(N_A, N_B, T, p) + pV(N_A, N_B, T, p) - TS(N_A, N_B, T, p) \quad (2.1)$$

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.[39, 40, 41]

Under these approximations, the relevant thermodynamic potential is the energy. The energy normalized by the total number of particles in the system ($N = N_A + N_B$): $\bar{E}(x_A, x_B)$ 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/atoms.

Solving the Kohn-Sham equation in the density functional theory 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 2-1 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* construction can be performed. The construction consists in finding a

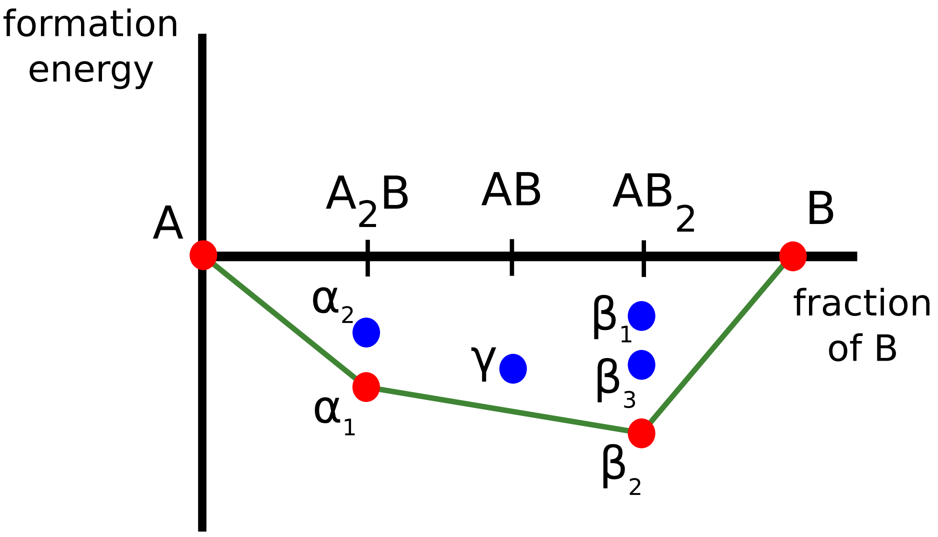


Figure 2-1: Convex hull construction for an A-B system. The points represent different phases. The green line is the convex hull. The red points are the most stable phases or ground states and blue points are unstable phases according to the construction.

convex envelop containing all the points in the plot. This envelop called the convex hull (or hull) is plotted in green in Figure 2-1. The phases present on this convex hull (red points) 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 in 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... The qhull package is a very efficient software performing this construction in any dimension.[42]

2.2 Measure of stability or instability versus the convex hull

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 Figure 2-2.a). 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/at. 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 energy above the hull. It is the 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/at. 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 figure 2-2.b

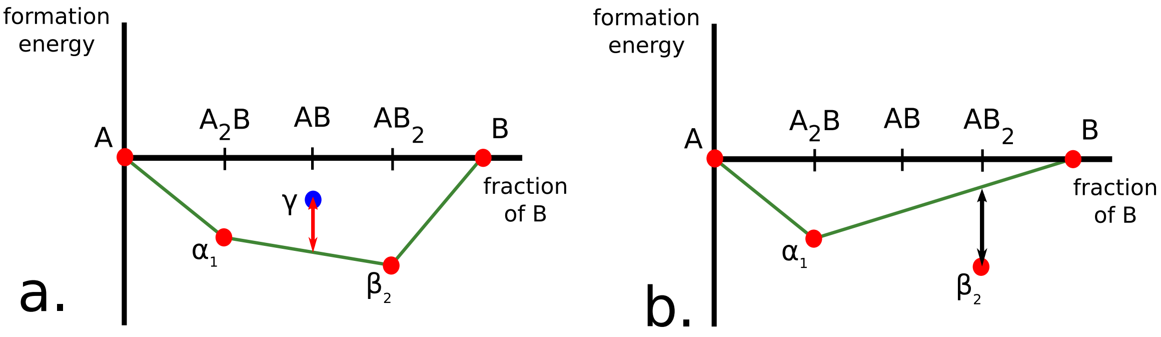


Figure 2-2: Illustration of different measure of stability from the convex hull construction. The energy above the hull is illustrated for the unstable phase γ by the red double arrow in a. The inverse distance to the hull is represented for the stable phase β_2 by the black double arrow in b.

2.3 Stability versus oxygen gas

Many of the compounds studied in this thesis are oxides. 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. However, 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

$$\varphi(N_A, N_B, \mu_O, T, p) \quad = \quad G - \mu_O N_O \tag{2.2}$$

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

$$\bar{\varphi}(N_A, N_B, \mu_O, T, p) \quad = \quad \frac{G - \mu_O N_O}{N}$$

This is a situation very similar to 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.[43] The normalized grand canonical potential is then:

$$\bar{\varphi}(N_A, N_B, \mu_O, T, p) \quad = \quad \frac{E - \mu_O N_O}{N}$$

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. On the contrary, the oxygen chemical potential can be decreased (i.e., be more reducing) by increasing the temperature or lowering the oxygen partial pressure.

In this thesis, all oxygen chemical potentials will be given with reference to the oxygen chemical potential of air at room temperature (298K, 0.21atm). A correspondence between an oxygen chemical potential and an estimated temperature and partial pressure of oxygen can be obtained using the oxygen gas entropy from the JANAF tables.[44] GGA being known to overbind the oxygen molecule, we use the experimentally fitted oxygen molecule energy from Wang et al.[4]

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. Figure 2-3 shows such an oxygen chemical potential stability window for different well known binary oxides. Experimentalists can use such a table as an indicator of how oxidizing or reducing the conditions will need to be for a predicted compound.

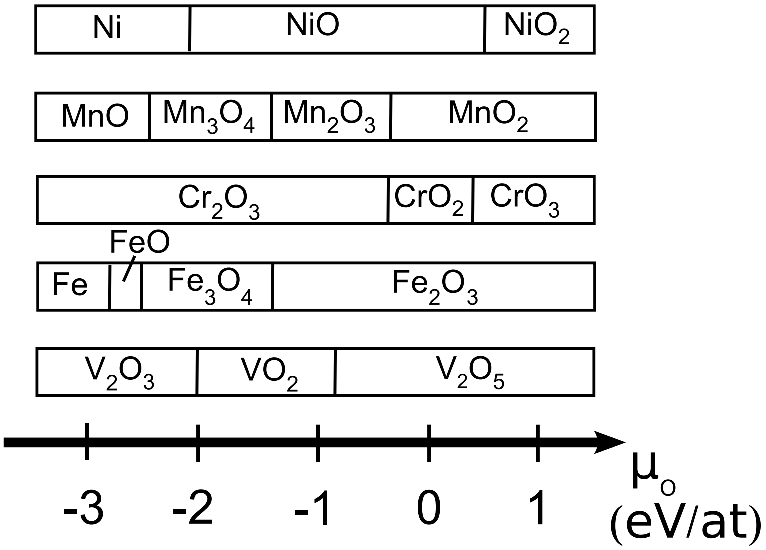


Figure 2-3: Oxygen chemical potential range for some binary transition metal oxides. This graph indicates the range of oxygen chemical potential in which common binary transition metal oxides would be stable. This data has been obtained by DFT computations on the binary oxides present in the ICSD. The zero oxygen chemical potential corresponds here to air at room temperature. Some phases only stable on a very narrow range (i.e. Cr_8O_{21}, Cr_5O_{12}, V_3O_7, V_3O_5, V_5O_9) have been removed for the sake of clarity.

2.4 Mixing GGA+U and GGA computations

The GGA+U method has been presented in chapter 1 as a very efficient way to correct for the self-interaction error present in computations on transition metal oxides. Effectively, GGA+U tends to localize more the *d* electrons than in GGA computations and provide therefore a more physical picture of the bonding in transition metal oxides. On the other hand, in metals the electron delocalization induced by GGA is actually close to the real metallic bonding state and applying a U correction would only deviate the model from the reality. We stand therefore in a situation in which, for transition metals, GGA reproduces sufficiently well the energy in metallic systems but in oxides, only GGA+U does. As computations with two different hamiltonian (GGA and GGA+U) cannot be directly compared, it is impossible to compute energies and then evaluate phase stability when compounds of different nature are involved such as oxides and metals. In this thesis, we will use an approach relying on an energy shift of the GGA energies. This shift is based on a calibration on binary oxides formation energies from the metal. After applying this shift to GGA computed phases, all computed data can be compared and used to assess phase stability. More details about the approach and its efficiency can be found in Jain et al.[45]

Chapter 3

The compound and crystal structure prediction problem

One of the continuing scandals in the physical sciences is that it remains impossible to predict the structure of even the simplest crystalline solids from a knowledge of their composition

John Maddox, 1988

New compound discovery in solid state chemistry is traditionally achieved experimentally through a careful application of chemical principles, intuition, and serendipity.[46] New compounds are postulated, often on the basis of heuristic chemical rules, followed by synthesis, characterization and evaluation of the properties of technological interest. Such a process, while essential to progress in the field, is slow and prone to systematic bias (e.g., by focusing a large degree of effort on a small region in the composition space).

Using the techniques presented in chapters 1 and 2, solid phase stability can be accurately and efficiently predicted through *ab initio* computations in the DFT framework.[27, 47, 43] This offers the opportunity to accelerate the materials discovery process by orienting the experimentalist to computationally predicted compounds of potential interest. However, even with a perfect physical modeling of phase stability, discovering new compounds computationally remains an incredibly challenging task.

In this thesis, we will focus on stability at 0K and 0 pressure as presented in chapter 2. This chapter will introduce the compound and crystal structure prediction problem and will review the different approaches used to address it.

3.1 The crystal structure prediction problem

The crystal structure prediction problem consists in finding the most stable crystal structure at a given composition. For instance, if we consider the $BaTiO_3$ composition, we would expect to predict the most stable structure to be the so-called perovskite structure with octahedrally coordinated titanium atoms sharing edges and barium atoms occupying 12-fold coordinated sites (see figure 3-1).

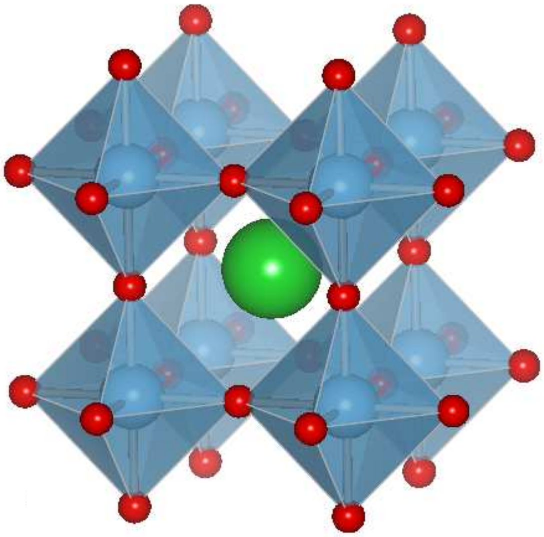


Figure 3-1: the $BaTiO_3$ perovskite crystal structure. Barium atoms are green, titanium in blue and oxygen in red.

3.1.1 Crystal structure prediction as an optimization problem

As pointed out in several recent reviews, the main direction pursued nowadays to address this crystal structure prediction problem consists in approaching it as an optimization problem.[48, 49, 50] The relevant thermodynamic potential at zero temperature and pressure being the energy, solving the crystal structure prediction problem consists in finding the crystal structure with the lowest energy at the targeted composition. Mathematically, this is equivalent to solve:

$$\underset{a,b,c,\alpha,\beta,\gamma,x_1,y_1,z_1,\ldots,x_n,y_n,z_n}{argmin} \{E(a,b,c,\alpha,\beta,\gamma,x_1,y_1,z_1,\ldots,x_n,y_n,z_n)\} \tag{3.1}$$

Where the variables are: a, b, c the three unit cell vector length, α, β, γ the three unit cell vector angles and x_i, y_i, z_i the atomic coordinates of each of the n atoms in the cell and the function E is the energy of the compound in the crystal structure defined by the variables.

There are two important pieces to this problem: a model to evaluate the energy and a method to search this energy function (very often called the *energy landscape*) for a global minimum.

Energy models can vary from the crudest empirical pair potential to the more elaborate *ab initio* quantum mechanics computations. Empirical potentials are much cheaper computationally than *ab initio* computations but are less portable (e.g., they are often accurate only for a few crystal structures).

A very common approach to reduce the size of the energy landscape is to assume a fixed topology by assuming a fixed underlying crystal lattice (e.g., fcc or bcc). Only the specific decorations on this underlying lattice have to be explored then. This approach is very often coupled with the cluster expansion technique.[51, 52, 53] A cluster expansion consists in building a function providing the energy for each single decoration of the lattice. This energy function is expressed as a sum of point, pair, triplets,... interactions called clusters and is fitted using a limited number of *ab initio* computations. When the fitting procedure is completed, the cluster expanded energy can be evaluated extremely rapidly and the space of lattice decorations can be explored for stable structures.[54, 55] The major drawback of using cluster expansion for crystal structure prediction is the need to know a priori the underlying lattice.

In this thesis, our aim will be to solve the crystal structure prediction problem using *ab initio* computations and without any knowledge of the underlying lattice. The principal techniques used therefore in the literature will be reviewed in the next section.

3.1.2 Methods to solve the crystal structure optimization problem

The search for a global minimum on the energy landscape is far from easy. This landscape is indeed very large, complex and present many local minima.[56] This problem has been tackled by many different methods. The two most common ones being simulated annealing and genetic algorithms.

While used in many fields as an optimization tool, simulated annealing is inspired by the metallurgical process in which a material in a metastable state such as a glass or a crystal with many defects is kept at a high temperature for a given time to permit its transformation in the stable crystalline state. In the case of structure prediction, a Monte-Carlo scheme is most often used.[57, 58] A starting atomic configuration is modified by different perturbations: moving individual atoms, exchanging atoms, and changing the unit cell shape. Those perturbations are accepted if they follow certain criteria. At high temperature, perturbations are easily accepted allowing the system to explore the energy landscape, easily jumping above energy barriers. During the optimization, the temperature is decreased slowly with the hope that the system will end up in the global minimum at the end of the procedure. The final minimum found being somewhat sensitive to the starting configuration, it is common to run a few of those simulated annealing runs using different initial configurations. A variant of simulated annealing consists in the basin hopping technique.[59, 60] The algorithm is similar to simulated annealing but every time a new configuration is proposed after a Monte-Carlo move, the system is relaxed by a local minimization technique, such as a conjugate gradient, to the nearest local minimum. Instead of jumping between different points on the energy landscape, the jumps are now between local minima or basins.

Genetic algorithms, on the other hand, are inspired by the biological process of evolution and the idea of the surviving of the fittest. A genetic algorithm consists in computing energies for a set of crystal structure candidates called a *population*. From those energies, we select the lowest energy structures (*selection* step) and combine them through a *mating* procedure to form a new population. Energies are computed for this new population and followed by a new selection and mating process. The repetition of those steps is expected to converge after several generations to the ground state solution by selecting the characteristic of the structure leading to the lowest energy. In addition to mating, it is common to add a mutation step, through some perturbations of the best solutions, to insure some diversity in the structural pool. The mating procedure between crystal structures can consist in cutting the unit cell of both parents in two pieces and combining pieces of both parents to form the unit cell of the child. We should point out that genetic algorithms have many variable parameters that influence dramatically the result of the optimization. Among those are the mutation rate, the initial population, the selection threshold, etc... The first uses of genetic algorithm in the structure prediction field related to atomic clusters.[61] It is nowadays the major optimization tool for periodic systems such as crystal structure, with many variants along the same idea from several authors. [62, 63, 64, 65, 66, 67, 68] A recent study has shown that, for the Al-Sc test system, the genetic algorithm starts to be more efficient than basin hopping in finding the ground state for large unit cells (>10 atoms). [69]

While appealing as quite exhaustive searches, optimization approaches require very significant computational resources, especially for multi-component systems such as ternaries. For example, around a thousand energy evaluations were needed to find the high-pressure ground state for $MgSiO_3$ using a genetic algorithm.[65] They are therefore of interest if one specific composition needs to be studied but are not affordable for large scale searches, especially when *ab initio* energy models are used.

3.1.3 Heuristic or empirical structure prediction rules

The optimization method assumes no previous knowledge (except for the energy model). On the other hand, solid state chemists have been successfully using for long 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). [70]

Another common heuristic approach consists in building structure maps.[71, 72, 73] 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 compound forming this structure. For example, the "NaCl" or "rocksalt" structure prototype is formed not only by NaCl but also by CoO, AgBr etc... (see figure 3-2).

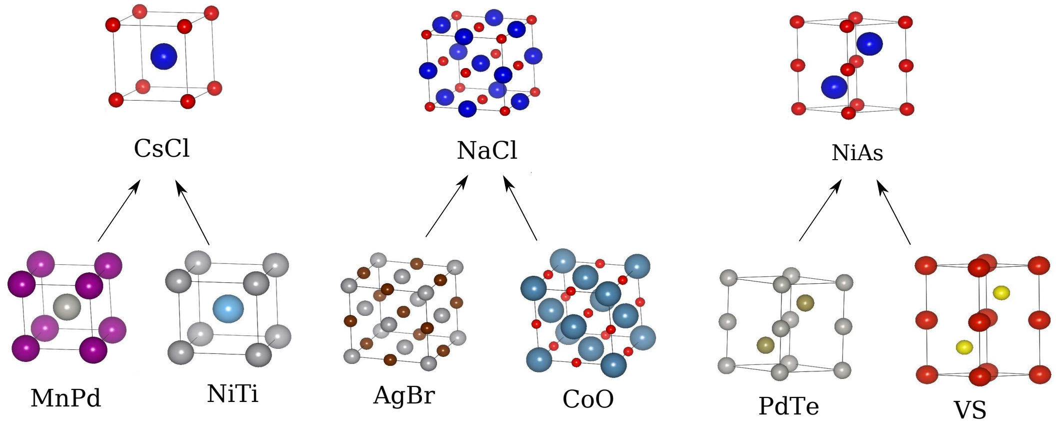
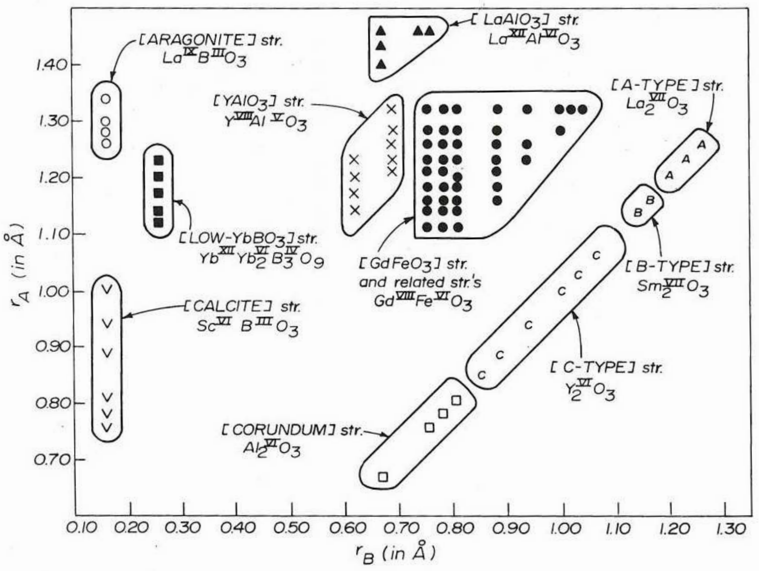


Figure 3-2: some examples of compounds and their crystal structure prototypes

Structure maps are constructed by plotting which crystal structure prototype forms depending on atomic factors of the constituents for many different chemical systems. These factors can be for instance ionic radius or a chemical scale 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. An example of structure map is shown in figure 3-3 for the oxide with ABO_3 stoichiometry. A plot of the ionic radius of A (r_A) vs the ionic radius of B (r_B) for the known ABO_3 oxides shows that crystal structure prototypes are grouped together. When established from known data such structure maps can be use to predict the structure of a new ABO_3 compound knowing the radii of the ion A and B.



Structure field map for $A^{3+}B^{3+}O_3$ compounds

Figure 3-3: structure map for the ABO_3 composition. Adapted from Muller et al.[1]

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., [74] 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.

3.2 The compound prediction problem

Many of the successes of structure prediction occurred for problems where previous experimental knowledge of the composition and lattice parameters was available. Two cases in point are the determination of the previously unknown crystal structure of Li_3RuO_4 and more recently of a new phase of high-pressure boron.[75, 76]

Nowadays, it is common to perform crystal structure prediction without constraint on the lattice parameters. However, the vast majority of the studies present in the literature assumes knowledge of the stoichiometry of interest. We will call the problem of finding a new compound (composition and crystal structure) the *compound prediction* problem by opposition to the crystal structure prediction problem consisting in finding the crystal structure at a fixed composition.

To our knowledge, only one paper from Trimarchi et al. addresses the compound prediction problem in an optimization framework by proposing a genetic algorithm without composition constraint.[67] This algorithm still needs to be proven to be scalable to the much larger compositional space of multi-component systems (i.e., ternaries or quaternaries). Most of the empirical approaches also do not address the full compound prediction problem. For instance, structure maps can only provide likely structures at a fixed composition.

3.3 Data mined approaches to compound and crystal structure prediction

As we presented in the previous sections, the main compound and crystal structure prediction techniques available are either quite unbiased and very expensive computationally, or knowledge-driven, cheap computationally but more qualitative. Recently, there have been efforts to combine rigorous knowledge-based methods used in the field of data mining, machine learning or statistical learning with the accuracy of DFT to solve the crystal structure prediction problem.[7]

Data mining or machine learning consists in using large data sets to mathematically extract patterns and rules.[77, 78] As human beings can perform this type of inference, machine learning is very often associated with the field of artificial intelligence. With the availability of computers and large, easy to query, data sets, data mining techniques are used nowadays in many fields: astrophysics, biology, fraud detection, movie recommendation systems, etc... Some common machine learning techniques are linear regression, Bayesian inference, support vector machine, neural networks,...

Curturalo et al. pioneered the use of data mining techniques in the crystal structure prediction field.[79] They used advanced linear regression analysis to predict energies of not yet computed compounds. Using a database of compounds already computed by DFT, the energy of a compound not yet computed was shown to be predictable by linear regression. Subsequently, Fischer et al. proposed to use data from an experimental crystal structure database to extract the chemical rules of structural stability using correlation between crystal structures at different composition.[80]

In the next chapter, we will present the technique developed by Fischer and show how its combination with high-throughput *ab initio* computations can be used to predict with minimal computational budget hundreds of new ternary oxides.

Chapter 4

Finding new ternary oxides by combining data mining and *ab initio* computations

Prediction is very difficult,
especially about the future.
Niels Bohr

In this chapter, we will show that new compounds can be predicted computationally on a large scale with a reasonable computational budget. Using a combination of a data mining algorithm with high-throughput DFT computations, we discovered around a hundred new ternary oxides. After presenting the algorithm based on based on correlations between co-existing crystal structure prototypes, the specific mathematical implementation is described. Finally, the results of the new ternary oxides search are presented and analyzed.

4.1 General principle of the algorithm

As presented in 3.1.3, crystalline inorganic compounds form in a limited set of crystal structure prototypes. 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 exist 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 observations, 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 section, we will present 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.

4.2 The data abstraction

In this chapter, we will assume that a prototype label has been assigned to all the compounds in the database. This prototyping step can be fully automatized and is discussed in more details in appendix A. After transformation of the raw database to a prototyped database. The data available will be in the form of a composition-crystal structure prototype pair for each compound.

composition	crystal structure prototype
CoO	NaCl
AgBr	NaCl
NaCl	NaCl
Al_2MgO_4	Al_2MgO_4
$LiMn_2O_4$	Al_2MgO_4
$FeAl_2O_4$	Al_2MgO_4

Table 4.1: An example of data present in a prototyped crystal structure database

We will use for the sake of simplicity 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 this composition bin 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 $x_{c_i} = nostructure$ 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 $\mathbf{X} = (x_{c_1}, x_{c_2}, \ldots, x_{c_n}, x_{E_1}, x_{E_2}, ..., x_{E_C})$ 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 figure 4-1).

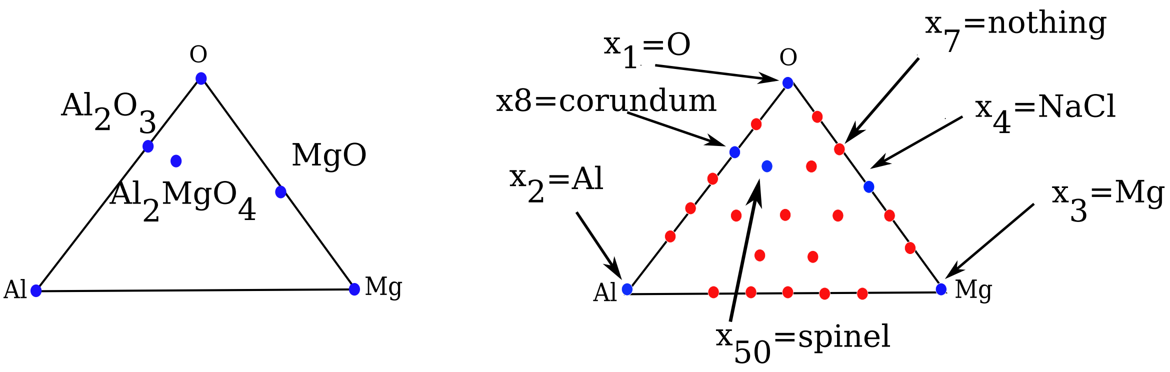


Figure 4-1: An example of how the information on the Al-Mg-O chemical system is projected onto the composition variables. All dots indicate composition bins. Red dots are composition bins without any known compound and blue dots are composition bins with a known compound crystallizing in a specific prototype marked by an arrow.

Any prototyped crystal structure database $\mathcal{D}$ can be represented in our formalism a collection of N $\mathbf{X}_\mathrm{i}$ instances, $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$

4.3 The probabilistic function and the new compound discovery procedure

The probability density $p(\mathbf{X})$ is the essential information we require as it 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:

$$p_{compound}(c_j) = \\ 1 - p(x_{c_j} = nostructure | x_{c_1}, x_{c_2}, ..., x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, ..., x_{E_1}, x_{E_2}, ..., x_{E_C}) \quad (4.1)$$

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

$$p(x_{c_j}|x_{c_1}, x_{c_2}, ..., x_{c_{j-1}}, x_{c_{j+1}}, \ldots, x_{c_n}, ..., x_{E_1}, x_{E_2}, ..., x_{E_C}) \tag{4.2}$$

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 Figure 4-2.

We should stress that, to the contrary of most optimization techniques, this approach can therefore 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.

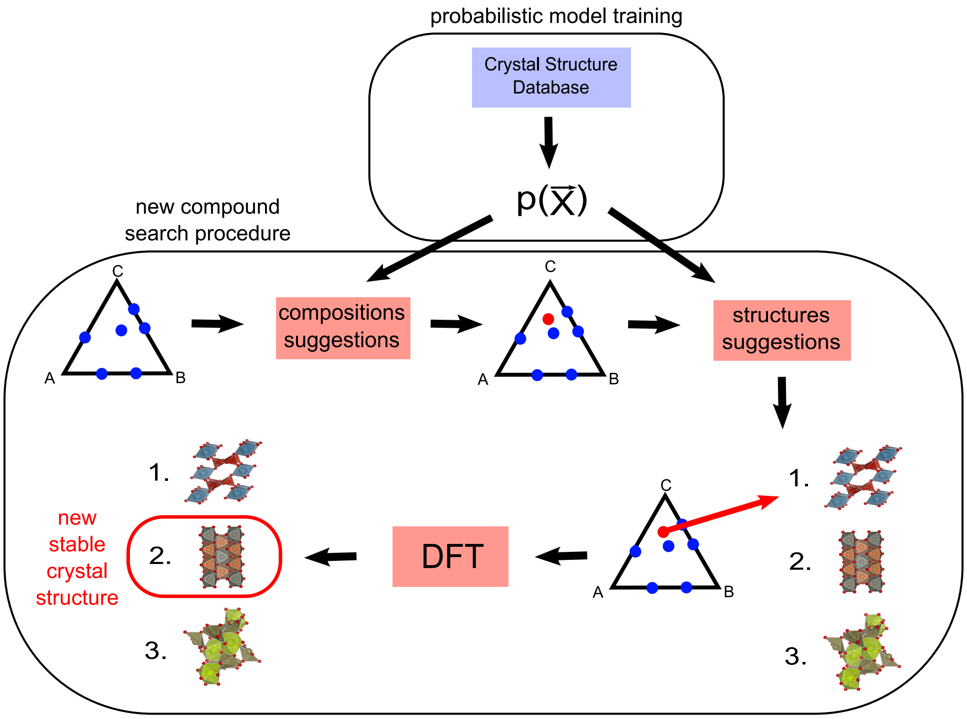

probabilistic model training
Crystal Structure Database
$p(\vec{X})$
new compound search procedure
C
A
B
compositions suggestions
structures suggestions
1.
2.
3.
DFT
new stable crystal structure

Figure 4-2: 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.

4.4 The 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 around 100 crystal structure prototype 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 mechanics as the cumulant expansion.[81] The cumulant expansion can be presented starting with the identity:

$$p(\mathbf{X}) = \prod_i g_i(x_{c_i}) \prod_{j<k} g_{jk}(x_{c_j}, x_{c_k}) \prod_{l<m<n} g_{lmn}(x_{c_l}, x_{c_m}, x_{c_n}) \dots \tag{4.3}$$

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 have trivially:

$$g_i(x_{c_i}) = p(x_{c_i}) \tag{4.4}$$

A two variables probability function will be

$$p(x_{c_i}, x_{c_j}) = p(x_{c_i})p(x_{c_j})g_{ij}(x_{c_i}, x_{c_j}) \tag{4.5}$$

Implying that

$$g_{ij}(x_{c_i}, x_{c_j}) = \frac{p(x_{c_i}, x_{c_j})}{p(x_{c_i})p(x_{c_j})} \tag{4.6}$$

The general form for a cumulant over the variable X_α is:

$$g_\alpha(x_\alpha) = \frac{p(x_\alpha)}{\prod_{\beta \subset \alpha} g_\beta(x_\beta)} \tag{4.7}$$

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

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

$$p(\mathbf{X}) = \frac{1}{Z} \prod_i p(x_{c_i}) \prod_{j<k} \frac{p(x_{c_i}, x_{c_j})}{p(x_{c_i})p(x_{c_j})} \tag{4.8}$$

Where Z is a normalization constant.

4.5 Estimating the probabilistic function from available data

Having decided on the form of an approximated probability function (equation (4.8)), we still need to estimate the values of this function parameters. Using a database $\mathcal{D}$, we will search for the values $p(x_{c_i}, x_{c_j}|\mathcal{D})$ and $p(x_{c_i}|\mathcal{D})$ in best agreement with the data. One can see this process called parameter estimation as a fit of the model to the available data.

In this section, 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 purpose, we will first present derivations for the single variable case and will generalize later to the multi-variable case.[82]

4.5.1 Single variable multinomial parameter estimation by maximum likelihood

Let us assume a random variable X that can take on n possible values $x \in \{v_1, v_2, ..., v_q\}$. Assuming we have a database $\mathcal{D}$ of N observed values for $\mathcal{D} = \{x_1, x_2, ..., x_N\}$, we would like to infer the probability function $p(x|\mathcal{D})$. For each of the q values possible of X, we assign a parameter with the value of the probability function. We have then 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.[83] The best estimate for the parameter is the one maximizing the (log)-likelihood of the data l.

$$\begin{aligned}
l(\mathcal{D}, \boldsymbol{\theta}) &= log\, p(\mathcal{D}|\boldsymbol{\theta}) \\
&= log\, p(x_1, x_2, ..., x_N|\boldsymbol{\theta}) \\
&= \sum_{t=1}^{N} log\, p(x_t|\boldsymbol{\theta}) \qquad (4.9) \\
&= \sum_{x} n(x)\, log\theta_x
\end{aligned}$$

This derivation has been performed assuming that all the x_i observations are independent. $n(x)$ indicates the number of time the value x is observed in the data $\mathcal{D}$. Maximizing the likelihood function in 4.9, under the constraint that $\sum\limits_{x}\theta_x = 1$, can be shown to lead to

$$\theta_x^{ML} = \frac{n(x)}{\sum\limits_{x'} n(x')} \tag{4.10}$$

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.

4.5.2 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 $p(\boldsymbol{\theta}|\mathcal{D})$. As an illustration, if one is observing a coin toss leading to 1001 heads and 999 tails, a maximum likelihood approach would find out that the probability for heads should be 0.5005. A Bayesian approach on the contrary 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 $p(\boldsymbol{\theta}|\mathcal{D})$ peaked on 0.5005 but allowing some spread and non-zero values for values close to 0.5005. A rather complete presentation of the Bayesian approach to probability can be found in Jaynes[84].

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:

$$p(x|D) = \int p(x|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}, \mathcal{D}) d\boldsymbol{\theta} \tag{4.11}$$

The parameters θ_x are now defined as

$$\theta_x = p(x|\boldsymbol{\theta}, \mathcal{D}) \tag{4.12}$$

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

$$
\begin{aligned}
p(\boldsymbol{\theta}|D) &= p(\mathcal{D}|\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{p(\mathcal{D})} && (4.13)\\
&= p(x_1, x_2, ..., x_N|\boldsymbol{\theta})\frac{p(\boldsymbol{\theta})}{p(x_1, x_2, ..., x_N)} && (4.14)\\
&= \lambda\prod_x \theta_x^{n(x)} p(\boldsymbol{\theta}) && (4.15)
\end{aligned}
$$

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

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 reason is the Dirichlet distribution:

$$p(\boldsymbol{\theta}) = \beta(\boldsymbol{\alpha})\prod_x \theta_x^{\alpha_x - 1} \quad (4.16)$$

Where $\beta(\boldsymbol{\alpha}) = \frac{\Gamma(\sum\limits_x \alpha_x)}{\prod\limits_x \Gamma(\alpha_x)}$ and Γ is the Gamma function.

Plugging the Dirichlet prior (equation (4.16)) in the expression of the posterior (equation (4.14)), we have:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \lambda\beta(\boldsymbol{\alpha})\prod_x \theta_x^{n(x)+\alpha_x - 1} \quad (4.17)$$

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 solved yet is the value of λ. We can use the normalization condition $\int p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} = 1$. Applying this constraint, it can be shown that

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathcal{D}) &= \Gamma(\sum_{x'} n(x') + \alpha_{x'})\prod_x \frac{\theta_x^{n(x)+\alpha_x - 1}}{\Gamma(n(x) + \alpha_x)} && (4.18)\\
&= C(n, \boldsymbol{\alpha})\prod_{x'} \theta_{x'}^{n(x')+\alpha_{x'}} && (4.19)
\end{aligned}
$$

When the part of the expression involving the Gamma function has been regrouped for notation purpose in $C(n, \boldsymbol{\alpha})$.

Now that we found the expression for $p(\boldsymbol{\theta}|\mathcal{D})$, we can evaluate the probability to observe a value v_i for the variable X.

$$\begin{aligned} p(x = v_i) &= \int \theta_{v_i} p(\boldsymbol{\theta}, D) d\boldsymbol{\theta} && (4.20) \\ &= C(n, \alpha) \int \theta_{v_i} \prod_{x'} \theta_{x'}^{n(x')+\alpha_{x'}} d\boldsymbol{\theta} && (4.21) \\ &= \frac{n(v_i) + \alpha_{v_i}}{\sum_{x'} n(x') + \alpha_{x'}} && (4.22) \end{aligned}$$

This final expression can be compared to the one obtained using the maximum likelihood (equation (4.10)). The way the prior influence 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 is very few data points, the prior will drive the probability.

While we have chosen the Dirichlet prior, we have still 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 available, [85, 86] there is a common choice of prior called the minimum information uniform Dirichlet prior consisting in setting α such as:

$$\alpha_x = \frac{1}{q} \tag{4.23}$$

4.5.3 Generalization to multiple variables

The results presented in the two previous sections can be generalized for multiple variables. Let say we have two variables X and Y and we want to estimate $p(x, y|\mathcal{D})$. $\mathcal{D}$ refers to a set of N observations $\mathcal{D} = \{(x, y)_1, (x, y)_2, ..., (x, y)_N\}$. 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 be directly used then with a multinomial defined on qr values. The maximum likelihood will be

$$\theta_{x,y}^{ML} = \frac{n(x, y)}{N} \tag{4.24}$$

The Bayesian estimate is

$$p(x = v_i, y = w_j|\mathcal{D}) = \frac{n(v_i, w_j) + \alpha_{v_i,w_j}}{N + \sum\limits_{x,y}\alpha_{x,y}} \tag{4.25}$$

And the minimum information Dirichlet prior is

$$\alpha_x = \frac{1}{qr} \tag{4.26}$$

4.6 Finding nature's missing ternary oxides

The previous sections presented how a probabilistic model incorporating correlations between crystal structures at different compositions can be built and used for structure prediction. In this section, we apply this approach to the discovery of previously unknown ternary oxides.

To search for new ternary oxides, we estimated a cumulant expansion probabilistic model (equation (4.8)) using the oxide experimental data available in the Inorganic Crystal Structure Database (ICSD, [87]) and the Bayesian estimation procedure presented in 4.5. The ICSD is the most complete crystal structure database for oxides. The 2006 version of the ICSD was searched for duplicate compounds using a an affine mapping technique described in Appendix A. Two entries are considered representing the same compound if their crystal structure have the same space group and can be transformed onto one another through an affine mapping. After this analysis, 616 unique binary and 4747 ternary oxides compounds were identified. These compounds were grouped by crystal structure prototype using the same affine mapping technique. 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, in total, 183 variables were used in the probability model.

4.6.1 Cross-validation test

We tested the compound discovery procedure outlined in the 4.3 using a classic cross-validation approach whereby some information is removed from the database and the quality of the predictions on this removed information is evaluated.[88] The cross-validation procedure consisted in grouping all the possible ternary oxides chemical systems into 20 random groups. Every cross-validation step involved removing all the systems included in one group from the database (the test set), estimating the model with the remaining groups (the training set) and testing how well the model performed in predicting back the compounds present in the test set. Each of these steps was repeated 20 times ensuring that every possible ternary oxide chemical system was present once in the cross-validation test. The cross-validation tests excluded structural prototypes unique to one composition as our data mining method can never predict such crystal structures.

A first cross-validation was performed to test the capability of the model to discover compound forming compositions using the compound probability $p_{compound}$ (equation (1)). This is a binary classification problem. The classifier must be able to discriminate between compound forming and non-compound forming compositions. A classical way to evaluate such a classifier is to plot a receiver operating characteristic (ROC) curve [89]. A ROC graph plots the true positive rate (i.e. the hit-rate) as a function of the false positive rate (i.e., the false alarm rate). Figure 4-3 plots in blue this ROC curve from the cross-validation of our model on the ternary oxides ICSD. The dashed red line indicates the random guess discrimination line. Any binary classifier above this line is more efficient than a random guess. The perfect classifier (false positive rate equals to zero and true positive rate to one) is in the upper left corner and indicated by a red dot. The position of the blue curve compared to the random guess discrimination line indicates the predictive power of our model in terms of compound forming compositions discovery.

Having tested the method to suggest compound forming compositions we focus on the structure prediction component. Here, the prediction consists of computing from the probabilistic model the l most likely structure prototypes for the given composition. Using the same 20-fold cross-validation procedure, a structure prediction was performed for all compositions for which a compound exists in the ICSD. Figure 4-4 shows the crystal structure prototype list length l needed to predict the correct ground state with a given probability during this cross-validation. The different curves are plotted for compositions sets that pass different $p_{compound}$ thresholds. Not surprisingly, the two steps of the compound search (composition and structure prediction) are linked. The list length needed depends on the $p_{compound}$ value. Compounds at very likely compositions (i.e. with high $p_{compound}$ values) generally need less candidates structures. The dashed line on the figure shows that a high success rate (95%) can be achieved with a reasonable number (between 4 and 20) of candidates.

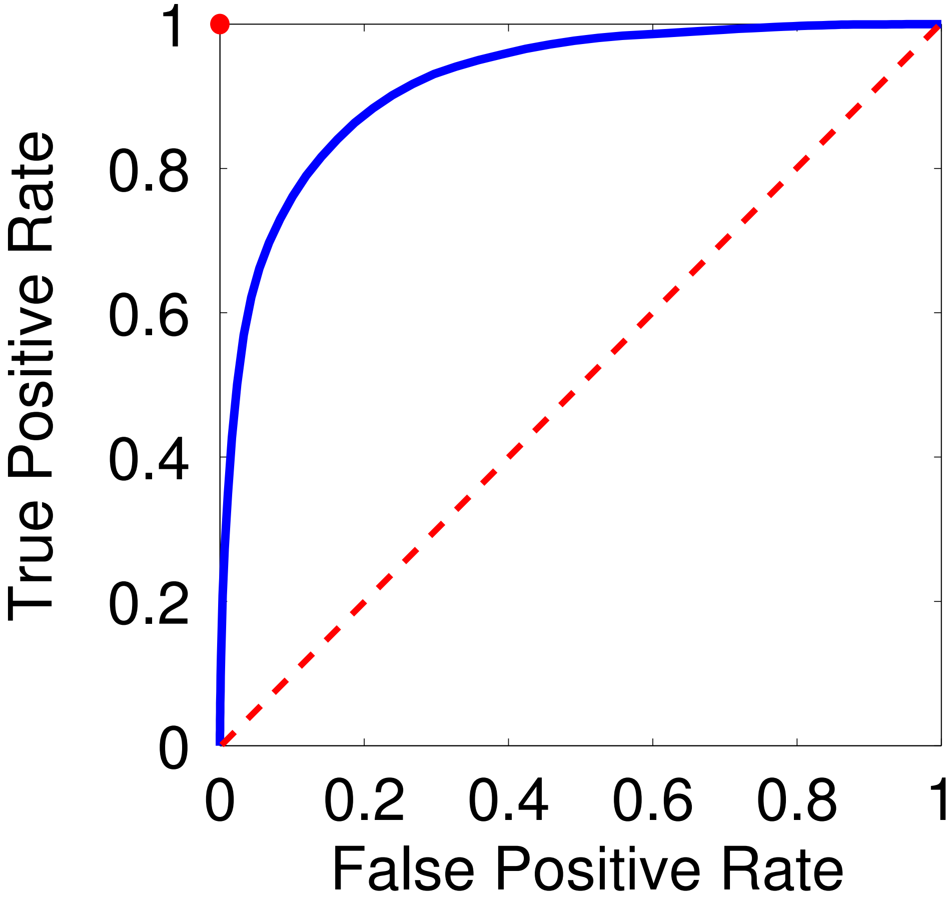


Figure 4-3: Receiver operating characteristic (ROC) plot for the compound forming composition predictions. The blue curve indicates the ROC curve obtained from our model during cross-validation on the ternary oxides ICSD. The dashed red line is the random guess discrimination line. Any data point above it belongs to a classifier more efficient than a random guess. The perfect classifier (false positive rate equals to zero and true positive rate to one) is in the upper left corner and indicated by a red dot.

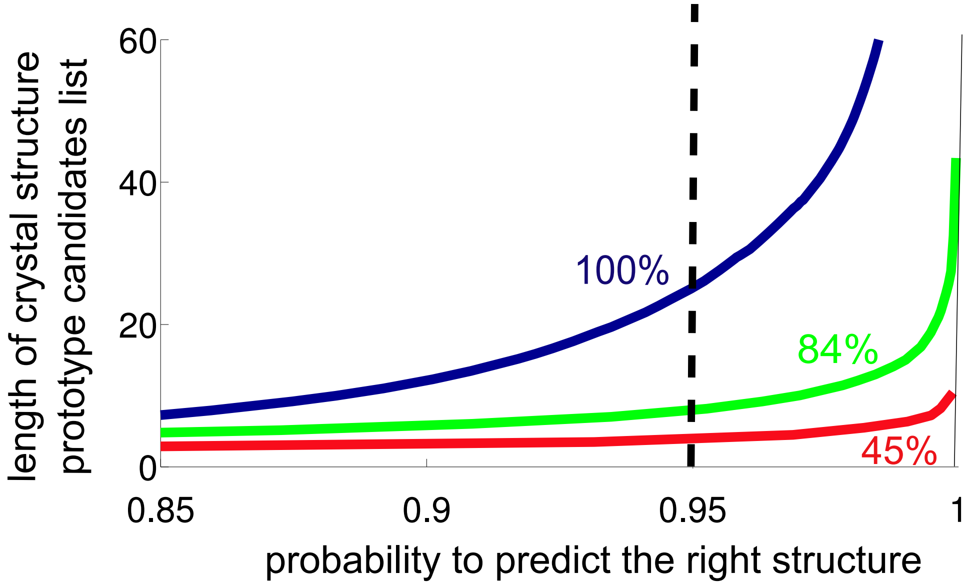


Figure 4-4: Cross-validation result for the prediction of the crystal structure at a compound-forming composition in the ICSD. Crystal structure candidate list length needed versus a certain probability to predict the right structure during cross-validation. The different curves correspond to different threshold for $p_{compound}$. Only compositions with a $p_{compound}$ high enough to recover 100%, 84% and 45% of the known ICSD compound forming compositions during Cross-Validation are taken into account in the corresponding curve. Compositions with higher $p_{compound}$ need fewer candidates crystal structure to find the right one. The dashed line shows that the method can predict with high success rate (95 percent) the right crystal structure prototype for reasonable number (between 4 and 20) of candidates.

4.6.2 New ternary oxides predictions

We searched then 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, or Lu. In these systems, we searched for compositions where no ternary oxide is given in the ICSD but for which the probability for forming a compound (equation (4.1)) 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 equation (4.2). 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* density functional theory. 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 chapter 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. This represents one 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 fully 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 4-5 indicates in blue how many new ternary oxides 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 1970's. The 355 new compounds suggested were obtained with about 55 days of computing on 400 Intel Xeon 5140 2.33Ghz cores. Those numbers show the potential for accelerating new compound discovery through combining data mining with DFT computations.

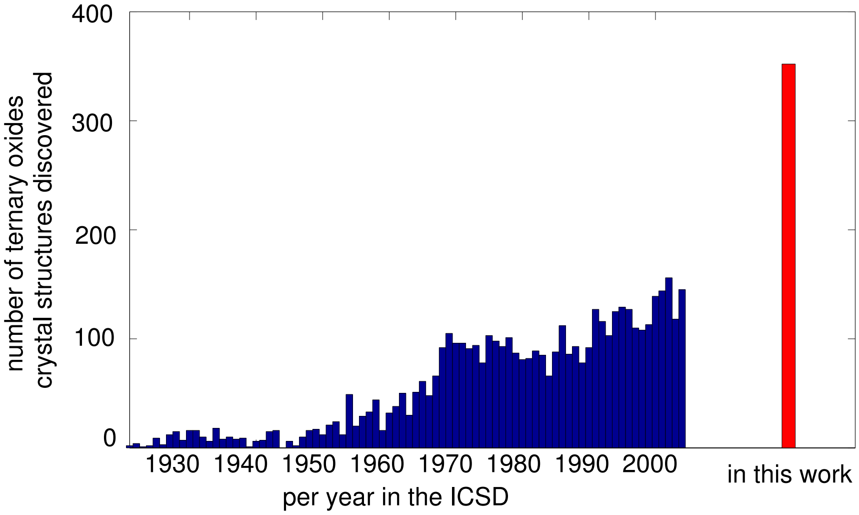


Figure 4-5: New ternary oxide discovery per year according to the ICSD. The blue bars indicate how many new ternary oxides were discovered per year from 1930 to 2005. The red bar shows how many new compounds have been discovered in this work.

While these predictions are made for chemical compositions which have no compounds in the ICSD database, we compared them to another major compound database available: the PDF4+ database from the International Center of Diffraction Data (ICDD).[90] This database contains compounds for which only a composition and an x-ray powder diffraction pattern are available but no structural information. Sixty four of the 355 compositions at which we predicted new compounds have a powder diffraction pattern in the PDF4+ database but without any crystal structure available. For those, our findings complete the current information available by assigning a stable crystal structure to the compound. In addition, 146 other predictions involve compositions presenting structural information in the PDF4+ database (but not in the ICSD).

A careful comparison between the structural data present in the PDF4+ database and our prediction has been carried out. The two structures (the predicted one and the one present in the PDF4+ database) have been compared using the affine mapping comparison scheme presented in Appendix A. Among the 146 predicted compounds with structural information in the PDF4+ database, 107 present a similar structure in the PDF4+ database. Eleven structures presented partial occupancy in the PDF4+ database and could not be compared directly to the predictions. For 22 compositions, while not the ground state structure according to DFT, the structure present in the PDF4+ was in the candidate list and had been computed. The vast majority of the PDF4+ structures lie close enough in energy (between 3 and 20 meV/at) to expect entropy or kinetics effects to explain their difference to the DFT ground state. However, one entry: $FeSnO_3$ presented a PDF4+ structure 590 meV/at above the ilmenite ground state we found. After verification, we discovered that this entry (04-008-6519) presented a transcription error while imported from the original paper it referred to[91]. This illustrates how combining crystal structure data mining with DFT can also help discovering errors present in crystal structure database.

Six compounds in total presented a crystal structure in the PDF4+ not present in any of the candidates we tested by DFT. A first set consists of the HoBrO, TmBrO, PrBrO and CeBrO compounds. The failure of the method to identify the adequate crystal structure comes here from the absence of this crystal structure prototype in our data set. On the other hand, two compounds LuP_3O_9 and DyP_3O_9 did have the PDF4+ crystal structure available in the data set (crystal structure prototype from VP_3O_9) but the probabilistic model suggested other structures. Interestingly, the predicted structure are found to be 9 meV/at for the Lu compound and 6 meV/at for the Dy compound lower in energy than the structure present in the PDF4+.

The different chemistries over which these new compounds are distributed can be analyzed. Figure 4-6 indicates the number of new compounds found for every A-B-O system where A is plotted on the x axis and B on the y axis. The elements are ordered according to their Mendeleev number.[72] This ordering allows us to directly spot the different chemical classes in which new compounds have been found. Figure 4-6 indicates that the predictions span many different chemistries. Most striking is the absence of any new compounds in mixtures of rare-earths. The difficulty to form energetically favorable ternary oxides containing two rare-earths relates to the important electrostatic component to phase stability in oxides. Indeed, many of the rare-earth compounds usually exhibit a +3 oxidation state and combining isovalent cations rarely leads to strong compound formation. Rather, solid solution mixing tends to be more common. Supporting that analysis is the fact that the only prediction we made in this chemistry is a $La_2Pr_2O_7$ compound combining La^{3+} and Pr^{4+} ions. Pr, along with Ce and Tb, are the only rare-earths exhibiting a +4 oxidation state in oxides.

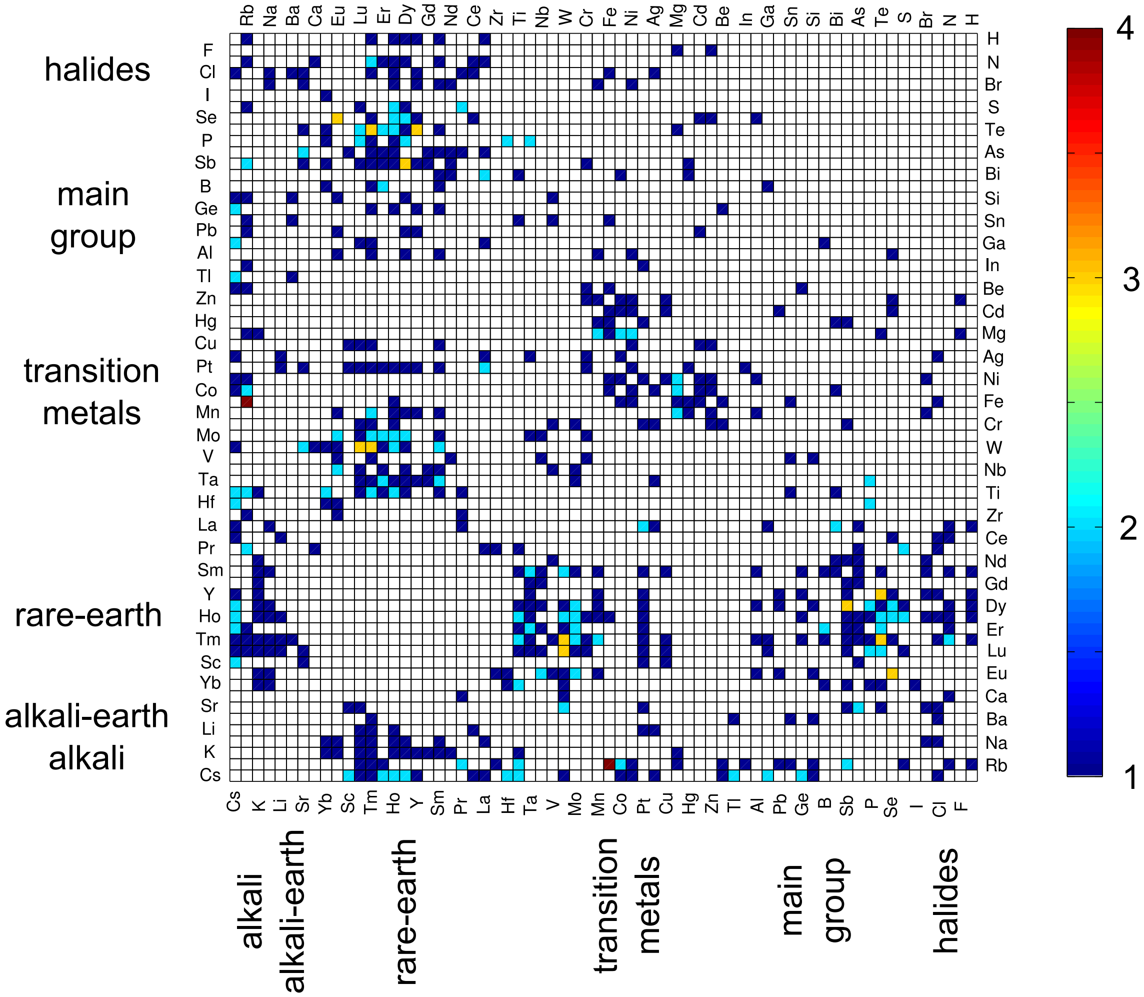


Figure 4-6: Distribution of the new compounds across chemical classes. This plot indicates the number of new compound discovered in this work for any A-B-O system with A along the x axis and B along the y axis. The elements are ordered according to their Mendeleev number.

A similar electrostatic effect can explain the absence of any predictions in the alkali-alkali and alkali-earth-alkali-earth corner. It is interesting to note that the ICSD also shows a lack of compounds in these two regions. The known ICSD compounds in these spaces are mostly disordered solid solution structures stabilized by entropic mixing effects.

The last major region without predictions is situated in the upper right corner of figure 4-6 and concerns oxides of two main group elements. Only 3 successes on 40 suggested compositions were obtained in this chemical space. However, analysis of the cross-validation results did not show any evidence suggesting a systematic failure of our approach in this region. Both the composition and structure prediction procedure did not perform worse for main group elements mixture compared to the whole ICSD. The composition prediction showed a 52% false positive rate vs 45% for the whole database. The structure prediction reached a 100% success rate with a list length of 4 vs 95% for the whole database. The difficulty of finding new main group-main group compounds is probably an indicator that almost all ternary oxides have already been discovered in this chemical space.

Many of the new compounds contain at least one rare-earth element mixed with any of the five other categories of elements (217 among the 355 predicted compounds). Indeed, rare-earth chemistry has not been explored as much as other chemistries and therefore presents more opportunities for our approach to identify unknown compositions and their crystal structures. Rare-earth compounds are however of great scientific and technological interest due to their unique catalytic, optical, or magnetic properties.[92] Among those, it is interesting to see that our algorithm picked up the correct $La_2Zr_2O_7$ structure for the $La_2Bi_2O_7$ compound present in the PDF4+ database.[93] Bi and Zr do not have very close chemical identity and a direct substitution from Zr to Bi would not be obvious to the chemist's intuition.

Another singular chemistry for which we predict many new compounds is the heavy alkali chemistry (Rb and Cs). Indeed, 56 of the 355 compounds contain one of these two elements. Here again, the fact that these two elements have been much less studied than for instance the lighter alkali elements opened up the possibility for discovery of many new compounds. Among these, it is striking to find many Co^{4+} compounds. While Co^{4+} requires very strong oxidation conditions in most chemistries, our analysis of the oxygen chemical potential in which the compound is stable indicates that it may be much easier to form Co^{4+} when combined with Cs or Rb. Table 4.2 compares the value for the oxygen chemical potential range for the predicted compounds with that of a simple CoO_2 binary compound (see chapter 2). CoO_2

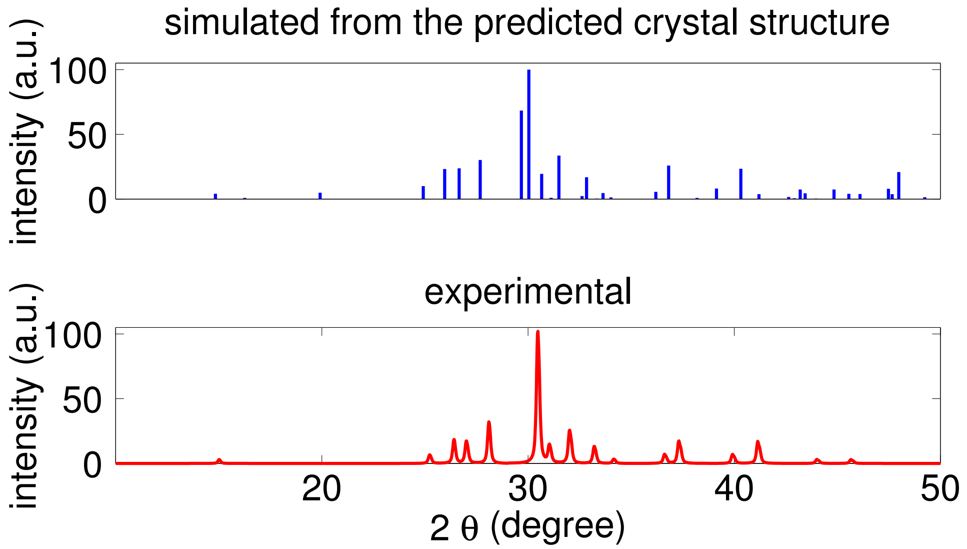


Figure 4-7: Comparison between the experimental and predicted powder diffraction pattern for $CoRb_2O_3$. The experimental pattern from the PDF4+ database is below in red. The simulated pattern from our predicted crystal structure is above in blue.

would not be stable in typical solid state synthesis conditions such as gaseous oxygen at 800C ($\mu_O \simeq -1.0$ eV/at). This is consistent with the impossibility to prepare this Co^{4+} binary oxide in this way. On the other hand, all the predicted heavy alkali Co^{4+} compounds should be stable in these conditions showing the stabilization effect of the heavy alkali. For instance, predicted compounds such as $Co_2Cs_6O_7$, $Co_2Rb_6O_7$ and $CoRb_2O_3$ should be synthesizable in relatively mild oxidizing conditions. The stabilizing effect of heavy alkali towards high oxidation state of a transition metal is a chemical rule known by some chemists and reported for instance by O'Keefe. [94] In addition, one of those three compositions ($CoRb_2O_3$) is linked to a powder diffraction pattern (without exact structural information) in the PDF4+ database (00-027-0515,[95]). Comparison between this PDF4+ pattern and the one simulated from our prediction shows a good agreement (see Figure 4-7).

compound	minimum μ_O	maximum μ_O
CoO_2	-0.04 eV/at	$+\infty$
$Co_2Cs_6O_7$	-2.9 eV/at	$+\infty$
$Co_2Rb_6O_7$	-3.1 eV/at	$+\infty$
$CoRb_2O_3$	-1.7 eV/at	$+\infty$

Table 4.2: Oxygen chemical potential range for different Co^{4+}compounds.

While searching for new compounds obviously results in a higher discovery rate for less common elements, our search also identified compounds in well-studied chemistries. Many of these new compounds, while exhibiting common elements, contain them in an oxidation state requiring such strongly oxidizing or reducing conditions that it could be possible that their synthesis was never attempted in favorable conditions. For instance, while it is surprising to make a prediction in a rather common chemical system such Ni-Zn-O, the predicted compound (a Ni^{3+} containing Ni_2ZnO_4 spinel), require a more oxidizing environment than the Ni^{2+} containing NiO-ZnO solid solution present in the ICSD.[96]

Similarly, while many SnO_2-TiO_2 solid solutions are in the ICSD, there is no known Sn^{2+}-Ti^{4+} compound. Our analysis predicts a $SnTiO_3$ ilmenite stable in reducing environment. 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 ferroelectric material.[97] However, our work shows that the ilmenite structure is more stable than the perovskite structure by a very significant 130 meV/atom.

We should emphasize that we also found compounds for which the extreme synthesis conditions, or the less-common element argument cannot apply to rationalize their presence in our list of novel predictions. For example, we predict two new compounds in the Mg-Mn-O system: $MgMnO_3$ and $Mg_2Mn_3O_8$, both compounds with a much more common +4 oxidation state for Mn. The PDF4+ database actually contains a powder diffraction pattern for $MgMnO_3$. Comparison of the powder diffraction pattern of our predicted structure to the one in this database shows a reasonable match and confirms our prediction of an ilmenite structure (00-024-0736,[98]). Lattice constants computed by GGA are typically a few percent off from the experimental lattice constants. As shown in 4-8, only one major peak in the 50 degree region does not match the powder diffraction pattern simulated from our prediction. This missing peak could be due to some disorder present in the experimental sample. On the other hand, the $Mg_2Mn_3O_8$ compound is totally unknown from literature and is predicted to crystallize in a very uncommon structure exhibited by $Co_2Mn_3O_8$.[99] This illustrates how our method can select unusual or less common structures and go beyond the "trying the usual suspect" approach often used in crystal structure prediction.

All of the compounds predicted have been analyzed separately and a full description, including thermodynamic data, cif file and remarks are available at `http://www.materialsgenome.org/ternaryoxidepredictions`. Researchers could benefit from access to theoretically predicted compounds in addition to the standard crystal structure and powder diffraction experimental database.

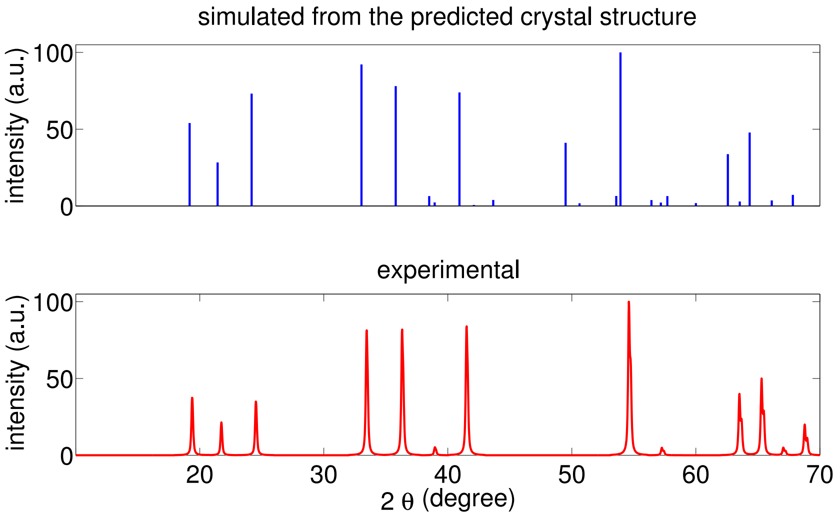


Figure 4-8: Comparison between the experimental and predicted powder diffraction pattern for $MgMnO_3$. The experimental pattern from the PDF4+ database is below in red. The simulated pattern from our predicted crystal structure is above in blue.

While our approach is fast and efficient for discovering new compounds, it has limitations. It is possible that we missed a true ground state due to the absence of its structure prototype from our database. By definition, our method cannot predict a compound crystallizing in an unknown structure prototype. However, finding a new stable compound through our method, while not guaranteeing to find the true ground state, indicates that in any case there is a stable unknown compound lying in this chemical system.

Chapter 5

Data mined ionic substitutions for the discovery of new compounds

The most important questions of life
are indeed, for the most part, really
only problems of probability.
Pierre Simon de Laplace

In Chapter 4, we presented a compound prediction algorithm based on correlations between the 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 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 Chapter 4 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 ternaries crystal structure prototypes are unique but up to 50% are unique in the case of quaternary prototypes.

In this chapter, 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. We describe the mathematical model and its training on an experimental crystal structures database. The model predictive power is then evaluated by cross-validation. Finally, the chemical rules captured by the model are discussed and compared to more traditional approaches based on ionic size or position in the periodic table.

5.1 The data mined ionic substitutions model

5.1.1 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. [100]

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

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

Our method follows an approach used in the field of machine translation.[102] 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 evaluates 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.

5.1.2 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 a n component vector:

$$\mathbf{X} = (X_1, X_2, ..., X_n) \tag{5.1}$$

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

$$\Omega = \{Fe^{2+}, Fe^{3+}, Ni^{2+}, La^{3+}, ...\} \tag{5.2}$$

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

$$p_n(\mathbf{X}, \mathbf{X}') = p_n(X_1, X_2, ..., X_n, X'_1, X'_2, ..., X'_n) \tag{5.3}$$

Knowing such a probability function allows to assess how likely any ionic substitution is. For example, by computing $p_4(Ni^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$, one can evaluate how likely Fe^{2+} in a lithium transition metal phosphate is to be substituted by Ni^{2+}. In this specific example, this value is expected to be high as Ni^{2+} and Fe^{2+} are both transition metals with similar charge and size. Actually, LiNiPO$_4$ and LiFePO$_4$ both form in the same olivine-like structure. On the other hand, the substitution of Fe^{2+} by Sr^{2+} would be less likely and $p_4(Sr^{2+}, Li^{1+}, P^{5+}, O^{2-}|Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$ should have a low value. We must point out that the probability function does not have any crystal structure dependence. The fact that the compound targeted for substitution forms an olivine structure does not influence the result of the evaluated probability. This is an approximation in our approach.

The probability function $p_n(\mathbf{X}, \mathbf{X}')$ is a multivariate function defined in a high-dimensional space and cannot be estimated directly. For all practical purposes, this function needs to be approximated. We follow here an approach successfully used in other fields such as machine translation, and based on the use of binary indicators f, so-called *feature functions.*[103] These feature functions are mathematical representations of important aspects of the problem. The only mathematical requirement for a feature function is to be defined on the domain of the probability function $(\mathbf{X}, \mathbf{X}')$ and return 1 or 0 as result. They can be as complex as required by the problem. For an ionic substitution model, one could choose for example as a feature function:

$$f(\mathbf{X}, \mathbf{X}') = \begin{cases} 1 & \text{if } \mathrm{C}a^{2+} \text{ substitutes for } \mathrm{B}a^{2+} \text{ in the presence of } O^{2-} \\ 0 & \text{else} \end{cases} \tag{5.4}$$

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

$$p_n(\mathbf{X}, \mathbf{X}') \approx \frac{e^{\sum_i \lambda_i f_i^{(n)}(\mathbf{X}, \mathbf{X}')}}{Z} \tag{5.5}$$

The λ_i indicate the weight given to the feature $f_i^{(n)}(\mathbf{X}, \mathbf{X}')$ in the probabilistic model. Z is a partition functionensuring the normalization of the probability function. The exponential form chosen in equation (5.5) follows a commonly used convention in the machine learning community.[104]

5.1.3 The binary feature model

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, ...).

Therefore, we will omit any reference to n in the probability and feature functions. Equation (5.5) becomes

$$p(\mathbf{X}, \mathbf{X}') \approx \frac{e^{\sum_i \lambda_i f_i(\mathbf{X}, \mathbf{X}')}}{Z} \tag{5.6}$$

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

$$f_k^{a,b}(\mathbf{X},\mathbf{X}') = \begin{cases} 1 & X_k = a \; and \; X_k' = b \\ 0 & else \end{cases} \tag{5.7}$$

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

$$f_k^{Ca^{2+},Ba^{2+}}(\mathbf{X},\mathbf{X}') = \begin{cases} 1 & X_k = Ca^{2+} \; and \; X_k' = Ba^{2+} \\ 0 & else \end{cases} \tag{5.8}$$

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

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

$$\lambda_k^{a,b} = \lambda_k^{b,a} \tag{5.9}$$

and

$$\lambda_k^{a,b} = \lambda_l^{a,b} \tag{5.10}$$

5.1.4 The 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.[105, 106] For instance, $CaTiO_3$ and $BaTiO_3$ both form cubic perovskite structures with Ca and Ba on equivalent sites. This translates in our mathematical framework as a specific assignment for the variables vector $(\mathbf{X}, \mathbf{X}') = (Ca^{2+}, Ti^{4+}, O^{2-}, Ba^{2+}, Ti^{4+}, O^{2-})$. We will follow the convention in probability theory designing specific values of the random variable vector $(\mathbf{X}, \mathbf{X}')$ by lower case letters $(\mathbf{x}, \mathbf{x}')$. An entire crystal structure database D will lead to m assignments: $(\mathbf{X}, \mathbf{X}') = (\mathbf{x}, \mathbf{x}')^t$ with $t = 1, ..., m$

$$D = \{(\mathbf{X}, \mathbf{X}') = (\mathbf{x}, \mathbf{x}')^1, (\mathbf{X}, \mathbf{X}') = (\mathbf{x}, \mathbf{x}')^2, ..., (\mathbf{X}, \mathbf{X}') = (\mathbf{x}, \mathbf{x}')^{m-1}, (\mathbf{X}, \mathbf{X}') = (\mathbf{x}, \mathbf{x}')^m\} \tag{5.11}$$

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

Using these assignments obtained from the database, we follow the commonly used maximum-likelihood approach to find the adequate weights from the database available.[83] The weights maximizing the likelihood to observe the training data are considered as the best estimates to use in the model. For notations purpose we will represent the set of weights by a weight vector $\boldsymbol{\lambda}$.

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

$$\begin{aligned} l(D, \boldsymbol{\lambda}) &= \sum_{t=1}^{m} log\, p((\mathbf{x}, \mathbf{x}')^t|\boldsymbol{\lambda}) && (5.12) \\ &= \sum_{t=1}^{m}[\sum_{i} \lambda_i f_i((\mathbf{x}, \mathbf{x}')^t) - log\, Z(\boldsymbol{\lambda})] && (5.13) \end{aligned}$$

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

$$\boldsymbol{\lambda}_{ML} = \underset{\boldsymbol{\lambda}}{argmax}\, l(D, \boldsymbol{\lambda}) \tag{5.14}$$

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

5.1.5 The compound prediction process

When the substitution probabilistic model in equation (5.5) has been trained, it can be used to predict new compounds and their structures from a database of existing compounds. The procedure to predict a compound formed by species a,b,c and d is presented in figure 5-1. For each compound containing $(x_i^1, x_i^2, x_i^3, x_i^4)$ as ionic species, the probability to form a new compound by substitution of a, b, c and d for x_i^1, x_i^2, x_i^3 and x_i^4 is evaluated by computing $p(a, b, c, d|x_i^1, x_i^2, x_i^3, x_i^4)$. If this probability is higher than a given threshold σ, the substituted structure is considered. If this new compound candidate is charge balanced and previously unknown, it can be added to our list of new compounds candidates. If not, the algorithm goes to the next $i+1$ compound in the crystal structure database. The substitutions proposed by the model do not have to be isovalent. However, all suggested compounds have to be charge balanced.

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

5.2 Analysis of the model

A binary feature model based on the ternary and quaternary ionic compounds present in the Inorganic Crystal Structure Database (ICSD, [87]) 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

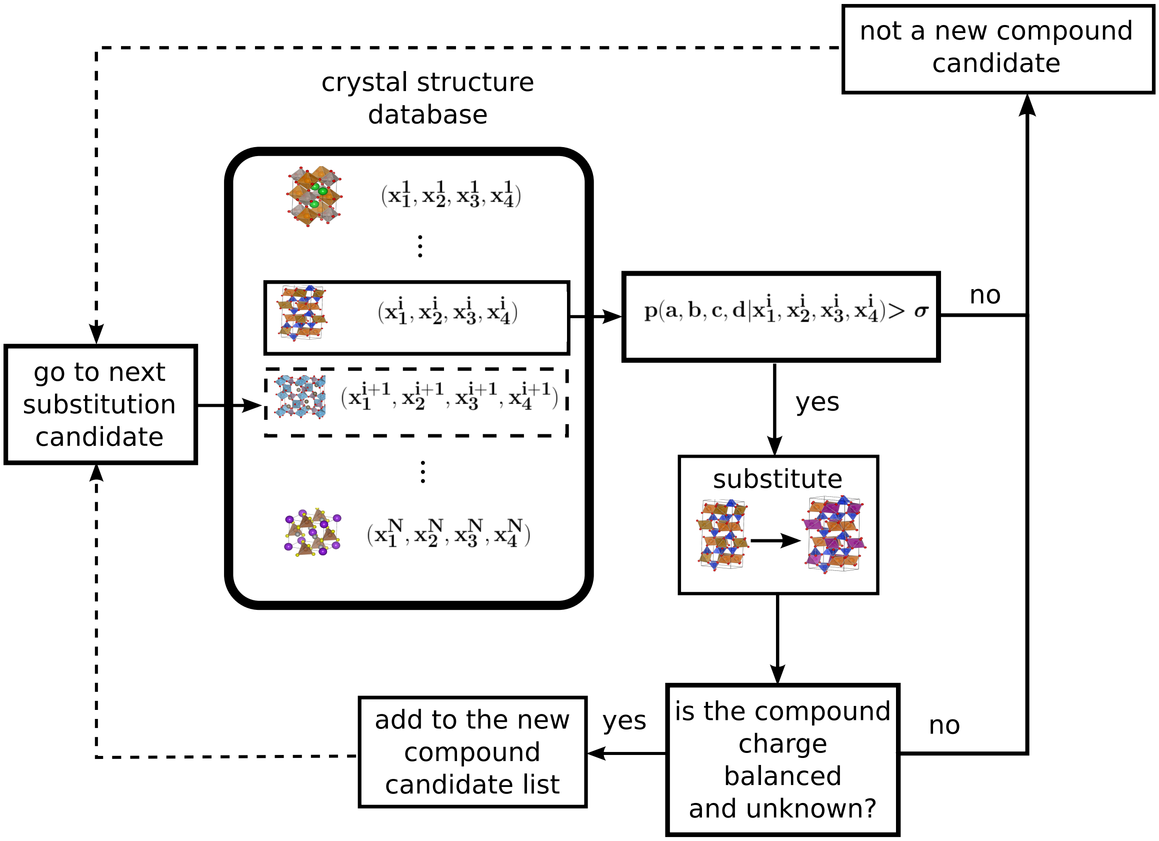

not a new compound candidate
crystal structure database
$(x_1^1, x_2^1, x_3^1, x_4^1)$
$(x_1^i, x_2^i, x_3^i, x_4^i)$
$(x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_4^{i+1})$
$(x_1^N, x_2^N, x_3^N, x_4^N)$
$p(a,b,c,d|x_1^i, x_2^i, x_3^i, x_4^i) > \sigma$
no
yes
go to next substitution candidate
substitute
add to the new compound candidate list
yes
is the compound charge balanced and unknown?
no

Figure 5-1: Procedure to predict new compounds formed by the a,b,c and d species using the substitutional probabilistic model.

partially occupied sites) are considered. Crystal structure similarity was found by an affine mapping technique as presented in Appendix A and used to obtain the database D of m assignments (equation (5.11)) necessary to train the model. A binary feature model was fitted on this data set using a maximum likelihood procedure as presented in 5.1.4.

5.2.1 Cross-validation on quaternary ICSD compounds

The procedure to discover new compounds using the probabilistic model was presented in 5.1.5. Using this procedure, we evaluated the predictive power of this approach by performing a cross-validation test.[88] Cross-validation consists in removing part of the data available (the test set) and training the model on the remaining data set (the training set). The model built in this way is then used to predict back the test set and evaluate its performance. We divided the quaternary ordered and ionic chemical systems from the ICSD in 3 equal-sized groups. We performed 3 cross-validation tests using all compounds in one of the group as test set and the remaining quaternary and ternary compounds as training set. This extensive cross-validation tested 2967 compounds in total. The cross-validation tests excluded compounds forming in prototypes unique to one compound, as our substitution strategy by definition cannot predict compounds in such unique prototypes. We also only considered substitution leading to charge balanced compounds.

Figure 5-2 indicates the false positive and true positive rates for a given threshold σ. The true positive rate (TP$_{rate}$) indicates the fraction of existing ICSD compound that are indeed found back by the model (i.e., true hits):

$$TP_{rate}(\sigma) = \frac{TP(\sigma)}{P} \tag{5.15}$$

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

$$FP_{rate}(\sigma) = \frac{FP(\sigma)}{N} \tag{5.16}$$

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

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

The clear separation between the two curves in figure 2 shows that the model is indeed predictive and can effectively distinguish between the substitutions leading to an existing compound and those leading to non-existing ones. Moreover, figure 5-2 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 on the figure by a dashed line.

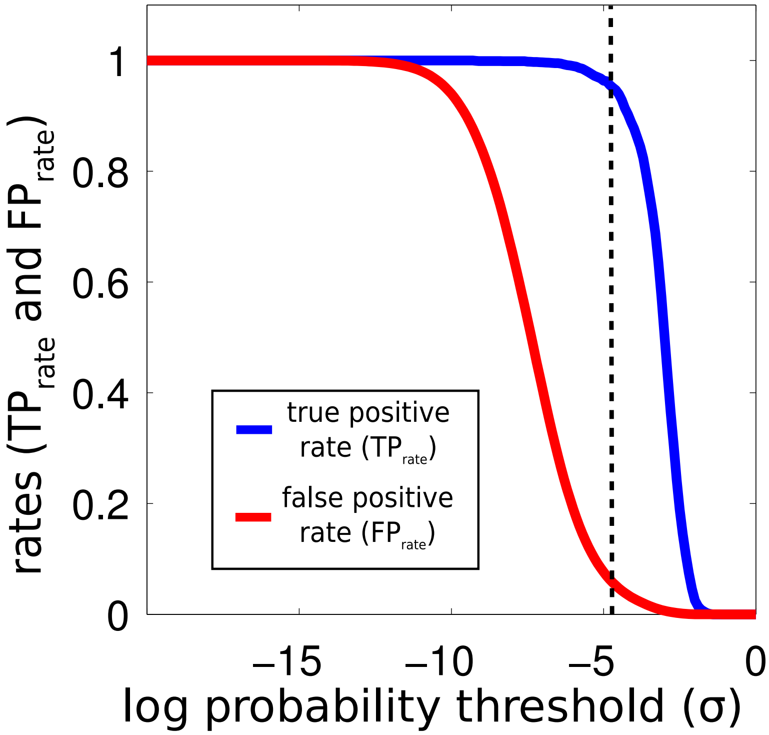


Figure 5-2: true positive rate (TP_{rate}, blue line) and false positive rate (FP_{rate}, red line) in function of the probability threshold (σ) logarithm during cross-validation.

These cross-validation results can also be used to compare our knowledge based method to a brute force approach in which all charge balanced substitutions from known compounds would be attempted. The brute force approach would require the testing of 884,037 compound candidates to recover the full set of known compounds during cross-validation. Using our model, recovering 95% of those known compounds would require testing only 53,251 candidates (i.e., only 6% of the number of brute force candidates).

5.2.2 Ionic pair substitution analysis

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

$$\begin{aligned} g_{ab} &= \frac{p(X_1 = a, X'_1 = b)}{p(X_1 = a)p(X_1 = b)} && (5.17) \\ &= \frac{p(X_1 = a, X'_1 = b)}{\sum\limits_j p(X_1 = a, X'_1 = x'_j) \sum\limits_j p(X_1 = b, X'_1 = x'_j)} && (5.18) \\ &= \frac{\frac{1}{Z} e^{\lambda_1^{a,b}}}{\frac{1}{Z} \sum\limits_j e^{\lambda_1^{a,x'_j}} \frac{1}{Z} \sum\limits_j e^{\lambda_1^{b,x'_j}}} && (5.19) \end{aligned}$$

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.

This pair correlation measures the increased probability to observe two ions at equivalent positions in a particular crystal structure over the probability to observe each of these ions in nature. Two ions which substitute well for each other will have a pair correlation higher than one (g_{ab}>1) while ions which rarely substitute will have a pair correlation lower than one (g_{ab}<1). The pair correlation is therefore a useful quantitative measure of the tendency for two ions to substitute for each other.

Figure 5-3 plots the logarithm (base 10) of this pair correlation for the 60 most common cations in the ICSD (the pair correlation for all the ionic pairs is presented in supplementary information). Positive values indicate a tendency to substitute while negative values on the contrary show a tendency not to substitute. The ions are sorted by their element Mendeleev number.[72] This ordering relates to their position in the periodic table. Therefore, the different ions are automatically clustered by chemical classes (alkali, alkali-earth, rare-earth, transition metals and main group elements).

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

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

While transition metals in general tend to substitute easily for each other, two subgroups of strong pair correlation can be observed: the early transition metals (Zr^{4+}, Ti^{4+}, Ta^{5+}, Nb^{5+}, V^{4+}, V^{5+}, W^{6+}, Mo^{6+}) and late transition metals (Cr^{3+}, Mn^{2+}, Mn^{3+}, Fe^{2+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Hg^{2+}, Cd^{2+}, Zn^{2+}). This separation in

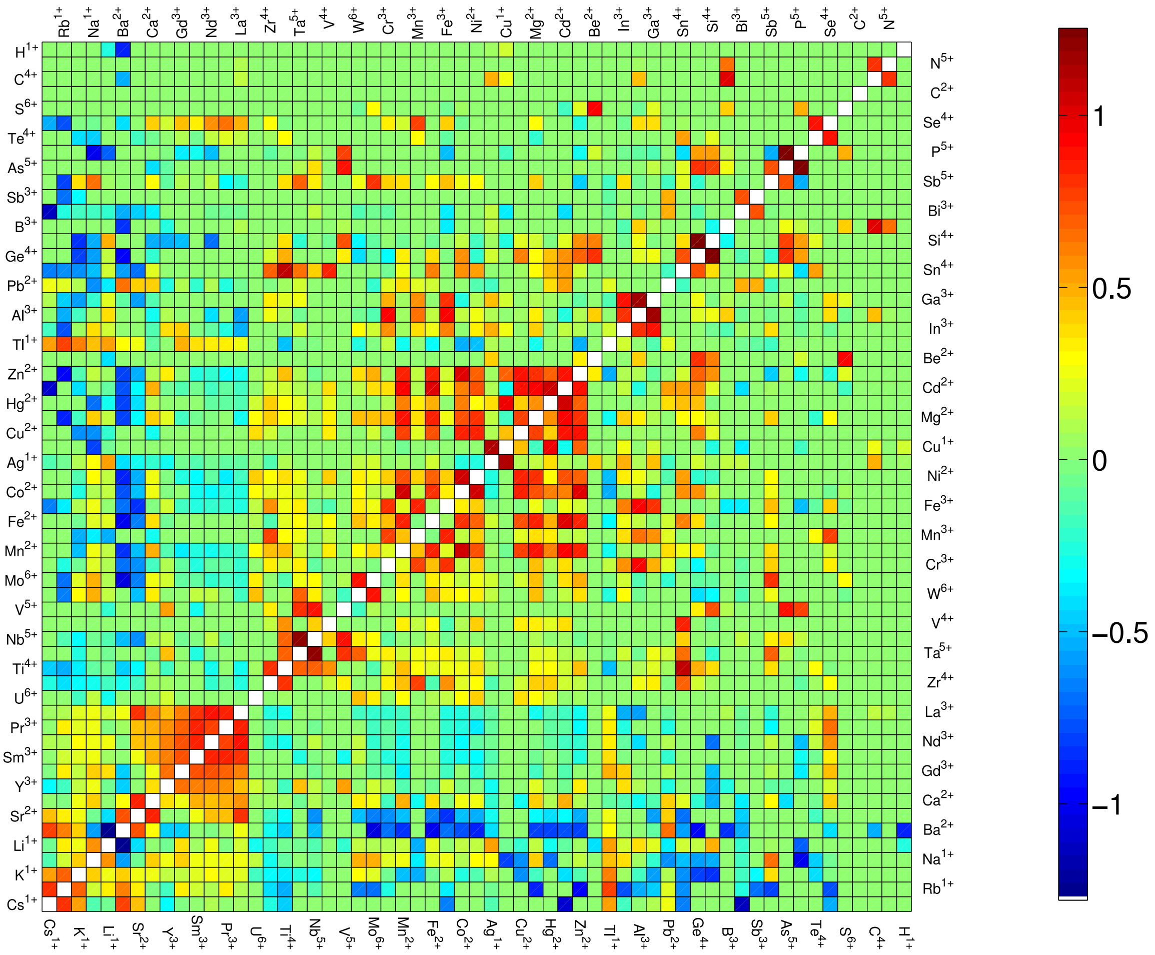


Figure 5-3: Logarithm (base 10) of the pair correlation g_{ab} for each ion couple a,b. Equation (5.17) was used to evaluate the pair correlation g_{ab}. The ions are sorted according to their element's Mendeleev number. Only the 60 most common ions in the ICSD are presented in this graph. These correlation coefficients were obtained by training our probabilistic model on the ICSD. Positive values indicate a tendency to substitute while negative values on the contrary show a tendency to not substitute. The symmetry of the pair correlation ($g_{ab} = g_{ba}$) is reflected in the symmetry of the matrix. .

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.[107]

On the other hand, the main group elements do not have an 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. 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 big +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+}, ...

Some ions tend to form very specific structures and local environment. 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.

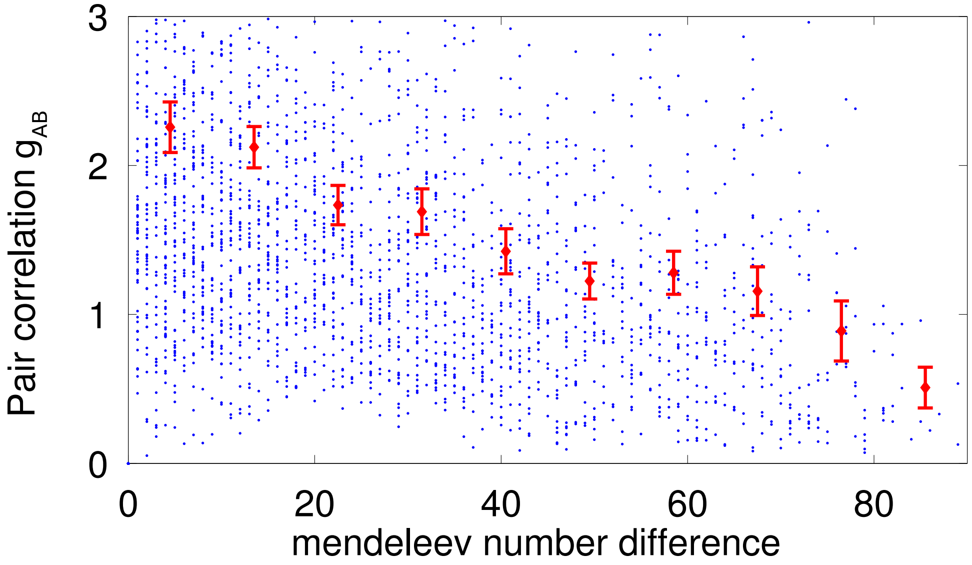


Figure 5-4: Pair correlation g_{ab} in function of the difference in Mendeleev number between the two ions a and b. Equation (5.17) was used to evaluate the pair correlation g_{ab}. The blue points are the raw data obtained from fitting the model on the ionic compounds in the ICSD. To distinguish the general trend from the scatter, the data has also been binned in 10 equally sized bins along the Mendeleev number difference axis. Each red point indicates the pair correlation mean for each bin with a 95% confidence interval as error bar. The pair correlation tends to decrease as the Mendeleev number difference increase.

5.2.3 chemical and size effects over the substitution tendencies

The previous analysis shows that strong or weak substitution tendencies can be often rationalized using chemical arguments (i.e. the relative position of the ionic pair in the periodic table). To study this effect, figure 5-4 plots the ionic pair correlation defined in equation (5.17) as function of the difference in Mendeleev number between the two ions. A relation is observed between this difference in Mendeleev number and the pair correlation. Higher pair correlation are associated with smaller differences in Mendeleev numbers. However this is only true on average and a large spread is observed around the mean values.

Some very interesting outliers can be pointed out. For instance, Cr^{6+} and S^{6+} while significantly distant from each other in the periodic table can easily substitute due to their common tendency to form tetrahedral polyanionic compounds (sulfates and chromates). Ti^{4+} and Sn^{4+} are also two ions with a high pair correlation coef-

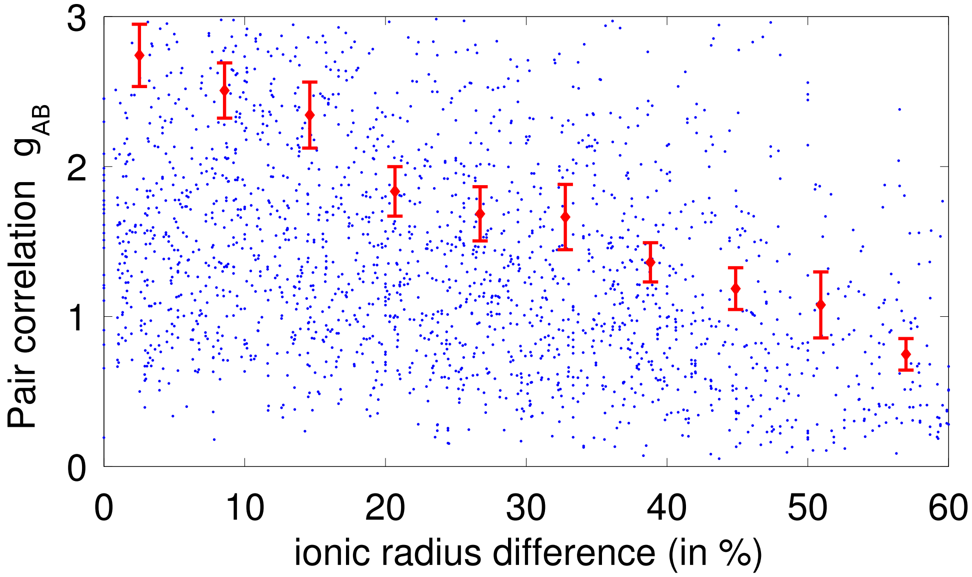


Figure 5-5: Pair correlation g_{ab} in function of the difference in ionic size between the two ions a and b. Equation (5.17) was used to evaluate the pair correlation g_{ab}. The ionic size difference is computed as the difference in ionic size divided by the size of the largest of the two ions. This gives relative ionic radius differences. The ionic size for the two ions are obtained from the Shannon radii table and for the coordination 6 have been used.[2] The blue points are the raw data obtained from fitting the model on the ionic compounds in the ICSD. To distinguish the general trend from the scatter, the data has also been binned in 10 equally sized bins along the Mendeleev number difference axis. Each red point indicates the pair correlation mean for each bin with a 95% confidence interval as error bar. The pair correlation tends to decrease as the difference in ionic size increase.

ficient despite an important difference in Mendeleev number. Conversely, $Rh^{3+}$and Co^{3+}, while in the same column of the periodic table, do not substitute strongly (g_{ab}=0.98).

In addition to chemical effects, size effects are also very often used to estimate how likely an ionic substitution is. Ions of similar size tend to be considered easier to substitute for each other. In Figure 5-5, the pair correlation g_{ab} is plotted in function of the difference in ionic size between the two ions. The ionic size used is the 6-fold coordinated size according to Shannon.[2] A clear relation between the two quantities can be observed. The highest pair correlations tend to be found for smaller differences in ionic size. As for the chemical effects, there is an important spread around the general trend. Again, S^{6+} and Cr^{6+} do not follow the general trend. Those two highly substitutable ions (g_{ab}=9.7) have a 50% difference in their ionic size.

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

5.2.4 Online ionic substitution model

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

5.3 Discussion

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

In addition, the substitution rules do not depend on structural factors. However, how easy a chemical substitution is will depend somewhat on the specific structure. Some crystal structure sites will accommodate for instance a wider range of ions with different size without major distortion. Perovksites are a good example of structures where the specific size tolerance factor is established (see for instance Zhang et al.[108]). In some sense, our model is "coarse grained" over structures.

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

Those simplifying assumptions are however very useful in the sense that they allow the model to capture rules from data dense regions and use them to make predictions in data sparse regions. The substitution rules learned from ternary chemical systems can be used to predict compounds in the much less populated quaternary space. Likewise, substitution rules learned from very common crystal structure prototypes can be learned and used to make predictions in uncommon crystal structures. It is this capacity for this simpler model to make predictions in sparser data regions which constitutes its main advantage versus more powerful models such as the one presented in chapter 4.

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.[104]

A limitation of the model lies in its inability to predict totally new crystal structures. Indeed, any new compound will be proposed by ionic substitutions from a compound with an already known crystal structure. This usual limitation to crystal structure prediction methods based on data mining (shared by the model presented in chapter 4) is however compensated by their much smaller computational requirements than a more exhaustive search based on optimization such as with a genetic algorithm.

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

Chapter 6

A High-throughput *ab initio* computational search for new lithium-ion battery cathode materials

Computers are useless. They can only give you answers.

Pablo Picasso

In the previous chapters, we showed how *ab initio* computations can be performed on a large scale enabling the high-throughput computational discovery of compounds and their crystal structures when combined with data mining techniques. In this chapter, we will present an application of this methodology for the discovery of new lithium-ion cathode battery materials.

The lithium-ion battery technology is nowadays ubiquitous in portable electronics. As new applications are considered (e.g., electrical and hybrid vehicles), new requirements emerge and the improvement of current lithium-ion technology is nowadays the focus of major research efforts. The discovery of new battery materials, outperforming the current commercial solutions in terms of energy density, cost, power density and safety, is one of the strategy to achieve major breakthrough in this field.[109, 110, 111, 112] In this chapter, the basic principles of lithium-ion batteries will be briefly exposed. The material properties of importance for a lithium-ion battery cathode will be briefly presented along with a discussion of their *ab initio* computations. Finally, we will show how battery materials can be screened computationally on a large scale, allowing experiments to focus on the most promising compounds and chemistries.

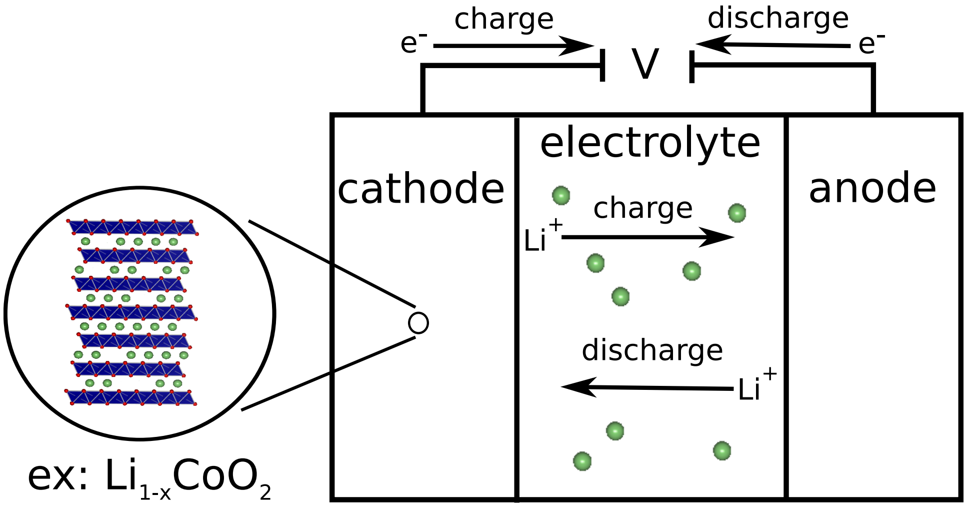


Figure 6-1: Schema of a lithium-ion battery. A zoom at the atomic scale on the crystal structure of a typical cathode material ($LiCoO_2$) is shown. Oxygen is in red. Cobalt sits in blue CoO_6 octahedra and lithium is in green.

6.1 Lithium-ion batteries

A battery is an electrochemical device used to store energy.[113] Two types of batteries exist: primary batteries, which can only be discharged once, and secondary (or rechargeable) batteries which can be charged and discharged multiple times. In this thesis, we will focus on rechargeable batteries. A battery consists in two electrode materials (the cathode and the anode) separated by an electrolyte. Figure 6-1 shows a schema for a lithium-ion battery.

In the case of lithium-ion battery, the electrolyte is a poor electron conductor but a good lithium ion conductor. The cathode is typically a transition metal oxide (e.g., $LiCoO_2$) and the anode is metallic lithium in laboratory settings or graphitic carbon in commercial cells. During charge, an electrical potential difference is applied between the cathode and anode. The lithium ions flow then through the electrolyte from the cathode to the anode. The $LiCoO_2$ cathode is depleted from lithium and the graphitic carbon is filled with lithium according to the two chemical reactions:

$$LiCoO_2 \rightarrow x\,Li + Li_{1-x}CoO_2 \tag{6.1}$$

$$x\,Li + C_y \rightarrow Li_xC_y \tag{6.2}$$

During discharge, the opposite reactions occurs. The difference between the high lithium chemical potential at the anode and the low chemical potential of lithium in the cathode makes the lithium ions flow from the anode to the cathode. The lithium intercalates back into the $LiCoO_2$ cathode and the graphitic carbon anode looses lithium. An electrical potential is then available between the cathode and anode. The following reactions occur:

$$Li_{1-x}CoO_2 + x\ Li \rightarrow LiCoO_2 \tag{6.3}$$

$$Li_xC_y \rightarrow C_y + x\ Li \tag{6.4}$$

Please not that during these charge and discharge processes, the transition metal present in the oxide cathode is oxidized/reduced between Co^{3+} and Co^{4+}.

In this work, we will only consider *insertion* batteries. Insertion materials are materials that intercalate or deintercalate lithium ions in a fixed structural framework. For instance, the layered crystal structure of $LiCoO_2$ is conserved while lithium ions are removed and replaced by vacancies. When the structural framework is not conserved, one speaks of *conversion* electrodes. Insertion batteries tend to be more reversible and exhibit less hysteresis between charge and discharge.[114] Only insertion systems will be considered in this work. We will also focus on the cathode material.

Below, some of the general key requirements for a lithium-ion insertion battery cathode material have been listed.[110] Depending on the application, the weight attributed to those requirements can vary.

- A high energy can be stored per unit of mass (Wh/kg) or volume (Wh/l). This requires a high *capacity* (i.e., the amount of charge that can be inserted between charge and discharge) per unit mass (mAh/g) or volume (mAh/cc) and a high *voltage*. The energy density is the product of the capacity and the voltage. Typical energy densities for a cathode materials are from 600 Wh/kg ($LiFePO_4$) to 800 Wh/kg (layered nickel cobalt manganese oxides).

- The material needs to be *cyclable* (i.e., present good reversibility). The material needs to be able to be charged and discharged many times with minimal capacity fade.

- The voltage should not be too high. While a high voltage is usually sought to maximize the energy density, current commercial electrolytes decompose around a voltage of 4.5V (with reference to lithium metal anode).

- Lithium needs to be able to diffuse fast enough in the bulk material. This is required for high rate charge and discharge (i.e., high power density).

- The material needs to be a good electronic conductor. Poor conductors can be used (e.g., phosphates) but large amount of inactive conductive additives such as carbon black are required then to insure good electrical contact in the cathode powder.

- The material in the discharged state needs to be stable enough versus oxygen evolution to minimize safety hazard.

- The material needs to be produced at a reasonable cost.

6.2 *Ab initio* computed properties

Some of the key properties mentioned in the previous section are controlled by intrinsic factors of the cathode materials at the atomistic scale. *Ab initio* computations have been used extensively to understand and predict those properties.[115]

Through a thermodynamical analysis, the average equilibrium voltage ($< V >$) of a cathode can be related to the difference in the Gibbs free energy between its charged state (delithiated phase) and discharged state (lithiated phase).[116] If $Li_{x_1}X$ and $Li_{x_2}X$ are respectively the lithiated and delithiated states, the average voltage between these two lithiation states is

$$< V >= \frac{-[G(Li_{x_2}X) - G(Li_{x_1}X) - (x_2 - x_1)G(Li)]}{(x_2 - x_1)F} \tag{6.5}$$

Where $G(Li)$ is the Gibbs free energy of the lithium metal anode and F the Faraday constant.

Because the contributions of entropy and volume effects to the Gibbs free energy difference are small, one may approximate the voltage using energy differences computed at zero temperature and pressure. As presented in chapter 1, such an energy computation can be performed by DFT. The average voltage is then

$$< V >\approx \frac{-[E(Li_{x_2}X) - E(Li_{x_1}X) - (x_2 - x_1)E(Li)]}{(x_2 - x_1)F} \tag{6.6}$$

As cathode materials contain very often transition metals, the self-interaction problem needs to be accounted for by a method such as GGA+U.[117] More recently, it has been shown that voltage can be accurately predicted also by the new hybrid Heyd-Scuseria-Ernzerhof (HSE) approach.[118, 3] Figure 6-2 demonstrates the good agreement between voltages computed using GGA+U or the HSE functionals and

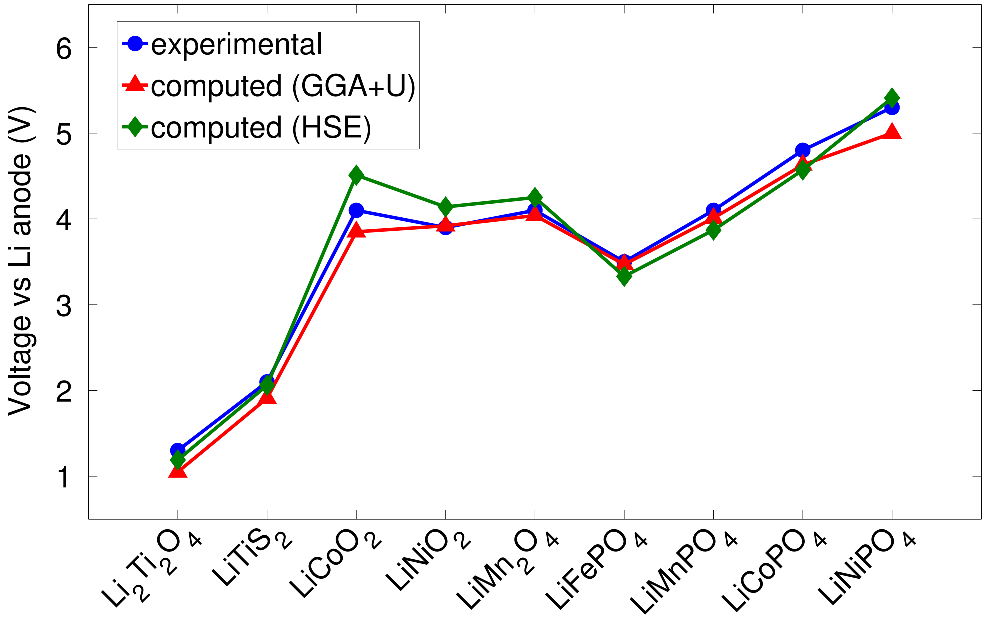


Figure 6-2: Comparison between experimental and computed (using GGA+U or HSE) voltages. As titanium does not require a U value to reproduce accurate energies, no U value was used for $Li_2Ti_2O_4$ and $LiTiS_2$. The data is from Chevrier et al.[3]

the experimental voltages for a variety of cathode materials. In average the computations deviate from experiment by only a few hundred mV. The HSE method has the advantage over GGA+U that it does not rely on a fitted parameter for each element as GGA+U. HSE computations are however computationally more expensive (in average 40 times more expensive than GGA+U).

Safety is one of the key considerations in the design of a lithium-ion battery cathode. Charged cathodes can be a safety hazard. A typical runaway reaction in a lithium battery starts by an overheating event (e.g., caused by an internal or external short), which causes the charged cathode to release oxygen. The released oxygen can potentially combust the flammable organic electrolyte and ultimately leads to fire. The safety of a cathode material can be evaluated by computing the critical oxygen chemical potential at which the charged cathode will release oxygen (see Chapter 2). This methodology has been successfully used to compare the safety of delithiated $LiMnPO_4$ and delithiated $LiFePO_4$.[119] Figure 6-3 plots different known cathode materials along a scale representing the critical oxygen chemical potential of their charged state. The ranking of the different materials follows the known experimental tendencies. $LiNiO_2$ and $LiCoO_2$ at the far left end are both unsafe materials while $LiFePO_4$ is one of the safest cathode material.

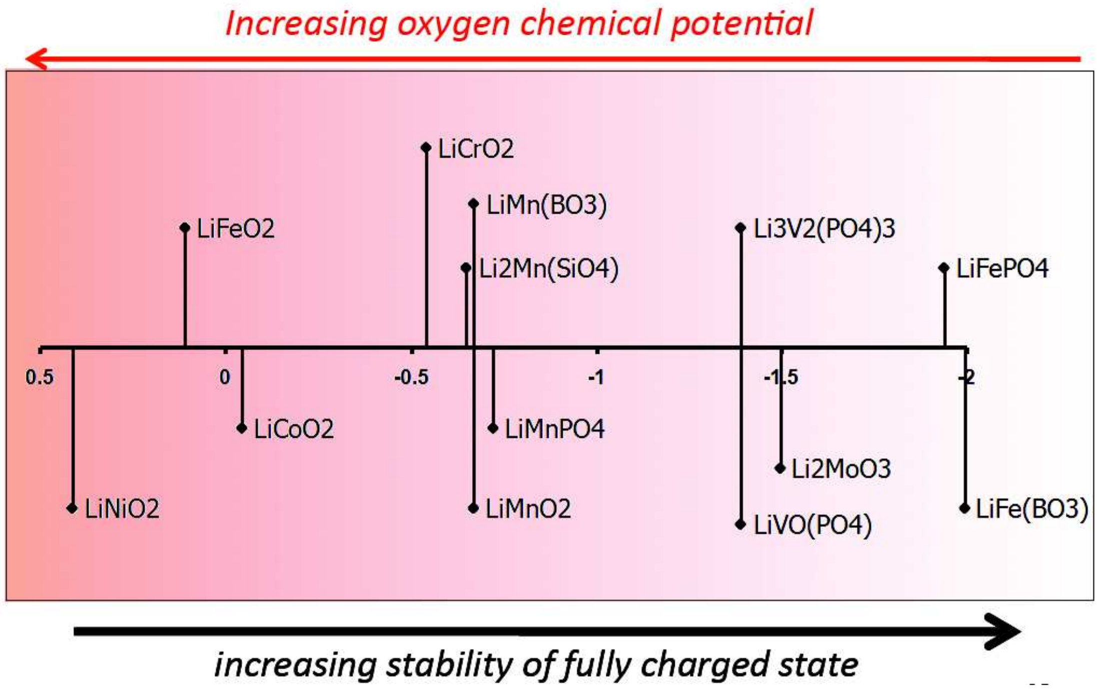


Figure 6-3: Critical oxygen chemical potential (in meV/at) for the charged state of known battery compounds.

For a battery material to charge and discharge at a high enough rate (i.e., to provide good power density), the lithium ions need to be able to diffuse fast enough in the material. Modeling diffusion in cathode materials involves the computations of energy barriers and the study of the vacancy and lithium ordering during discharge.[120, 121] A simpler but often sufficient approach is to focus only on the energy barrier. Small enough energy barriers to the lithium diffusion is a necessary requirement for fast diffusion. As a rule of thumb, the lowest lithium migration barrier needs to be lower than 500-600 meV for any battery material to be a good enough diffuser.[122] Migration barriers can be computed in the Nudge Elastic Band (NEB) framework with DFT computations. This approach has been used to predict that $LiFePO_4$ is an extremely fast one-dimensional diffuser with an activation barrier around 200 meV.[123] Materials design can be driven by DFT diffusion computations as shown by the crystal structure engineering performed by Kang et al. to improve the rate capability of layered $Li(Ni_{0.5}Mn_{0.5})O_2$.[124]

As well as fast lithium transport, fast electron transport is important for high rates in electrode materials. Most of cathode materials are most likely polaronic conductors. The limiting factor for the electronic conductivity is then the mobility of the polaron. The polaron migration barrier can be computed *ab initio* as shown for

$LiFePO_4$.[125]

Other important battery properties are more difficult to directly access through *ab initio* computations. Cyclability for instance tends to be difficult to predict directly as many, not always well understood, phenomena can be at the origin of the capacity fade during the cycling of the battery.[126] Some indications of possible poor cyclability can be obtained by DFT computations though. For instance, a large instability versus decomposition product of the metastable delithiated state(s) (i.e. a tendency to convert instead of intercalate) and/or large volume change during delithiation are often indications of a possible irreversibility.

6.3 Electrochemical testing

The main experimental tool used on a cathode battery material is the galvanostatic electrochemical testing.[127] A battery (also called cell) is built using a powder of the cathode material mixed with a binder and carbon. The carbon is added to provide good electrochemical contact between the grains of the cathode material. The cell is then typically constructed using lithium metal as anode and an organic solvent with a lithium salt as electrolyte.

A galvanostatic electrochemical test consists in imposing a current density to the cell while measuring the potential response (with respect to the lithium metal anode). A voltage window is set in which this test is performed and results from charge and discharge are observed. The results are usually plotted as a voltage versus capacity graph. Figure 6-4 shows a typical galvanostatic charge (in red) and discharge curve (in blue). This test provide important information on the material. One can directly determine how much capacity is accessible and at which voltage. Reversibility can also be assessed by observing if the capacity obtained during charge is recoverable during discharge. The hysteresis between charge and discharge is also of interest. Kinetic processes tend to make the voltage higher during charge and lower the voltage during discharge. This phenomena called *polarization* makes the charge and discharge curve not superposing. This test depends greatly on the rate at which the battery is charged and discharged (i.e., on the current density applied). A common convention in the battery community is to use the concept of C-rate. A rate of C/10 implies that the battery is fully charged in 10 hours. Similarly, a C/100 rate implies a current density that will fully charge the battery in 100 hours. Lower rates are closer to a equilibrium process and will show less polarization.

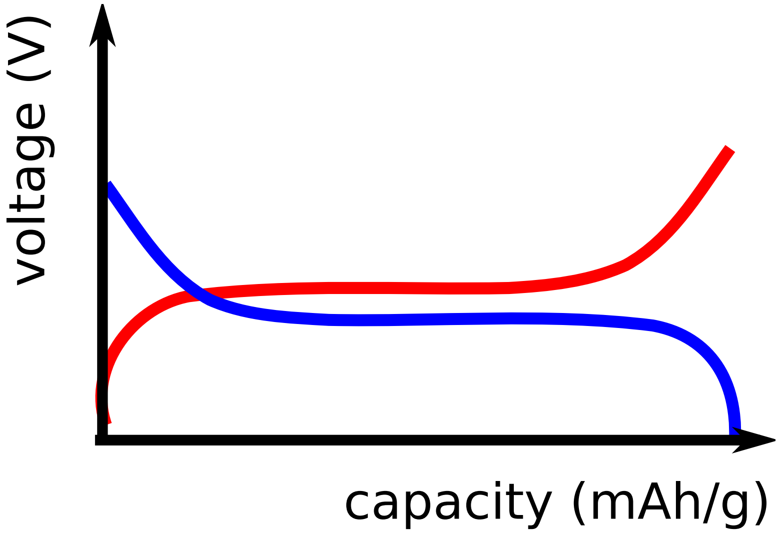


Figure 6-4: Typical capacity-voltage curve during galvanostatic electrochemical testing. The curve during charge is in red. The curve during discharge is in blue.

6.4 High-Throughput *ab initio* search for new cathode materials

Combining the predictive power of *ab initio* computations in the battery field with the possibility for performing high-throughput DFT computations, we have embarked in a large scale computational search of new cathode materials. A flow-chart for this high-throughput project is presented in Figure 6-5.

The starting point of our analysis is a database of known compounds (i.e., the ICSD), and a set of allowed elements (with their available oxidation states). Elements too expensive (e.g., Pt and Au) or too toxic (e.g., Cd) are excluded. A periodic table of technologically acceptable (in black) and unacceptable elements (in grey) is presented in Figure 6-6.

We consider stability, voltage and lithium diffusion to be quantities that are reliable enough through DFT to be considered in our high-throughput screening. The thermodynamical stability (in the lithiated and delithiated state) is computed using the convex hull construction presented in Chapter 2. In this work, we will especially concentrate on cathode materials stable in their lithiated (discharged) state. A technological reason motivates this choice. The anode used nowadays in commercial lithium-ion batteries is based on carbon and is synthesized without any lithium intercalated. The cathode is therefore expected to be the lithium source and contain lithium (i.e., being discharged) when synthesized. On the other hand, the charged

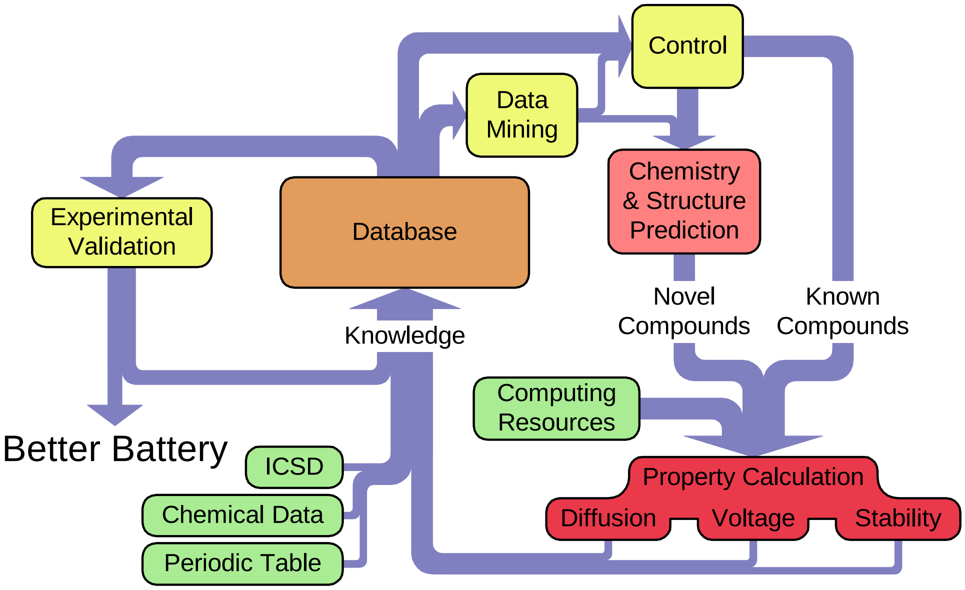


Figure 6-5: flow chart for the high-throughput computational cathode search project

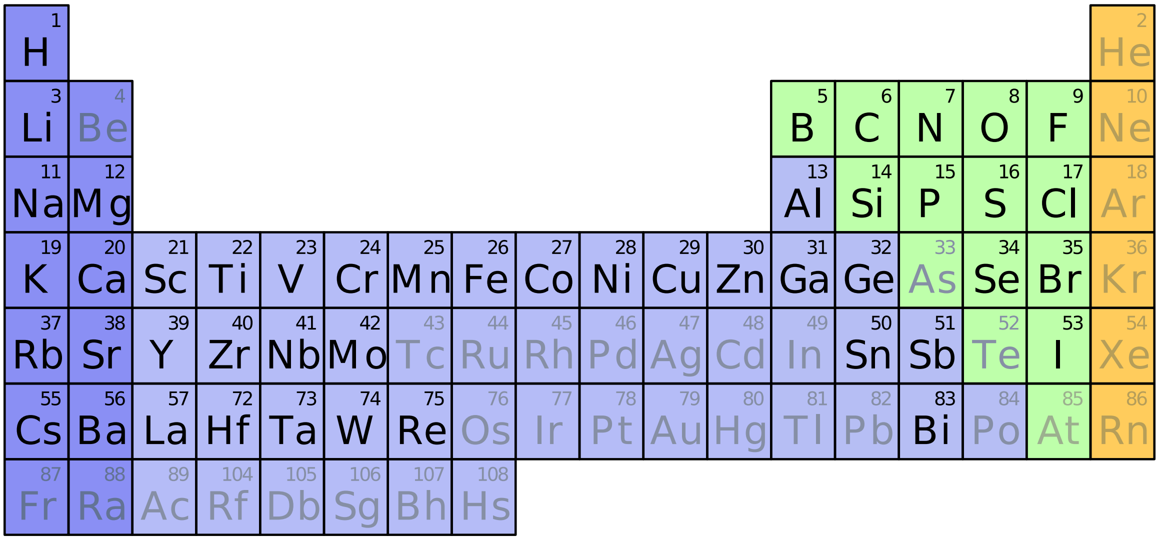


Figure 6-6: periodic table of the elements with technologically acceptable elements in black and non-acceptable elements in grey. Indications on toxicity and cost is also given. Please note that rare-earth and trans-uranides are not present and will not be considered in this thesis battery related work.

state is the object of a special focus on the evaluation of its critical oxygen chemical potential (i.e., its intrinsic safety). Voltages are computed by removing a certain amount of lithium (e.g., corresponding to a maximum capacity or to the formation of integer oxidation states). The choice of the vacancy-lithium ordering for delithiated compounds still containing lithium is made using a fixed lattice enumeration algorithm.[128] An Ewald summation is used to choose the ordering with the lowest electrostatics. From the capacity and voltage, the gravimetric and volumetric energy density can directly be computed.

Stability and voltages are properties requiring only simple DFT ionic relaxations. They can therefore be evaluated through the high-throughput *ab initio* methodology presented in Chapter 1. Lithium diffusion however is more challenging to perform on a large scale. Even the simple approach of computing lithium energy barriers through NEB computations turns out to be not trivial to perform without human intervention. This thesis will not go into the details of the issues with the scaling up of NEB computations. Lithium diffusion has not been used yet in a fully high-throughput way but is envisioned in the future. Energy barriers can however be computed on the limited set of compounds satisfying the previous already limiting conditions on voltage and stability.

We should stress that all battery properties cannot be directly computed *ab initio* on a large scale. Our investigations only focused on some necessary but not sufficient conditions that new cathode materials should meet. After the high-throughput discovery of a new material, low-throughput computational and experimental work is still required to evaluate its performances as a battery material.

Two types of compounds can be analyzed through our high-throughput approach: known and novel compounds. The known compounds can be directly extracted from a crystal structure database such as the ICSD. Searching for overlooked battery materials among already known compounds can be successful in the identification of new promising battery materials. For instance, a previously known monoclinic $LiMnBO_3$ has been identified by our high-throughput project.[129] However, such findings are rare in an extensively studied field such as lithium-ion batteries. Going beyond known compounds, using crystal structure prediction tools, as presented in chapter 4 and 5, is therefore paramount. We will now present in the following chapter a new family of promising cathode materials discovered through this high-throughput search.

Chapter 7

Carbonophosphates: a new family of cathode materials discovered by high-throughput *ab initio* computations

As outlined in Chapter 6, high-throughput *ab initio* computations can be used to search for potential new battery cathode materials. In this chapter, we present an example of a new family of promising battery compounds discovered *in silico* through this approach. The carbonophosphate compounds are described along with the computational results motivating their study as cathode materials. Experimental results on their synthesis and electrochemical testing are also provided.

7.1 The carbonophosphates and their *ab initio* predicted cathode properties

From our dataset of known ICSD battery materials and possible novel compounds (generated through the compound prediction algorithms presented in chapter 4 and 5), we searched for stable materials with gravimetric energy density higher than $LiFePO_4$ (i.e., >600 Wh/g), an operating voltage lower than 4.5V (to avoid electrolyte decomposition) and sufficient stability to make synthesis possible. A special attention was accorded to compounds unstable in their lithium form but stable in their sodium form. Indeed, it is common to synthesize the material in the sodium form and subsequently form a metastable lithium containing compound (while retaining

formula	Energy above the hull for A=Na (meV/at)	Energy above the hull for A=Li (meV/at)
$A_3Fe(CO_3)(PO_4)$	0	37
$A_3Mn(CO_3)(PO_4)$	0	65
$A_3Co(CO_3)(PO_4)$	0	54
$A_3Ni(CO_3)(PO_4)$	0	44

Table 7.1: Energy above the hull for the $A_3M(CO_3)(PO_4)$ (with A=Na, Li, and M=Co, Mn, Fe, Ni) compounds in the sidorenkite structure.

the sodium crystal structure) by a process called *lithium-sodium ion-exchange*. Such an ion-exchange has been performed for instance to form the unstable lithium layered $LiMnO_2$ by exchanging sodium by lithium in the stable layered $NaMnO_2$.[130] Other examples of materials synthesized by ion-exchange in the battery field are $Li_3V_2(PO_4)_3$ rhombohedral NASICON and $Li(Ni_{0.5}Mn_{0.5})O_2$.[124, 131]

A new family of cathode materials containing carbonates $(CO_3)^{2-}$ and phosphates $(PO_4)^{3-}$ groups was discovered through this systematic search. This *carbonophosphate* family is composed of compounds of formula $A_3M(CO_3)(PO_4)$ (with A=Na, Li, and M=Co, Mn, Fe, Ni). While the sodium version of the manganese and iron compounds are known rare minerals, respectively named sidorenkite and bonshtedite,[132, 133, 134] they have never been artificially synthesized, and their lithium versions have not been known to exist. The sidorenkite $Na_3Fe(PO_4)(CO_3)$ and bonshtedite $Na_3Mn(PO_4)(CO_3)$ minerals are isostructural and crystallize in a monoclinic P21/m (11) space group. Figure 1 shows their crystal structure in a 2x2x2 supercell. The unit cell is drawn in black. Each transition metal octahedron (in purple) shares 4 vertices with tetrahedral PO_4 groups (in grey) and an edge with a CO_3 group (in brown). The two-dimensional subunits extending along the (100) plane is composed of connected transition metal octahedra, PO_4 groups and CO_3 groups. Alkali metals (in yellow) connect two of those two-dimensional subunits. The alkalis occupy two different sites: coordinated by 7 and 6 oxygen atoms respectively.

The computed stability data is presented in Table 7.1. The Na version for M=Co, Mn, Fe and Ni are computed to be stable at 0K with respect to decomposition to any compound present in the ICSD. However, the Li versions are all predicted to be moderately unstable.

The instability predicted for the lithium versions indicates the need to synthesize those compounds by first forming the Na version and performing a Li-Na ion exchange. Table 7.2 shows the lattice parameters (lattice vectors lengths and angles) for the computed and ICSD compounds. A very good agreement is found between

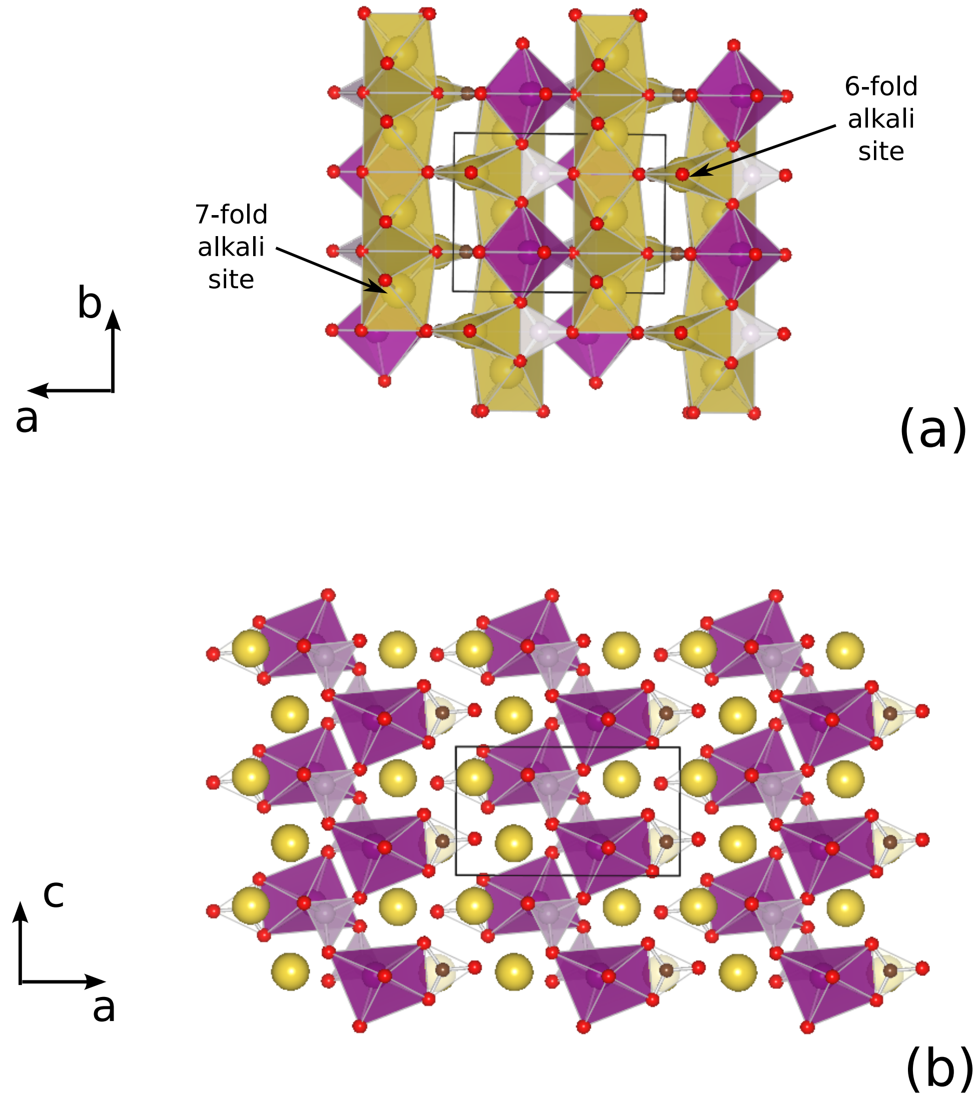


Figure 7-1: The structure $Na_3Mn(PO_4)(CO_3)$ of the (M=Mn, Fe, Co, Ni) compounds, viewing down from c axis (a) and b axis (b). Alkalis are in yellow and oxygen in red. PO_4 groups are grey and CO_3 groups are brown. Transition metal MO_6 octahedra are in purple.

the experimentally reported sodium iron carbonophosphates and GGA+U computations. Lattice lengths are computationally overestimated by less than 1%. In addition, computations indicate that during Li-Na exchange important relaxations will happen. Around 11% volume change, major contraction along the a axis (5%), and deviations from the 90 degrees lattice angles of the Na version are computationally predicted. These relaxations break the few symmetries present in the crystal structure and makes the lithium exchanged structure triclinic. The difference in size between Li and Na explain this relaxation.

	A=Na, ICSD	A=Na, GGA+U	A=Li, GGA+U
a	8.956 A	9.014 A	8.546 A
b	6.629 A	6.644 A	6.421 A
c	5.149 A	5.189 A	5.103 A
α	90	90	93.4
β	89.5	89.7	96.4
γ	90	90	93.4
volume	305.700 A^3	310.741 A^3	277.146 A^3
space group	P21/m	P21/m	P1

Table 7.2: Lattice vectors and angles for iron sidorenkites $A_3Fe(CO_3)(PO_4)$ computed for A=Na and A=Li, and reported in the ICSD for A=Na.

Voltage computations (vs lithium metal) for the $Li_3M(CO_3)(PO_4)$ (M=Mn, Fe, Co, Ni) compounds predict only the Fe (Fe^{2+}/Fe^{3+}: 3V), Mn (Mn^{2+}/Mn^{3+}: 3.3V; Mn^{3+}/Mn^{4+}: 4.1V) and Co (Co^{2+}/Co^{3+}: 4.1V) based compounds to be electrochemically active in the voltage range suitable to the current electrolyte technology (<4.5V) (see Figure 7-2). The carbonophosphate manganese phase is the most promising in terms of energy density due to its two available redox couples. When 2 electrons are available, the full theoretical capacity of these carbonophosphates is around 230 mAh/g. Theoretical energy density of 859 Wh/kg and 2376 Wh/l are computed for the $Li_3Mn(CO_3)(PO_4)$ compound. This would be a 45% and 15% improvement in terms of gravimetric and volumetric energy density compared to the most successful polyanion based cathode battery currently used: $LiFePO_4$.[135] Volume changes during delithiation are predicted to be low for both compounds 1.1% for Fe (after one lithium per iron removed) and 2.4% for Mn (after 2 lithium per manganese was removed) indicating a potentially good cyclability.

Lithium ion migration barriers can be evaluated *ab initio* using the nudge elastic

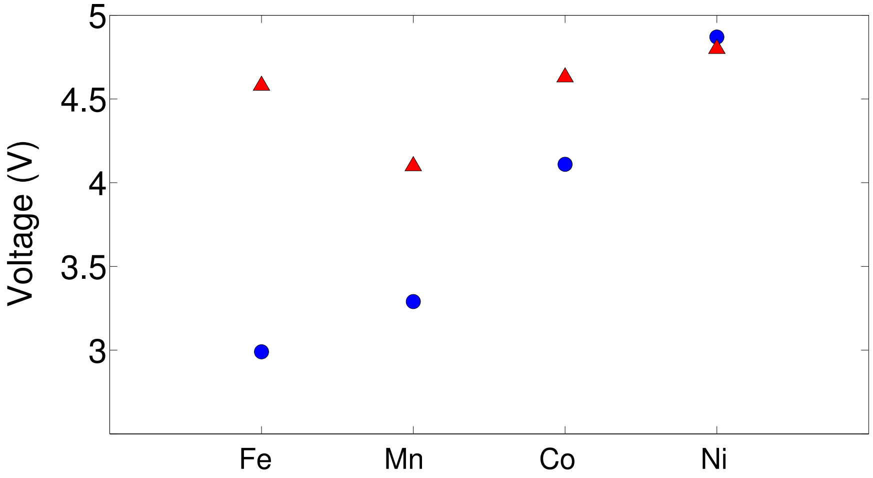


Figure 7-2: Voltages for the +2/+3 (blue circles) and +3/+4 (red triangles) redox couples for the delithiation of $Li_3M(CO_3)(PO_4)$ in the sidorenkite crystal structure.

band (NEB) framework. To the contrary of the rest of the DFT results, these energy barriers have been evaluated using GGA without any U parameter to avoid charge ordering complications when GGA+U is used with NEB. Activation barriers for the fully lithiated ($Li_3Mn(CO_3)(PO_4)$) and partially delithiated ($Li_2Mn(CO_3)(PO_4)$) manganese compound have been computed at 410 meV and 400 meV. Similar values were found for the iron compound (410 and 390 meV). Those values indicate that lithium diffusivity should not limit the activity of the manganese and iron sidorenkites.

7.2 Experimental synthesis and electrochemical testing

Motivated by the previous computational results, we decided to synthesize both $Li_3Fe(CO_3)(PO_4)$ and $Li_3Mn(CO_3)(PO_4)$ through Li-Na ion-exchange from the stable sodium phases. The two sodium based compounds while reported as minerals had never been artificially synthesized. These experimental results have been obtained by Hailong Chen.

$Na_3Fe(CO_3)(PO_4)$ and $Na_3Mn(CO_3)(PO_4)$ were synthesized hydrothermally. In a representative synthesis of $Na_3Mn(CO_3)(PO_4)$, 0.002 mole $MnNO_3{\cdot}4H_2O$ was dissolved in 5 ml water to form a clear solution A. 0.002 mole of $(NH_4)_2HPO_4$ and 2 g of Na_2CO_3 were dissolved in 10 ml of water to form a clear solution B. Solution A was then quickly added to solution B under fast magnetic agitation. The mixture slurry was then transferred to a Teflon-lined autoclave and heated at 120°C for 20 hours in an Ar flushed glove box. After the autoclave was slowly cooled down to room temperature, the products were washed with distilled water for several times, followed by drying in a vacuum oven at 40°C overnight. For the synthesis of $Na_3Fe(CO_3)(PO_4)$, the procedure is the same except that $FeSO_4{\cdot}7H_2O$ was used as the transition metal source and the mixing of the solutions were done in an Ar flushed glove box to prevent possible oxidation of Fe^{2+} species to Fe^{3+}.

Phase-pure $Na_3Fe(CO_3)(PO_4)$ and $Na_3Mn(CO_3)(PO_4)$ were successfully obtained following this hydrothermal route. Figure 7-3.a shows the XRD pattern and Rietvelt refinements of the synthetic $Na_3Fe(CO_3)(PO_4)$ using the structure model proposed by Khomyakov et al.[132] Lattice parameters extracted from the refinement are: a = 8.9406(9) Å, b = 6.5844(0) Å, c=5.1687(5) Å, $\alpha = \gamma = 90^o$, and $\beta = 89.45(7)^o$. The synthetic material was tested by inductively coupled plasma (ICP) analysis and showed ratios of elements fairly close to the formula stoichiometry: Na:Fe:P= 2.93 : 1.11 : 0.96. The synthetic $Na_3Fe(CO_3)(PO_4)$ are light green precipitation in solution after hydrothermal reaction and the color changes to very light brown after washed with distilled water. The change of color may indicate slight oxidation of Fe^{2+} on the surface due to the short exposure to air during washing. To avoid significant oxidation of Fe^{2+}, the hydrothermal reaction and the following washing process were done in a Ar flushed glove box and the powder were dried in a vacuum oven at 40°C overnight. Figure 7-3.b shows the XRD and refinement results for the synthetic $Na_3Mn(CO_3)(PO_4)$ sample. The cell parameters extracted from the refinement are: a=8.977(5), b= 6.734(6),c= 5.158(9) $\alpha = \gamma = 90^o$, and β=90.170. The synthetic $Na_3Mn(CO_3)(PO_4)$ was also tested by ICP and the ratio of elements are: Na:Mn:P= 3.01: 1: 0.98, The color of the powder is white. The Mn compound appears stable when exposed to air and the synthesis and washing can be done without using glove box. In this study, we still kept all manganese processes in the Ar glove box, for the sake of comparison with the Fe compound.

Li-Na ion exchange for both samples were performed with a conventional procedure using 2M LiBr in 1-hexanol solution. The ion-exchange for the Fe sample was done at 110°C for 3 days, and 90°C for 28 days for the Mn sample.

Scanning electron microscopy (SEM) pictures for both iron and manganese samples before and after ion-exchange show very uniform particle sizes around 300 nm with plate-like morphology. The ion-exchange does not alter the particle size distribution.

Figure 7-4 shows the powder XRD pattern after ion-exchange for both the iron and manganese sample. The pattern keeps the first major peak corresponding to the (100)

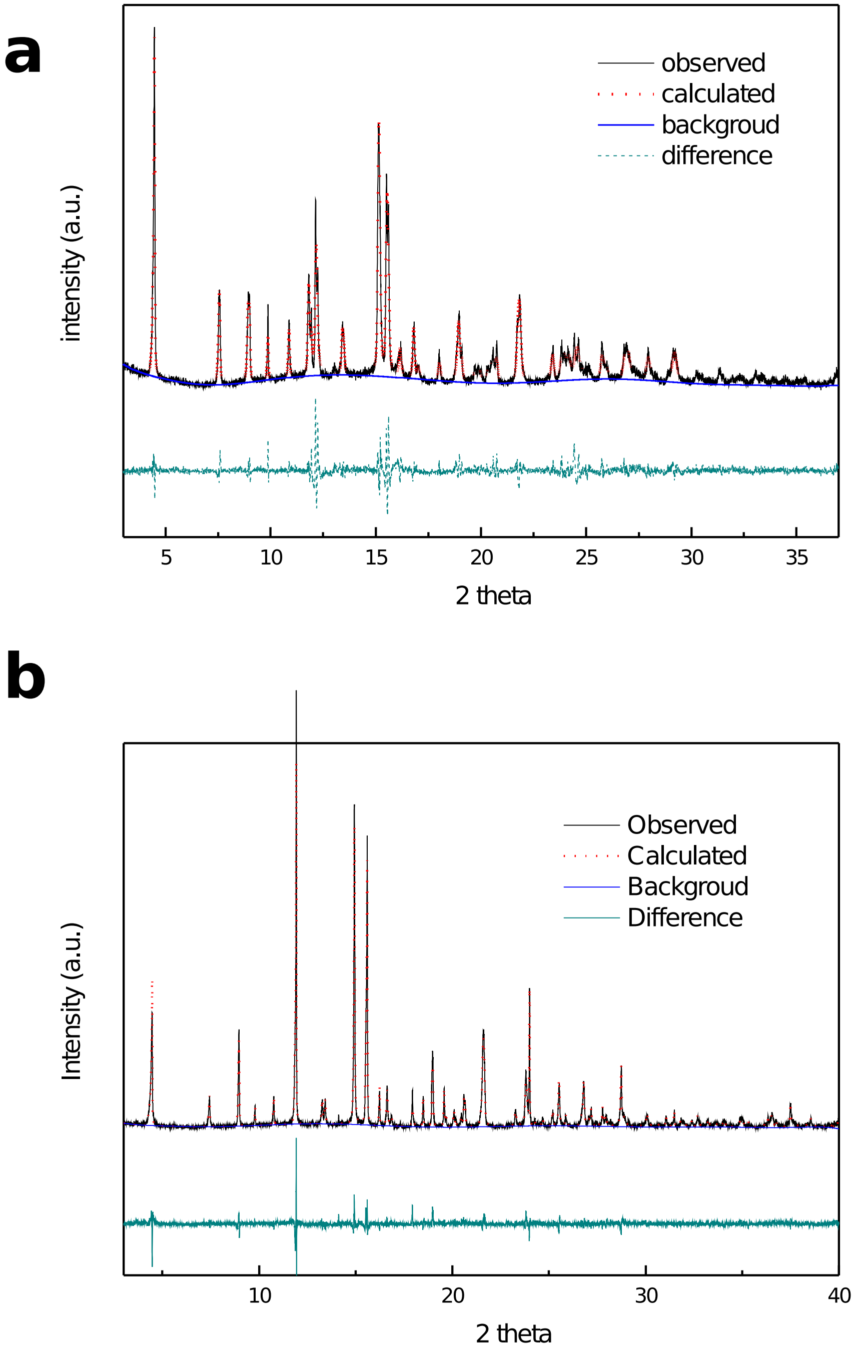


Figure 7-3: Powder XRD pattern for $Na_3Fe(CO_3)(PO_4)$ (a) and $Na_3Mn(CO_3)(PO_4)$ (b). Radiation used is from a synchrotron (wavelength=0.7A).

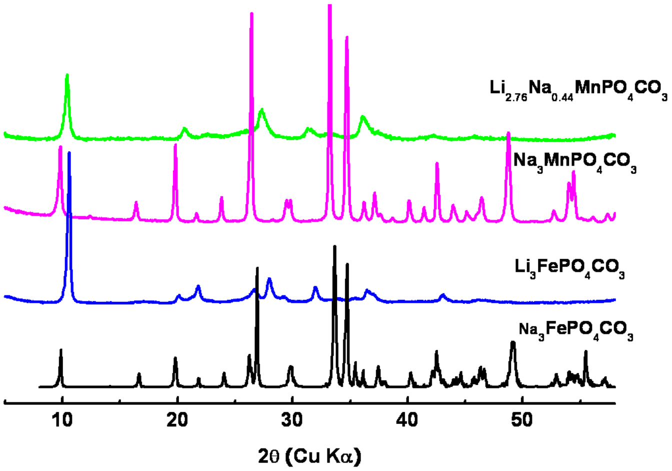


Figure 7-4: Powder XRD pattern for the iron and manganese samples after ion-exchange. Cu Kα radiation was used (λ=1.5108 A).

reflexion and characteristic of the sidorenkite structure (at 5° for λ=0.7A and 10° for λ=1.5108A) but the rest of the pattern changes significantly. The low symmetry of the structure and the relaxation to an even less symmetric triclinic cell after ion-exchange can explain this dramatic change in XRD pattern. The XRD line shape of the ion-exchanged samples becomes much broader than the Na precursors. Since the particle size did not change significantly, this peak broadening could be due to the high density defects, such as stacking faults, caused by the ion-exchange process. Direct evaluation of the lattice parameters after ion-exchange using the XRD pattern could not be performed due to the poor quality of the signal. Transmission electron microscopy (TEM) of the $Li_3Fe(CO_3)(PO_4)$ sample was used to obtain the lattice parameters: a=8.487 A, b=6.4222 A, c=4.9445 A, α=90.46°, β=95.19°, γ=96.93° and to confirm the relaxation to a triclinic space group of the $Li_3Fe(CO_3)(PO_4)$ compound.

Atomic content characterization using ICP was also performed on both samples after ion-exchange. The results showed that ion-exchange are complete for the Fe sample, with the element ratios being Li:Na:Fe:P=2.95:0.08:1.01:0.95. However, the ion-exchange for the Mn samples is more difficult. With a much longer exchange time, only a part of Na is replaced by Li, with the element ratios being Li:Na:Mn:P= 2.76:0.44 :0.86: 0.94.

The ion-exchanged $Li_3Fe(CO_3)(PO_4)$ and $Li_3Mn(CO_3)(PO_4)$ samples have been tested electrochemically. Results for a galvanostatic test of the $Li_3MnPO_4CO_3$ compound at C/100 are presented in Figure 7-5. Two cusps close to the corresponding

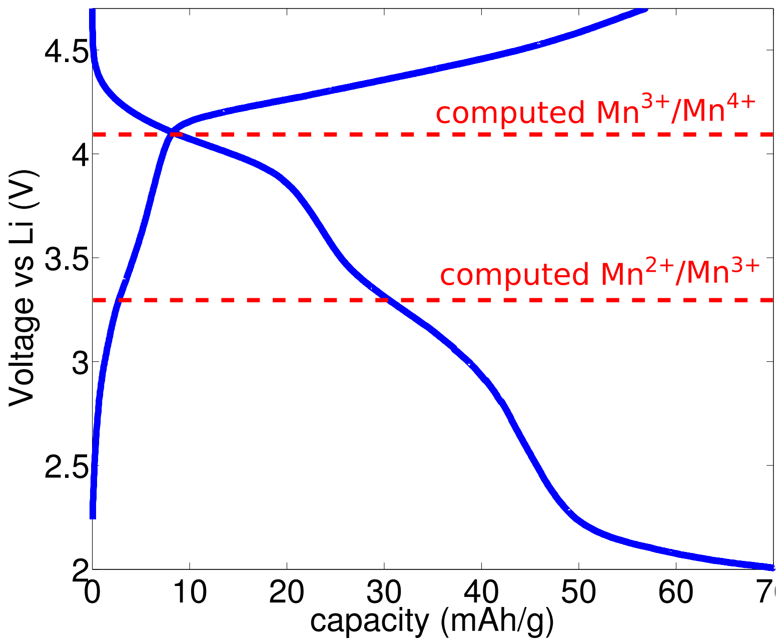


Figure 7-5: Galvanostatic profile for the manganese lithium compound (the formula from ICP is $Li_{2.76}Na_{0.44}Mn_{0.86}(CO_3)(PO_4)$). The electrochemical test was performed at a slow C/100 rate between 2-4.8 V at room temperature. The showed charge-discharged profile is after 12 cycles. The computed values from GGA+U for the two active couples (Mn^{2+}/Mn^{3+}: 3.3V and Mn^{3+}/Mn^{4+}: 4.1V) are drawn by a dashed red line.

Mn^{2+}/Mn^{3+} and Mn^{3+}/Mn^{4+} computed voltages are present in the curve. The capacity is however much smaller than the theoretical capacity (231 mAh/g) even at the very low rate used.

For the sake of comparison, we performed galvanostatic testings of the $Li_3Fe(CO_3)(PO_4)$ compound. Figure 7-6 shows those results for cells cycled at a C/10 rate, at room temperature (a) and at elevated temperature (b). A total capacity of 100 mAh/g (close to theoretical 115 mAh/g) is observed for the high-temperature sample. On the other hand, the room temperature sample does not show full capacity. This indicates a kinetic limitation to the lithium insertion/extraction. The voltage is very close to the computed GGA+U voltage (Fe^{2+}/Fe^{3+}: 3V) and a good reversibility is observed.

The increase in capacity observed at high temperature compared to room temperature testing for $Li_3Fe(CO_3)(PO_4)$ indicates that kinetic factors are limiting the material's performances. Bulk lithium diffusion should be fast enough to not be the likely limiting factor, according to the DFT computations. The study of other possible limiting factors such as the electronic conductivity (i.e., the polaron mobility,

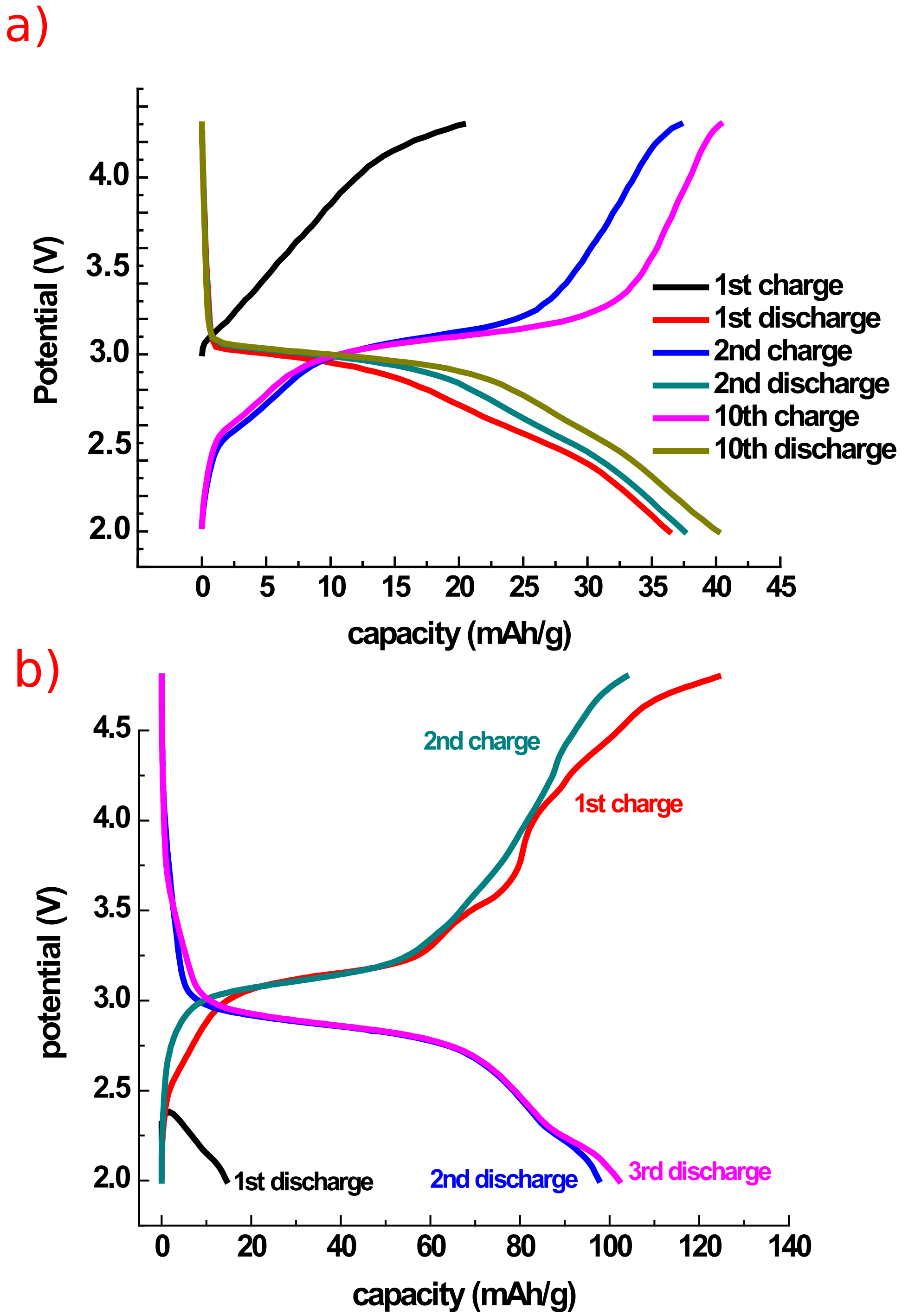


Figure 7-6: The voltage profile of $Li_3Fe(CO_3)(PO_4)$ cycled at various rates between 2-4.3 V at room temperature with a C/10 rate at room temperature (a) and 60°C (b).

[125]) and/or Li surface diffusion [136] will be needed in the future.

The manganese compound shows a more ideal voltages (3.2 and 4.1V) than the iron compound, and is able to carry a two-electron reaction in a commercial electrolyte. Two electrons phosphates compounds being able to work in a 3 to 4.5V voltage window are rare, especially when a 2+ to 4+ couple is sought. Recent attempts to use the two potentially available electron in $Li_2FeP_2O_7$ and $Li_2MnP_2O_7$ did not succeed due to the high voltage of the Mn^{3+}/Mn^{4+} and Fe^{3+}/Fe^{4+} couples.[137, 138] The poor electrochemical performances of the manganese compound compared to the iron version could be due to many factors. The most obvious difference between the two compounds is the presence of residual sodium. According to ICP, the manganese sample shows 26% residual Na on the alkali sites after ion-exchange. Partially exchanged iron samples (with 15% Na on alkali sites) showed poor electrochemical performances (room temperature capacity around 15 mAh/g). Improving the ion-exchange rate of the Mn compound and obtaining a fully ion-exchanged $Li_3Mn(CO_3)(PO_4)$ sample might improve the achieved capacity.

7.3 Synthesis of $Na_3Co(CO_3)(PO_4)$, $Na_3Ni(CO_3)(PO_4)$

The nickel and cobalt compounds in the sidorenkite structure are less interesting than the manganese version as lithium-ion battery due to their high voltage (see Figure 7-2). They have never been observed in nature (even as minerals) but have been predicted by DFT to be stable (see Table 7.1). To test this prediction, we performed a similar hydrothermal synthesis of those nickel and cobalt compounds. Figure 7-7 shows the XRD pattern for the four $Na_3M(CO_3)(PO_4)$ (M=Fe, Mn, Co, Ni). The similarity of the XRD patterns show the isostructurality of the different compounds and confirm the DFT prediction.

7.4 Chemical exploration of the sidorenkite structure

The presence of a tetrahedral and triangular polyanionic group in a compound is unusual. High-throughput DFT computations can be used to perform a computational chemical exploration, searching for other chemistry stabilizing the sidorenkite structure. All the possible $A_zM(YO_3)(XO_4)$ compounds formed in the sidorenkite structure can be generated and tested for stability in DFT. A is an alkali (A=Na,

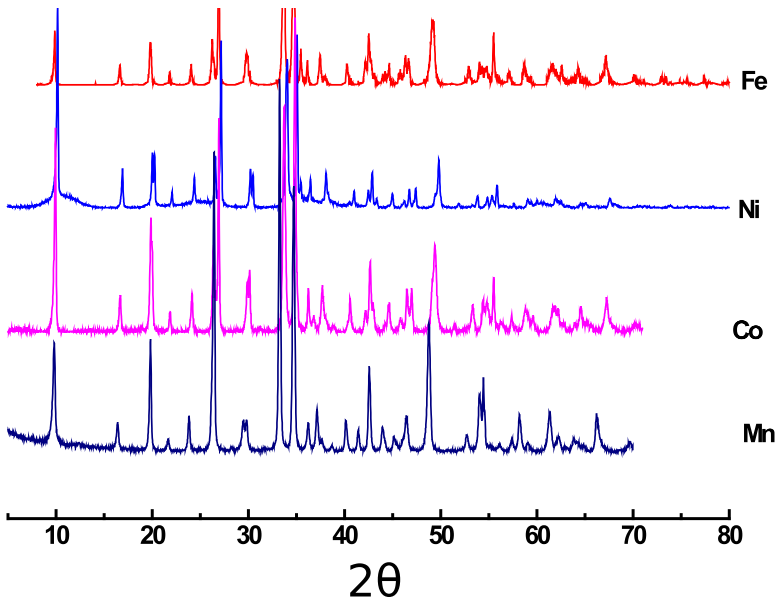


Figure 7-7: XRD pattern for the four $Na_3M(CO_3)(PO_4)$ (M=Fe, Mn, Co, Ni). Cu Kα radiation was used (λ=1.5108 A).

Li), Y a tetrahedron forming element (Y=P, As, Si, S), Y a triangular planar forming element (Y=C,B) and M a redox active metal (M=Mn, Fe, Ni,...). The amount z of alkali can be used to insure charge balance of the compound. It would be experimentally very challenging to attempt synthesis of all the possible compounds formed by these substitutions but high-throughout DFT computations can perform this task and help the experimentalist focus on the most stable combination(s).

Chapter 8

Conclusions and future work

As the computational power available to researchers continues to grow and *ab initio* methods become more accurate and predictive, an opportunity exists to accelerate the materials discovery process by computationally identifying the most technologically promising chemistries, prior to any time-consuming and expensive synthesis.

One of the challenges in this new paradigm is the prediction of new compounds and their crystal structures. This thesis presented two compound prediction methods based on data mining: one based on correlations between crystal structure prototypes and another based on likely ionic substitutions. The power of these data mining approaches when combined with DFT computations was demonstrated through a large-scale search for new ternary oxides. However, the compound and crystal structure prediction problem is not yet fully solved and many challenges lie ahead. For instance, methods that are able to predict compounds that crystallize in previously unknown crystal structure prototypes are still not available when only a limited computational budget is available. A data mining approach (in the form presented in this thesis or in any other form) combined with an optimization algorithm, such as a genetic algorithm, might be one route to successfully address this problem in the future.

In terms of applications, the high-throughput *ab initio* search for new lithium-ion cathode materials is not yet complete. Experimental efforts on a handful of new compounds are currently being pursued. Among those, lithium manganese carbonophosphate was presented as an example of a successful *in silico* compound discovery. However, the full theoretical potential of this novel material has not yet been demonstrated and future computational and experimental efforts are necessary to improve its current performance.

Beyond the simple screening of a computational materials database to search for compounds with certain properties, the large amount of data provided through high-throughput computations can also be used to better understand the factors governing important battery properties. While many of the cathode materials design rules are most often based on a few experimental results, we believe an analysis on a larger and more homogeneous data set will improve the fundamental understanding of the factors governing key battery properties (e.g., voltage or safety). Such a high-throughput analysis is currently ongoing in the field of phosphate cathode materials.

While the lithium-ion battery field is especially suited to high-throughput *ab initio* computations, the methods presented in this thesis are also useful in tackling many of the materials challenges faced by our society. A publicly available database of all computationally accessible materials properties for known and predicted compounds, known as the "materials genome" project, is currently being constructed. The materials design process will greatly benefit from such a database in various areas, including, for example, photovoltaics and transparent conducting oxides.

Appendix A

The crystal structure prototyping

Both data mining algorithms presented in chapter 4 and 5 need access to a database for which all the compounds have been classified to their corresponding crystal structure prototype. The Inorganic Crystal Structure Database (ICSD) we used in this thesis is only partially prototyped. In this appendix, we will present a brief literature review on prototyping algorithms and results we obtained on the prototyping of the ICSD database.

A.1 The crystallographic approach

A.1.1 Crystallographic definition of crystal structure similitude

The International Union for Crystallography (IUCr) defined in a report the different degree of similitude between inorganic structures.[139]

• Two structure are *isopointal*: if they have the same space-group and have the same Wyckoff position sequence after standardization. Two remarks can be made here. Nothing is said about the actual atomic positions and the cell parameters. Two isopointal structure can have very different geometric arrangements and atomic coordinates. A case in point is FeS_2 pyrite and CO_2. These are isopointal but have very different atomic environment. In addition, the Wyckoff sequence is dependent on the representation. Indeed, origin shifts or rotation of the unit cell can modify this sequence. A standardization procedure is then needed to check if two structures are isopointal.

• Two structures are *isocongurationnal* (or *isotypic*): if they are isopointal and if their geometric parameters (axial ratios, angles between the cell axes, and values of the atomic positions) are similar. We can note that similar is a very vague notion that could vary from problem to problem and from author to author.

A.1.2 The standardization approach

The main standardization algorithm is the one developed by Gelato and Parthe.[105] As there is more than one way to represent a crystal structure, the standardization procedure tries every single allowed shifts and rotations in the space group of the structure. A number (called a standardization coefficient) is associated with each of these representations following a certain number of rules. The representation with the lowest of these standardization coefficients is the standardized representation. When all the entries are represented in their standardized representation, they can be more easily directly compared. In the majority of cases this method works perfectly fine. However, it is not a robust method. Indeed a slight change of one of the parameters can sometimes change drastically the representation chosen. This is a known source of false negative where two structures that are similar are missed by the standardization. A case in point are the structures of $CeCu_2$ and HgK_2 which are very similar but do not have the same wyckoff sequence in their standard representation.

A.1.3 The affine mapping approach

Due to the lack of robustness of the standardization procedure an other kind of approach based affine mappings has been developed. An affine mapping is a geometrical transformation consisting in a linear transformation followed by a translation.

The first paper introducing this approach was from Burzlaff et al.[140] In this paper, two structures are isotypic if they have the same space group and an an affine mapping can transform one into another with a small misfit. This misfit is evaluated by some function evaluating the difference in angles, length of the unit cell axes, and in position of the atoms.

Not only is this approach robust, but it can be also extended to structure close to each other but not belonging to the same space groups. Indeed small changes in atomic positions can break a symmetry element and make two structures very similar geometrically to not be considered in the same prototype. A group of crystal structures that can be transformed into each other by an affine mapping but do not have the same space group are called an *aristotype*.

The paper from Burzlaff developed the theory behind the classification by affine mapping but do not provide an algorithm to actually compute these affine mappings. It is only recently that Hundt et al. proposed an algorithm to compute these mappings.[106] We must note, however, that this algorithm does not give you the best possible mapping. It only tells you if there is one in the given tolerances.

Let us assume we are trying to fit a structure A to a structure B. The algorithm consists in four steps:

1. Reduce the two structure to the same density by a uniform scaling of their unit cell axes.

2. Search for all the not too distortive linear transformations mapping any quadruplets of a given species (the fitting specie) in structure A to the unit cell in structure B.

3. Each of these allowed transformations is applied to the structure A atoms (all of them) and followed by a translation that brings the fitting species to any of the corresponding atom in structure B. If one of these transformations followed by a translation brings in correspondence all the atoms from A to an atom from B in a given tolerance (atomic mist tolerance), we consider that the structures are similar.

4. Finally if a mapping is found, its inverse must be checked to actually fit B to A.

Measuring how distortive a transformation is has been done in the affine mapping literature by evaluating how much the angles and the cell length change. The problem with this approach is that it will depend on the crystal structure representation chosen. The changes in angles and axial length will be different in a conventional and a primitive cell for example. We adopted to face that problem another approach based on the analysis of the transformation matrix. This matrix can be decomposed in a product of the closest unitary matrix (in the Fröbenius norm sense) and a symmetric positive definite matrix. The decomposition involved is called the polar decomposition and is used in the image processing field.[141] The distortion cannot come from the unitary matrix part and must be included only in the second matrix that we will call therefore the stretch matrix. We want now to define a measure of the distortion for this stretch matrix. This measure must obviously be invariant to the references axes. Moreover, it should give a zero value for a uniform scaling (all 3 eigenvalues equal) and treat equally the three eigenvalues (no direction are favored). One of the expression satisfying this is the second invariant of the deviatoric tensor:

$$(a_1 - a_2)^2 + (a_2 - a_3)^2 + (a_1 - a_3)^2$$

Where a_1, a_2, a_3 are the eigenvalues of the stretch matrix. If the value of this invariant is lower than a given tolerance we accept the transformation. From our tests, we have established that a tolerance of 0.1 is reasonable for prototype finding.

The affine mapping algorithm can be used online to compare crystal structures provided as cif files: `http://www.materialsgenome.org/structurematcher/`

A.1.4 Prototyping of the ICSD

We used the affine mapping algorithm to prototype the entire ICSD. The ICSD has been prototyped but only partially and with a non-robust standardization algorithm which requires some human oversight.[142] We decided to prototype the whole ICSD database following the affine mapping algorithm proposed in the previous section.

The ICSD contains around 100,000 entries. Actually, around 50,000 if we consider only the structure without partial occupancies. A first search for duplicate entries (i.e., entries with the same composition and the same crystal structure prototype), let only 26,732 entries left. From these 26,732 entries, we found around 11,443 crystal structure prototypes.

Bibliography

[1] O. Muller and R. Roy. *The major ternary structural families*. Springer-Verlag, 1974.

[2] R. D. Shannon. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallographica Section A*, 32(5):751–767, September 1976.

[3] V. L. Chevrier, S. P. Ong, R. Armiento, M. K. Y. Chan, and G. Ceder. Hybrid density functional calculations of redox potentials and formation energies of transition metal compounds. *Physical Review B*, 82(7):075122, August 2010.

[4] L. Wang, T. Maxisch, and G. Ceder. Oxidation energies of transition metal oxides within the GGA+U framework. *Physical Review B*, 73(19):1–6, 2006.

[5] O. Kubaschewski, C. B. Alcock, and P. J. Spencer. *Thermochemical Data*, chapter 5, pages 257–323. Pergamon Press, sixth edition, 1993.

[6] E. A. Carter. Challenges in modeling materials properties without experimental input. *Science*, 321(5890):800–3, August 2008.

[7] G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, and S. Curtarolo. Data-Mining-Driven Quantum Mechanics for the Prediction of Structure. *MRS Bulletin*, 31(December):981–985, 2006.

[8] J. Hafner, C. Wolverton, and G. Ceder. Toward Computational Materials Design: The Impact of Density Functional Theory on Materials Research. *MRS Bulletin*, 31(September):659–668, 2006.

[9] N. W. Ashcroft and N. D. Mermin. *Solid State Physics*. Brooks Cole, 1976.

[10] R. M. Dreizler and E. Gross. *Density Functional Theory: An Approach to the Quantum Many-Body Problem*. Springer-Verlag, 1991.

[11] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.

[12] K. Burke. The ABC of DFT. 2007.

[13] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, 1996.

[14] W. Foulkes, L. Mitas, R. Needs, and G. Rajagopal. Quantum Monte Carlo simulations of solids. *Reviews of Modern Physics*, 73(1):33–83, January 2001.

[15] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):864–871, 1964.

[16] W. Kohn and L. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):1131–1138, 1965.

[17] J. Perdew, K. Burke, and M. Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review letters*, 77(18):3865–3868, October 1996.

[18] H. Monkhorst and J. Pack. Special points for Brillouin-zone integrations. *Physical Review B*, 13(12):5188–5192, 1976.

[19] G. Kresse and J. Furthmuller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane. *Computational Materials Science*, 6:15–50, 1996.

[20] P. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994.

[21] M. Payne, M. Teter, D. Allan, T. Arias, and J. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Reviews of Modern Physics*, 64(4):1045–1097, 1992.

[22] V. Anisimov, J. Zaanen, and O. Andersen. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Physical Review B*, 44(3):943–954, 1991.

[23] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+U method. *Journal of Physics: Condensed Matter*, 9:767–808, 1997.

[24] S. L. Dudarev, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Physical Review B*, 57(3):1505–1509, January 1998.

[25] F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan, and G. Ceder. First-principles prediction of redox potentials in transition-metal compounds with LDA+U. *Physical Review B*, 70:235121, December 2004.

[26] M. Cococcioni and S. de Gironcoli. Linear response approach to the calculation of the effective interaction parameters in the LDA+U method. *Physical Review B*, 71(3):1–16, January 2005.

[27] S. Curtarolo, D. Morgan, and G. Ceder. Accuracy of methods in predicting the crystal structures of metals: A review of 80 binary alloys. *Computer Coupling of Phase Diagrams and Thermochemistry*, 29(3):163–211, September 2005.

[28] W. Setyawan and S. Curtarolo. High-throughput electronic band structure calculations: Challenges and tools. *Computational Materials Science*, 49(2):299–312, August 2010.

[29] J. Greeley, T. F. Jaramillo, J. Bonde, J. K. Nø rskov, and I. B. Chorkendorff. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature materials*, 5(11):909–13, 2006.

[30] A. Jain, S.-A. Seyed-Reihani, C. C. Fischer, D. J. Couling, G. Ceder, and W. H. Green. Ab initio screening of metal sorbents for elemental mercury capture in syngas streams. *Chemical Engineering Science*, 65(10):3025–3033, May 2010.

[31] C. Ortiz, O. Eriksson, and M. Klintenberg. Data mining and accelerated electronic structure theory as a tool in the search for new functional materials. *Computational Materials Science*, 44(4):1042–1049, February 2009.

[32] T. R. Munter, D. D. Landis, F. Abild-Pedersen, G. Jones, S. Wang, and T. Bligaard. Virtual materials design using databases of calculated materials properties. *Computational Science & Discovery*, 2(1):015006, November 2009.

[33] A. Jain, G. Hautier, C. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder. A High-Throughput Infrastructure for Density Functional Theory Calculations. *submitted to Journal of Computational Materials Science.*

[34] S. Curtarolo. *Coarse-Graining and Data Mining Approaches to the Prediction of Structures and their Dynamics*. PhD thesis, Massachusetts Institute of Technology, 2003.

[35] D. Wood and A. Zunger. A new method for diagonalising large matrices. *Journal of Physics A*, 18(0305):1343–1359, 1985.

[36] E. R. Davidson. Matrix Eigenvector Methods. In G. H. F. Diercksen and S. Wilson, editors, *NATO Advanced Study Institute on Methods in Computational Molecular Physics*, page 95, 1983.

[37] H. B. Callen. *Thermodynamics and an Introduction to Thermostatistics*. Wiley, 1985.

[38] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, USA, 1987.

[39] G. Ceder, A. Ven, C. Marianetti, and D. First-principles alloy theory in oxides. *Modelling and Simulation*, 8:311–321, 2000.

[40] A. Van De Walle and G. Ceder. First-principles computation of the vibrational entropy of ordered and disordered Pd_3V. *Physical Review B*, 61(9):5972–5978, March 2000.

[41] F. Zhou, T. Maxisch, and G. Ceder. Configurational Electronic Entropy and the Phase Diagram of Mixed-Valence Oxides: The Case of Li_xFePO_4. *Physical Review Letters*, 97:155704, October 2006.

[42] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, December 1996.

[43] S. P. Ong, L. Wang, B. Kang, and G. Ceder. Li-Fe-P-O_2 Phase Diagram from First Principles Calculations. *Chemistry of Materials*, 20(5):1798–1807, March 2008.

[44] M. W. Chase. *NIST-JANAF Thermochemical Tables.* American Institute of Physics, Woodbury, NY, 1998.

[45] A. Jain, G. Hautier, S. P. Ong, C. Moore, C. C. Fischer, and G. Ceder. Accurate Formation Enthalpies by Mixing GGA and GGA+U Calculations. *in preparation.*

[46] F. J. DiSalvo. Challenges and opportunities in solid-state chemistry. *Pure and Applied Chemistry*, 72(10):1799–1807, 2000.

[47] A. R. Akbarzadeh, C. Wolverton, and V. Ozolins. First-principles determination of crystal structures, phase stability, and reaction thermodynamics in the Li-Mg-Al-H hydrogen storage system. *Physical Review B*, 79(18):1–10, May 2009.

[48] M. O'Keeffe. Aspects of crystal structure prediction: some successes and some difficulties. *Physical chemistry chemical physics : PCCP*, pages 10–15, June 2010.

[49] S. M. Woodley and R. Catlow. Crystal structure prediction from first principles. *Nature materials*, 7(12):937–46, December 2008.

[50] J. C. Schon, K. Doll, and M. Jansen. Predicting solid compounds via global exploration of the energy landscape of solids on the ab initio level without recourse to experimental information. *Physica Status Solidi (B)*, 247(1):23–39, January 2010.

[51] F. Ducastelle. *Order and Phase Stability in Alloys, Volume 3 (Cohesion and Structure)*. North Holland, 1991.

[52] J. M. Sanchez, F. Ducastelle, and D. Gratias. Generalized cluster description of multicomponent systems. *Physica A: Statistical and Theoretical Physics*, 128:334–350, 1984.

[53] G. Ceder. A derivation of the Ising model for the computation of phase diagrams. *Computational Materials Science*, 1(2):144–150, 1993.

[54] G. L. W. Hart. Verifying predictions of the $L1_3$ crystal structure in Cd-Pt and Pd-Pt by exhaustive enumeration. *Physical Review B*, 80(1):1–5, July 2009.

[55] V. Blum and A. Zunger. Structural complexity in binary bcc ground states: The case of bcc Mo-Ta. *Physical Review B*, 69(2):20103, January 2004.

[56] A. R. Oganov and M. Valle. How to quantify energy landscapes of solids. *The Journal of Chemical Physics*, 130(10):104504, March 2009.

[57] J. C. Schön and M. Jansen. First Step Towards Planning of Syntheses in Solid-State Chemistry: Determination of Promising Structure Candidates by Global Optimization. *Angewandte Chemie International Edition in English*, 35(12):1286–1304, July 1996.

[58] D. J. Wales and H. A. Scheraga. Global optimization of clusters, crystals, and biomolecules. *Science*, 285(5432):1368–72, August 1999.

[59] D. J. Wales and J. P. K. Doye. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *The Journal of Physical Chemistry A*, 101(28):5111–5116, July 1997.

[60] Z. Li and H. A. Scheraga. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84(19):6611–5, October 1987.

[61] D. Deaven and K. Ho. Molecular geometry optimization with a genetic algorithm. *Physical Review Letters*, 75:288–291, 1995.

[62] S. M. Woodley, P. D. Battle, J. D. Gale, and C. Richard A. Catlow. The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation. *Physical Chemistry Chemical Physics*, 1(10):2535–2542, 1999.

[63] N. L. Abraham and M. I. J. Probert. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Physical Review B*, 73(22):1–6, June 2006.

[64] A. R. Oganov and C. W. Glass. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *The Journal of Chemical Physics*, 124(24):244704, June 2006.

[65] A. R. Oganov and C. W. Glass. Evolutionary crystal structure prediction as a tool in materials design. *Journal of Physics: Condensed Matter*, 20(6):064210, February 2008.

[66] G. Trimarchi and A. Zunger. Global space-group optimization problem: Finding the stablest crystal structure without constraints. *Physical Review B*, 75(10):1–8, March 2007.

[67] G. Trimarchi, A. Freeman, and A. Zunger. Predicting stable stoichiometries of compounds via evolutionary global space-group optimization. *Physical Review B*, 80(9):1–4, September 2009.

[68] A. Kolmogorov, S. Shah, E. Margine, A. Bialon, T. Hammerschmidt, and R. Drautz. New Superconducting and Semiconducting Fe-B Compounds

Predicted with an Ab Initio Evolutionary Search. *Physical Review Letters*, 105(21):1–4, November 2010.

[69] M. Ji, C.-Z. Wang, and K.-M. Ho. Comparing efficiencies of genetic and minima hopping algorithms for crystal structure prediction. *Physical Chemistry Chemical Physics: PCCP*, 12(37):11617–23, October 2010.

[70] L. Pauling. The principles determining the structure of complex ionic crystals. *Journal of the American Chemical Society*, 51:1010–1026, 1929.

[71] P. Villars. A three-dimensional structural stability diagram for 998 binary AB intermetallic compounds. *Journal of the Less Common Metals*, 92(2):215–238, August 1983.

[72] D. G. Pettifor. Structure Maps in Alloy Design. *Journal Of The Chemical Society Faraday Transactions*, 86(8):1209–1213, 1990.

[73] D. G. Pettifor. Structure maps revisited. *Journal of Physics: Condensed Matter*, 15:13–16, 2003.

[74] D. Morgan, J. Rodgers, and G. Ceder. Automatic construction, implementation and assessment of Pettifor maps. *Journal of Physics: Condensed Matter*, 15:4361–4369, 2003.

[75] T. S. Bush, C. R. A. Catlow, and P. D. Battle. Evolutionary Programming Techniques for predicting Inorganic Crystal Structures. *Journal of Materials Chemistry*, 5(8):1269–1272, October 1995.

[76] A. R. Oganov, J. Chen, C. Gatti, Y. Ma, Y. Ma, C. W. Glass, Z. Liu, T. Yu, O. O. Kurakevych, and V. L. Solozhenko. Ionic high-pressure form of elemental boron. *Nature*, 457(February):863–868, 2009.

[77] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* The MIT Press, 2001.

[78] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning.* Springer, 2003.

[79] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder. Predicting Crystal Structures with Data Mining of Quantum Calculations. *Physical Review Letters*, 91(13):1–4, 2003.

[80] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder. Predicting crystal structure by merging data mining with quantum mechanics. *Nature materials*, 5(8):641–6, August 2006.

[81] T. Morita. Cluster Variation Method of Cooperative Phenomena and its Generalization I. *Journal of the Physical Society of Japan*, 12(7):753–755, 1957.

[82] C. C. Fischer. *A Machine Learning Approach to Crystal Structure Prediction.* PhD thesis, Massachusetts Institute of Technology, 2007.

[83] S. R. Eliason. *Maximum Likelihood Estimation: Logic and Practice.* Sage Publications, Inc, 1993.

[84] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[85] R. S. J. Lynch and P. K. Willett. Adaptive Bayesian classification using noninformative Dirichlet priors. *IEEE Transactions on Systems, Man, and Cybernetics*, 33(3):2812–2815, 2003.

[86] W. Buntine. Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, volume 91, pages 52–60. Citeseer, 1991.

[87] ICSD. Inorganic Crystal Structure Database, 2006.

[88] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*, chapter 4, pages 80–113. Springer, 2009.

[89] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[90] ICDD. Powder Diffraction File PDF4+, 2008.

[91] K. S. Roh, K. H. Ryu, and C. H. Yo. Nonstoichiometry and Physical Properties of the $SrSn_{1-x}Fe_xO_{3-y}$ System. *Journal of Solid State Chemistry*, 142(2):288–293, 1999.

[92] G. Scarel, A. Svane, and M. Fanciulli. *Scientific and technological issues related to rare earth oxides: An introduction*, volume 106, pages 1–14. Springer, 2007.

[93] S. Uma and J. Gopalakrishnan. Synthesis of Novel Oxide Pyrochlores $A_2BB'O_7$ (A = La, Nd; BB' = Pb, Sn, Bi), by Alkali Melt Route. *Journal of Solid State Chemistry*, 105:595–598, 1993.

[94] M. O'Keeffe and B. G. Hyde. Stoichiometry and the structure and stability of inorganic solids. *Nature*, 309(5967):411–414, May 1984.

[95] M. Jansen and R. Hoppe. Neue Oxocobaltate (IV):$Cs_2[CoO_3]$, $Rb_2[CoO_3]$ und $K_2[CoO_3]$. *Zeitschrift für anorganische und allgemeine Chemie*, 408:75–82, 1974.

[96] H. Kedesky and A. Drukalsky. X-Ray Diffraction Studies of the Solid State Reaction in the NiO-ZnO System. *Journal of the American Chemical Society*, 76(23):5941–5946, 1954.

[97] S. Matar, I. Baraille, and M. Subramanian. First principles studies of $SnTiO_3$ perovskite as potential environmentally benign ferroelectric material. *Chemical Physics*, 355(1):43–49, January 2009.

[98] B. Chamberland, A. W. Sleight, and J. F. Weiher. Preparation and characterization of $MgMnO_3$ and $ZnMnO_3$. *Journal of Solid State Chemistry*, 1(3-4):512–514, April 1970.

[99] A. Riou and A. Lecerf. Structure cristalline de $Co_2Mn_3O_8$. *Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry*, 31(10):2487–2490, 1975.

[100] D. Johrendt and R. Pöttgen. Pnictide oxides: a new class of high-T_C superconductors. *Angewandte Chemie (International ed. in English)*, 47(26):4782–4, January 2008.

[101] V. Goldschmidt. Die gesetze der krystallochemie. *Naturwissenschaften*, 14:477–485, 1926.

[102] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19:263–312, 1993.

[103] A. Berger, V. J. Della Pietra, and S. A. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–72, 1996.

[104] S. A. Della Pietra, V. J. Della Pietra, and J. Lafferty. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):1–13, 1997.

[105] E. Parthé and L. Gelato. The standardization of inorganic crystal-structure data. *Acta Crystallographica Section A*, 40:169–183, 1984.

[106] R. Hundt, J. C. Schön, and M. Jansen. CPMZ-an algorithm for the efficient comparison of periodic structures. *Journal of Applied Crystallography*, 39:6–16, 2006.

[107] E. Gaudin, F. Boucher, and M. Evain. Some Factors Governing Ag^+ and Cu^+ Low Coordination in Chalcogenide Environments. *Journal of Solid State Chemistry*, 160(1):212–221, August 2001.

[108] H. Zhang, N. Li, K. Li, and D. Xue. Structural stability and formability of ABO3-type perovskite compounds. *Acta crystallographica. Section B, Structural science*, 63(Pt 6):812–8, December 2007.

[109] G.-A. Nazri and G. Pistoia, editors. *Lithium Batteries: Science and Technology.* Springer, 2003.

[110] M. S. Whittingham. Lithium Batteries and Cathode Materials. *Chemical Reviews*, 104(10):4271–4302, October 2004.

[111] J. B. Goodenough and Y. Kim. Challenges for Rechargeable Li Batteries. *Chemistry of Materials*, 22(3):587–603, February 2010.

[112] B. L. Ellis, K. T. Lee, and L. F. Nazar. Positive Electrode Materials for Li-Ion and Li-Batteries. *Chemistry of Materials*, 22(3):691–714, February 2010.

[113] R. A. Huggins. *Advanced Batteries: Materials Science Aspects*. Springer, 2008.

[114] J. Cabana, L. Monconduit, D. Larcher, and M. R. Palacín. Beyond Intercalation-Based Li-Ion Batteries: The State of the Art and Challenges of Electrode Materials Reacting Through Conversion Reactions. *Advanced Materials*, 22:E170–E192, August 2010.

[115] Y. S. Meng and M. E. Arroyo-de Dompablo. First principles computational materials design for energy storage materials in lithium ion batteries. *Energy & Environmental Science*, 2(6):589, 2009.

[116] M. Aydinol, A. Kohan, G. Ceder, K. Cho, and J. Joannopoulos. Ab initio study of lithium intercalation in metal oxides and metal dichalcogenides. *Physical Review B*, 56(3):1354–1365, 1997.

[117] F. Zhou, M. Cococcioni, K. Kang, and G. Ceder. The Li intercalation potential of $LiMPO_4$ and $LiMSiO_4$ olivines with M=Fe, Mn, Co, Ni. *Electrochemistry Communications*, 6(11):1144–1148, 2004.

[118] J. Heyd, G. E. Scuseria, and M. Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics*, 118(18):8207, 2003.

[119] S. P. Ong, A. Jain, G. Hautier, B. Kang, and G. Ceder. Thermal stabilities of delithiated olivine MPO_4 (M=Fe, Mn) cathodes investigated using first principles calculations. *Electrochemistry Communications*, 4:1–4, 2010.

[120] A. Van Der Ven, M. K. Aydinol, and G. Ceder. First-Principles Evidence for Stage Ordering in Li_xCoO_2. *Journal of The Electrochemical Society*, 145(6):2149, 1998.

[121] J. Bhattacharya and A. Van Der Ven. Phase stability and nondilute Li diffusion in spinel $Li_{1+x}Ti_2O_4$. *Physical Review B*, 81(10), March 2010.

[122] K. Kang, D. Morgan, and G. Ceder. First principles study of Li diffusion in I-Li_2NiO_2 structure. *Physical Review B*, 79(1):1–4, 2009.

[123] D. Morgan, A. Van Der Ven, and G. Ceder. Li Conductivity in Li_xMPO_4 (M=Mn, Fe, Co, Ni) Olivine Materials. *Electrochemical and Solid-State Letters*, 7(2):A30, 2004.

[124] K. Kang, Y. S. Meng, J. Bréger, C. P. Grey, and G. Ceder. Electrodes with high power and high capacity for rechargeable lithium batteries. *Science*, 311(5763):977–980, 2006.

[125] T. Maxisch, F. Zhou, and G. Ceder. Ab initio study of the migration of small polarons in olivine Li_xFePO_4 and their association with lithium ions and vacancies. *Physical Review B*, 73(10):1–6, 2006.

[126] P. Arora, R. White, and M. Doyle. Capacity Fade Mechanisms and Side Reactions in Lithium-Ion Batteries. *Journal of the Electrochemical Society*, 145(10):3647, 1998.

[127] A. J. Bard and L. R. Faulkner. *Electrochemical methods.* John Wiley, 2001.

[128] G. L. W. Hart and R. W. Forcade. Algorithm for generating derivative structures. *Physical Review B*, 77(22):1–12, 2008.

[129] J. C. Kim, C. Moore, B. Kang, G. Hautier, A. Jain, and G. Ceder. Synthesis and electrochemical properties of monoclinic $LiMnBO_3$ as a Li intercalation material. *to be published in the Journal of The Electrochemical Society*, 2011.

[130] A. Armstrong and P. Bruce. Synthesis of layered $LiMnO_2$ as an electrode for rechargeable lithium batteries. *Nature*, 381:499–500, 1996.

[131] J. Gaubicher, C. Wurm, G. Goward, C. Masquelier, and L. Nazar. Rhombohedral Form of $Li_3V_2(PO_4)_3$ as a Cathode in Li-Ion Batteries. *Chemistry of Materials*, 12(11):3240–3242, 2000.

[132] A. Khomyakov. Sidorenkite, $Na_3Mn(PO_4)(CO_3)$, a new mineral. *International Geology Review*, 22(7):811–814, July 1980.

[133] T. A. Kurova, N. G. Shumyatskaya, A. A. Vornkov, and Y. A. Pyatenko. Refinement of the crystal structure of sidorenkite ($Na_3Mn(PO_4)(CO_3)$). *Mineralogiceskij Zhurnal*, 2:65–70, 1980.

[134] C. T. L. Tjy, T. N. Nadezhina, E. A. Pobedimskaya, and A. P. Khomyakov. Crystal chemical characteristics of bradleyite, sidorenkite, and bonshtedtite. *Mineralogiceskij Zhurnal*, 6(5):79–84, 1984.

[135] A. Padhi, K. Nanjundaswamy, and J. B. Goodenough. Phospho-olivines as Positive-Electrode Materials for Rechargeable Lithium Batteries. *Journal of the Electrochemical Society*, 144(4):1188, 1997.

[136] B. Kang and G. Ceder. Battery materials for ultrafast charging and discharging. *Nature*, 458(March):190–193, 2009.

[137] S.-i. Nishimura, M. Nakamura, R. Natsui, and A. Yamada. New lithium iron pyrophosphate as 3.5 V class cathode material for lithium ion battery. *Journal of the American Chemical Society*, 132(39):13596–7, October 2010.

[138] H. Zhou, S. Upreti, N. A. Chernova, G. Hautier, G. Ceder, and M. S. Whittingham. Iron and Manganese Pyrophosphates as Cathodes for Lithium-Ion Batteries. *Chemistry of Materials*, (6), December 2010.

[139] J. Lima-de Faria, E. Hellner, F. Liebau, E. Makovicky, and E. Parthe. Report of the International Union of Crystallography Commission on Crystallographic Nomenclature of Inorganic Structure Types. *Acta Crystallographica Section A*, 46:1–11, 1990.

[140] H. Burzlaff and Y. Malinovsky. A Procedure for the Clasification of Non-Organic Crystal Structures. I. Theoretical Background. *Acta Crystallographica Section A Foundations of Crystallography*, 53(2):217–224, March 1997.

[141] K. Shoemake and T. Duff. Matrix Animation and Polar Decomposition.

[142] R. Allmann and R. Hinek. The introduction of structure types into the Inorganic Crystal Structure Database ICSD. *Acta crystallographica. Section A, Foundations of crystallography*, 63(Pt 5):412–7, September 2007.