

# Finding Nature's Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory

Geoffroy Hautier, Christopher C. Fischer, Anubhav Jain, Tim Mueller, and  
Gerbrand Ceder\*

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

Received March 18, 2010. Revised Manuscript Received May 15, 2010

Finding new compounds and their crystal structures is an essential step to new materials discoveries. We demonstrate how this search can be accelerated using a combination of machine learning techniques and high-throughput *ab initio* computations. Using a probabilistic model built on an experimental crystal structure database, novel compositions that are most likely to form a compound, and their most-probable crystal structures, are identified and tested for stability by *ab initio* computations. We performed such a large-scale search for new ternary oxides, discovering 209 new compounds with a limited computational budget. A list of these predicted compounds is provided, and we discuss the chemistries in which high discovery rates can be expected.

## 1. Introduction

New compounds often form the basis for new materials discoveries. For example, recent advances in high-critical-temperature (high- $T_c$ ) superconductor or thermoelectric materials find their roots in the discovery of previously unknown crystalline phases.<sup>1,2</sup> It is of scientific and technological interest to know how many undiscovered compounds nature still has in store, as well as their compositions. Density functional theory (DFT) and its extensions<sup>3–5</sup> have been shown to model the ground-state and finite temperature behavior for a wide variety of compounds accurately. In particular, phase stability has been shown to be efficiently and accurately accessible through DFT computations in many different chemistries.<sup>6–13</sup>

As noted by Woodley et al., the current challenge in computationally predicting compounds is not the computation, but *what* to calculate.<sup>14</sup> Simply put, one must decide which chemical compositions likely form compounds and

warrant investigation, *and* one must have a method to suggest possible crystal structures of the compound. In this paper, we use large-scale *ab initio* calculations, together with machine learning techniques, to explore the ternary oxide chemical space in search of missing compounds. Several hundreds of potential new compounds and their crystal structure are identified and provided in the online information. We believe that these types of knowledge-driven high-throughput computational techniques can form the basis of accelerated materials discovery.

Algorithms for crystal structure prediction are typically based on optimization of the relevant thermodynamic potential in the space of atomic coordinates, using techniques such as genetic algorithms<sup>15–18</sup> or simulated annealing.<sup>19–21</sup> Although successful in several focused cases of compound prediction, coordinate-based optimization techniques are simply too computationally costly for tasks involving more than a handful of candidate chemical compositions. If, instead, one is searching in the space of both chemical composition and structure, an approach for deciding which compositions are likely to yield new compounds is needed, in addition to a fast structure prediction method.

In this paper, we use a recently introduced method based on data mining to search across all ternary oxides for new compounds.<sup>22</sup> Broadly speaking, the technique

\*Author to whom correspondence should be addressed. E-mail: gceder@mit.edu.

- (1) Snyder, G.; Toberer, E. *Nat. Mater.* **2008**, *7*, 105–114.
- (2) Johrendt, D.; Pöttgen, R. *Angew. Chem., Int. Ed.* **2008**, *47*, 4782–4.
- (3) Hafner, J.; Wolverton, C.; Ceder, G. *MRS Bull.* **2006**, *31*, 659–668.
- (4) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864–B871.
- (5) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- (6) Ong, S. P.; Wang, L.; Kang, B.; Ceder, G. *Chem. Mater.* **2008**, *20*, 1798–1807.
- (7) Ozoliņš, V.; Wolverton, C.; Zunger, A. *Phys. Rev. B* **1998**, *57*, 6427–6443.
- (8) Hart, G. L. W. *Phys. Rev. B* **2009**, *80*, 1–5.
- (9) Curtarolo, S.; Morgan, D.; Ceder, G. *CALPHAD: Comput. Coupling Phase Diagrams Thermochem.* **2005**, *29*, 163–211.
- (10) Akbarzadeh, A. R.; Wolverton, C.; Ozoliņš, V. *Phys. Rev. B* **2009**, *79*, 1–10.
- (11) Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I. B.; Norskov, J. K. *Nat. Mater.* **2006**, *5*, 909–913.
- (12) Hummelshøj, J. S.; et al. *J. Chem. Phys.* **2009**, *131*, 014101.
- (13) Levy, O.; Chepulskii, R. V.; Hart, G. L. W.; Curtarolo, S. *J. Am. Chem. Soc.* **2010**, *132* (2), 833–837.
- (14) Woodley, S. M.; Catlow, R. *Nat. Mater.* **2008**, *7*, 937–46.

- (15) Abraham, N. L.; Probert, M. I. *J. Phys. Rev. B* **2006**, *73*, 224104.
- (16) Woodley, S. M.; Battle, P. D.; Gale, J. D.; Catlow, C. R. A. *Phys. Chem. Chem. Phys.* **1999**, *1*, 2535–2542.
- (17) Oganov, A. R.; Glass, C. W. *J. Chem. Phys.* **2006**, *124*, 244704.
- (18) Glass, C.; Oganov, A.; Hansen, N. *Comput. Phys. Commun.* **2006**, *175*, 713–720.
- (19) Schön, J. C.; Jansen, M. *Angew. Chem., Int. Ed.* **1996**, *35*, 1287–1304.
- (20) Schön, J. C.; Jansen, M. *Z. Kristallogr.* **2001**, *216*, 307–325.
- (21) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368–1372.
- (22) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. *Nat. Mater.* **2006**, *5*, 641–646.

uses a machine learning approach to extract the “chemical rules” that govern structure selection from an experimental database of crystal structure information. Using this learned information, it suggests likely compound forming compositions and candidate structures in new systems. By combining this with an accurate energy model, such as DFT, to sort through the candidate list, we effectively predict several hundred new ternary oxide compounds.

## 2. Methodology

### 2.1. Compound and Structure Prediction Procedure.

We briefly present the method here. Let the variable  $x_{c_i}$  indicate what crystal structure is present at a given composition  $c_i$ . For example, if  $c_i$  represents the composition  $AB_2C_4$ , then  $x_{c_i}$  may have values such as spinel, olivine, etc. The condition “ $x_{c_i} = \text{nostructure}$ ” indicates the absence of a compound at the given composition. In addition, variables representing the system’s constituents (e.g.,  $E_i = \text{Ag, Cu, Na, etc.}$ ) are defined. With these definitions, any chemical system of  $C$  constituents and  $n$  compositions can be represented by a vector  $\vec{X} = (x_{c_1}, x_{c_2}, \dots, x_{c_n}, x_{E_1}, x_{E_2}, \dots, x_{E_C})$ , where the composition space is discretized using  $n$  composition bins.

The probability density  $p(\vec{X})$  is the essential information that we require, because it provides information regarding 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 ( $x_{c_j}$ ) is likely to be compound-forming. Mathematically, this is evaluated by computing the probability of forming a compound:

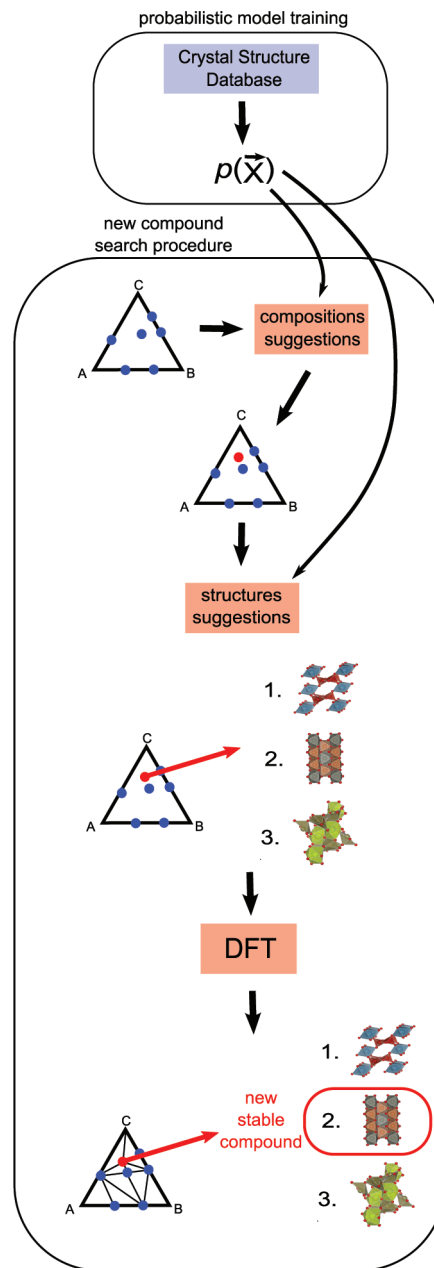
$$p_{\text{compound}}(x_{c_j}) = 1 - p(x_{c_j} = \text{nostructure} | x_{c_1}, x_{c_2}, \dots, x_{c_{j-1}}, x_{c_{j+1}}, \dots, x_{c_n}, x_{E_1}, x_{E_2}, \dots, x_{E_C}) \quad (1)$$

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

$$p(x_{c_j} | x_{c_1}, x_{c_2}, \dots, x_{c_{j-1}}, x_{c_{j+1}}, \dots, x_{c_n}, x_{E_1}, x_{E_2}, \dots, x_{E_C}) \quad (2)$$

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 1.

**2.2. Training and Testing of the Ternary Oxide Probabilistic Model.** To search for new ternary oxides, we trained the probabilistic model presented above on the oxide experimental data available in the Inorganic Crystal Structure Database (ICSD).<sup>23</sup> The ICSD is the most



**Figure 1.** 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 most likely to form a compound (denoted by the red dots). For those compositions, the most likely crystal structures are proposed using the same probabilistic model. These structure candidates are then tested for stability, using an accurate energy model, such as DFT.

complete crystal structure database for oxides. It contains ~5000 unique ternary oxide compounds distributed across 1581 chemical systems.

The 2006 version of the ICSD was searched for duplicate compounds, using a robust affine mapping technique.<sup>24,25</sup> Two entries are considered to represent the same compound if their crystal structures have the same space group and can

(23) ICSD, *Inorganic Crystal Structure Database*; Fachinformationszentrum Karlsruhe: Karlsruhe, Germany, 2006.

(24) Burzlaff, H.; Malinovsky, Y. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1997**, 53, 217–224.

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

be transformed onto each other 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, a total of 183 variables was used in the probability model.

The probability function  $p(\vec{X})$  is a complex multivariable function that must be approximated, for all practical purposes. In this work, we approximate it using a cumulant expansion truncated after the pair terms, following Fischer et al.<sup>22</sup>

$$p(\mathbf{X}) \approx \frac{1}{Z} \prod_i p(x_i) \prod_{j < k} \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \quad (3)$$

The parameters to be estimated from the available data are the point  $\{p(x_i)\}$  and pair  $\{p(x_i, x_j)\}$  terms. These parameters are estimated using a Bayesian estimation approach and a Dirichlet prior, as presented in Fischer et al.<sup>22</sup>

During predictions, only compositions allowed by the possible known oxidation states of the element were allowed.

We tested the compound discovery procedure outlined in the previous section 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. This cross-validation procedure confirmed that our approach is predictive. We find the correct crystal structure 95% of the time, using 5–20 DFT calculations (see Supporting Information).

**2.3. *Ab Initio* Density Functional Theory Computation Parameters.** All *ab initio* computations have been performed using the Vienna *ab initio* Simulation package (VASP)<sup>26</sup> within the projector augmented-wave approach,<sup>27</sup> using the generalized gradient exchange and correlation functional parametrized by Perdew, Burke, and Ernzerhof.<sup>28</sup> We used a U parameter<sup>29</sup> for some of the transition metals (V, Cr, Mn, Fe, Co, Ni, Cu, Nb, Mo, Ag, Ta, W). This U parameter was fitted following Wang's method<sup>30</sup> and using binary oxide enthalpy formation energies from the Kubaschewski tables.<sup>31</sup> The energy cutoff was 30% higher than the maximal energy cutoff specified by the pseudopotential. For oxides, oxygen has the maximal specified energy cutoff of 400 meV, constraining all oxides run to be performed with a cutoff energy of 520 meV.

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.<sup>32</sup> A  $\Gamma$ -centered grid was used for hexagonal cells.

All computations were performed with spin polarization. All magnetic moments were initialized in a ferromagnetic configuration with high spin for the 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. Two runs were performed with Co: one initialized high spin and the other initialized low spin. Only the computation with the lowest energy was considered.

Relaxation was performed using the AFLOW software.<sup>9</sup> The first relaxation attempt was performed using a mixture of RMM-DIIS<sup>33</sup> and Davidson algorithm<sup>34</sup> as the diagonalization scheme. If the first attempt did not converge, a second try was performed using only the Davidson algorithm, which is known to be slower but more stable. Mixing parameters were also modified for this second attempt.

Using these parameters, all element, binary, and ternary compounds present in the ICSD database were computed. The oxide compounds having partial occupancies were related to an ordered structure at the same or close composition using an enumeration technique similar to the one proposed by Hart et al.<sup>35</sup> and picking the structure with the lowest electrostatic energy, as determined via Ewald summation.<sup>36</sup>

In this work, any crystal structure containing more than 100 atoms in its unit cell was not considered for computation, because of resource limitations.

### 3. Results and Discussion

Using this probabilistic approach, we searched for new compounds in 2211 A–B–O systems, with A and B taken from the following pool of elements: 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 (eq 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 1261 possible compositions and exhibited a 45% true positive rate during cross-validation (see the Supporting Information). At these selected compositions, the most likely crystal structures were determined

(26) Kresse, G.; Furthmüller, J. *Comput. Mater. Sci.* **1996**, *6*, 15.

(27) Blöchl, P. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(28) Perdew, J.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(29) Anisimov, V. I.; Aryasetiawan, F.; Lichtenstein, A. I. *J. Phys.: Condens. Matter* **1997**, *9*, 767–808.

(30) Wang, L.; Maxisch, T.; Ceder, G. *Phys. Rev. B* **2006**, *73*, 1–6.

(31) Kubaschewski, O.; Alcock, C. B.; Spencer, P. J. In *Thermochemical Data*, Sixth ed.; Pergamon Press: New York, 1993; Chapter 5, pp 257–323.

(32) Monkhorst, H.; Pack, J. *Phys. Rev. B* **1976**, *13*, 5188.

(33) Wood, D. M.; Zunger, A. *J. Phys. A* **1985**, *18*, 1343–1359.

(34) Davidson, E. R. Matrix Eigenvector Methods. In *Methods in Computational Molecular Physics*; NATO ASI Ser. C, No. 113; Diercksen, G. H. F., Ed.; Kluwer: Dordrecht, The Netherlands, 1983; p 95.

(35) Hart, G. L. W.; Forcade, R. W. *Phys. Rev. B* **2008**, *77*, 1–12.

(36) Toukmaji, A.; Board, J. A. *Comput. Phys. Commun.* **1996**, *95*, 73–92.

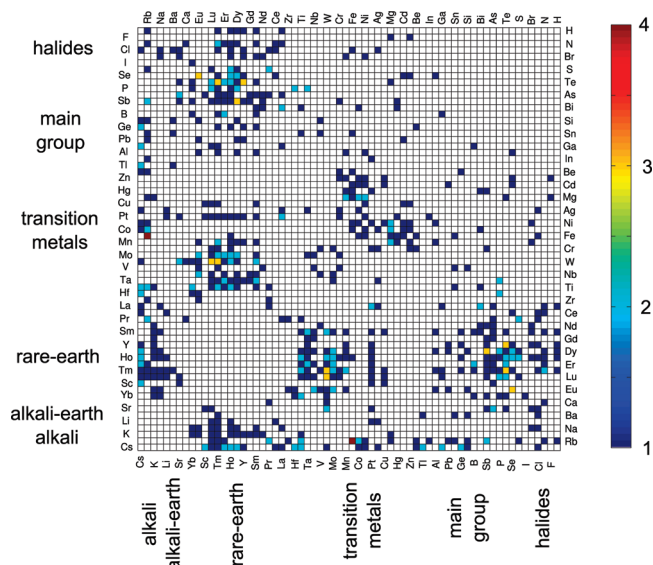


from the data-mined probability density using eq 2. The number of suggested crystal structures at each composition corresponds to the list length that gave 95% accuracy in cross-validation. This corresponds to a total of 5546 crystal structures whose energy needed to be calculated with *ab initio* DFT. All existing binary, ternary, and elemental structures in the ICSD were also calculated so that relative phase stability can be assessed. 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 1261 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 2211 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 1261 compositions bins would require substantially more computations (183 007) than the 5546 needed, while fully using the machine learned model.

While these predictions are made for chemical compositions that have no compounds in the ICSD database, we compared them to another major compound database available: the PDF4+ database from the International Centre for Diffraction Data (ICDD).<sup>37</sup> This database also contains compounds for which only a composition and a powder X-ray diffraction (XRD) 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 do not have 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 of these 146 predicted crystal structures to the PDF4+ entries revealed that only six of our predicted results did not match the structural data present in their corresponding PDF4+ entries (see the Supporting Information). Since no information at all from the PDF4+ database was used to train our model, this high success rate supports our method and strengthens the confidence in the rest of our predictions (209 compounds).

The different chemistries over which these new compounds are distributed can be analyzed. Figure 2 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 Mendelev number.<sup>38</sup> This ordering allows us to spot the different chemical classes in which new compounds



**Figure 2.** 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 Mendelev number.

have been found directly. Figure 2 indicates that the predictions span many different chemistries. Most striking is the absence of any new compounds in mixtures of rare-earth elements. The difficulty involved in forming energetically favorable ternary oxides that contain two rare-earth elements relates the important electrostatic component to the phase stability in oxides. Indeed, many of the rare-earth compounds usually exhibit a +3 oxidation state and combining isovalent cations rarely leads to the formation of highly stable compounds. 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  $\text{La}_2\text{Pr}_2\text{O}_7$  compound, which combined  $\text{La}^{3+}$  and  $\text{Pr}^{4+}$  ions. Pr, along with Ce and Tb, are the only rare-earth elements that exhibit a +4 oxidation state in oxides.

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 that have been stabilized by entropic mixing effects.

The last major region without predictions is situated in the upper right corner of Figure 2 and concerns the oxides of two main group elements. Only 3 successes from 40 suggested compositions were obtained in this chemical space. However, analysis of the cross-validation results did not show any evidence, which suggests a systematic failure of our approach in this region (see the Supporting Information). 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

(37) PDF4+, Powder Diffraction File; International Centre for Diffraction Data: Newtown Square, PA, 2008.

(38) Pettifor, D. G. *J. Phys. C: Solid State Phys.* **1986**, *19*, 285–313.

much as other chemistries; therefore, it presents more opportunities for our approach to identify unknown compositions and their crystal structures. However, rare-earth compounds attract great scientific and technological interest, because of their unique catalytic, optical, or magnetic properties.<sup>39</sup> Among those, it is interesting to see that our algorithm picked up the correct  $\text{La}_2\text{Zr}_2\text{O}_7$  structure for the  $\text{La}_2\text{Bi}_2\text{O}_7$  compound present in the PDF4+ database.<sup>40</sup> 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 possible discovery of many new compounds. Among these, it is striking to find many  $\text{Co}^{4+}$  compounds. While  $\text{Co}^{4+}$  requires very strong oxidation conditions in most chemistries, our analysis of the oxygen chemical potential in which the compound is stable (see the Supporting Information) indicates that it may be much easier to form  $\text{Co}^{4+}$  when combined with Cs or Rb. For instance, one should be able to synthesize predicted compounds such as  $\text{Co}_2\text{Cs}_6\text{O}_7$ ,  $\text{Co}_2\text{Rb}_6\text{O}_7$ , and  $\text{CoRb}_2\text{O}_3$  under relatively mild oxidizing conditions. The stabilizing effect of heavy alkali toward the high oxidation state of a transition metal is a chemical rule that is known by some chemists and has been reported, for instance, by O'Keefe.<sup>41</sup> In addition, one of those three compositions ( $\text{CoRb}_2\text{O}_3$ ) is linked to a powder diffraction pattern (without exact structural information) in the PDF4+ database. Comparison between this PDF4+ pattern and that simulated from our prediction shows good agreement (see the Supporting Information).

Although 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 under favorable conditions. For instance, while it is surprising to make a prediction in a rather common chemical system such  $\text{Ni-Zn-O}$ , the predicted compound (a  $\text{Ni}^{3+}$ -containing  $\text{Ni}_2\text{ZnO}_4$  spinel) requires a more oxidizing environment than the  $\text{Ni}^{2+}$ -containing  $\text{NiO-ZnO}$  solid solution present in the ICSD.<sup>42</sup>

Similarly, while many  $\text{SnO}_2\text{-TiO}_2$  solid solutions are in the ICSD, there is no known  $\text{Sn}^{2+}\text{-Ti}^{4+}$  compound. Our analysis predicts a  $\text{SnTiO}_3$  ilmenite that is stable in a reducing environment. This  $\text{SnTiO}_3$  ilmenite prediction is of techno-

logical interest, because  $\text{SnTiO}_3$  perovskite has been predicted through *ab initio* computation to be a good candidate lead-free ferroelectric material.<sup>43</sup> 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  $\text{Mg-Mn-O}$  system:  $\text{MgMnO}_3$  and  $\text{Mg}_2\text{Mn}_3\text{O}_8$ , both of which are compounds that have the much more common +4 oxidation state for Mn. The PDF4+ database actually contains a powder diffraction pattern for  $\text{MgMnO}_3$ . Comparison of the powder diffraction pattern of our predicted structure to that presented in this database shows a reasonable match and confirms our prediction of an ilmenite structure (see the Supporting Information). On the other hand, the  $\text{Mg}_2\text{Mn}_3\text{O}_8$  compound is totally unknown from literature and is predicted to crystallize in a very uncommon structure, exhibited by  $\text{Co}_2\text{Mn}_3\text{O}_8$ .<sup>44</sup> 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 data, and remarks) is available in the Supporting Information and on the Internet at <http://ceder.mit.edu/ternaryoxides>. It is our belief that researchers could benefit from access to theoretically predicted compounds, in addition to the standard crystal structure and powder diffraction experimental database.

While our approach is fast and efficient for discovering new compounds, it has limitations. It is possible that we missed a true ground state, because of 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.

#### 4. Conclusion

In summary, we have used a combination of data mining and first principles computations to look for undiscovered compound forming compositions in  $\text{A}_x\text{B}_y\text{O}_z$  systems and predicted their structure. The 355 new compounds suggested were obtained within ~55 days of computing on 400 Intel Xeon 5140 2.33 Ghz cores, indicating the efficiency by which new compounds can be discovered computationally. One can compare this to the experimental discovery rate as measured by the number of new entries each year in the ICSD, which has been averaging about 100 per year since the

(39) Scarel, G.; Svane, A.; Fanciulli, M. In *Scientific and Technological Issues Related to Rare Earth Oxides: An Introduction*; Springer: Berlin, 2007; Vol. 106, pp 1–14.

(40) Uma, S.; Gopalakrishnan, J. *J. Solid State Chem.* **1993**, *105*, 595–598.

(41) O'Keefe, M.; Hyde, B. G. *Nature* **1984**, *309*, 411–414.

(42) Kedesky, H.; Drukalsky, A. *J. Am. Chem. Soc.* **1954**, *76*, 5941–5946.

(43) Matar, S.; Baraille, I.; Subramanian, M. *Chem. Phys.* **2009**, *355*, 43–49.

(44) Riou, A.; Lecerf, A. *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1975**, *31*, 2487–2490.

1980s (see the Supporting Information). Hence, we believe that our proposed approach represents an exciting opportunity for novel materials discovery by theorists and experimentalists alike.

**Acknowledgment.** This work was supported by the NSF (under Contract No. DMR-0606276). The authors thank Robert Doe for helpful discussion and Charles Moore and Will Tipton for their help in the development of the ab initio high-throughput infrastructure. G.H. would like to thank the Belgian American Education Foundation (BAEF) and Total for their financial support. A.J. would like to acknowledge funding from the U.S. Department of Energy Department of Energy

Computational Science Graduate Fellowship (DOE CSGF) (under Grant No. DE-FG02-97ER25308).

**Supporting Information Available:** Details on the cross-validation procedure details and results, comparison of the predictions to the PDF4+ database, analysis of the main group mixture chemical region, more details on the  $\text{Co}^{4+}$  and  $\text{MgMnO}_3$  compounds, and a plot of the rate of new ternary oxides entering the ICSD. Stability results and crystallographic data (in cif format) are also provided for all predicted compounds in PDF format and via the Internet at <http://ceder.mit.edu/ternaryoxides>. This material is available free of charge via the Internet at <http://pubs.acs.org>.