Predicting the stability of ternary intermetallics with density functional theory and machine learning

Cite as: J. Chem. Phys. 148, 241728 (2018); https://doi.org/10.1063/1.5020223
Submitted: 21 December 2017. Accepted: 02 April 2018. Published Online: 23 April 2018

Jonathan Schmidt, Liming Chen, Silvana Botti, and Miguel A. L. Marques

COLLECTIONS

Paper published as part of the special topic on Data-Enabled Theoretical Chemistry

ARTICLES YOU MAY BE INTERESTED IN

SchNet – A deep learning architecture for molecules and materials
The Journal of Chemical Physics 148, 241722 (2018); https://doi.org/10.1063/1.5019779

Machine learning of molecular properties: Locality and active learning
The Journal of Chemical Physics 148, 241727 (2018); https://doi.org/10.1063/1.5005095

Alchemical and structural distribution based representation for universal quantum machine learning
The Journal of Chemical Physics 148, 241717 (2018); https://doi.org/10.1063/1.5020710
Predicting the stability of ternary intermetallics with density functional theory and machine learning

Jonathan Schmidt,1 Liming Chen,2 Silvana Botti,3 and Miguel A. L. Marques1,a)
1Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, D-06099 Halle, Germany
2Liris laboratory UMR CNRS 5205 Ecole Centrale de Lyon, University of Lyon, 36 avenue Guy de Collongue 69134 Ecully Cedex, France
3Institut für Festkörpertheorie und -Optik, Friedrich-Schiller-Universität Jena and European Theoretical Spectroscopy Facility, Max-Wien-Platz 1, 07743 Jena, Germany

(Received 21 December 2017; accepted 2 April 2018; published online 23 April 2018)

We use a combination of machine learning techniques and high-throughput density-functional theory calculations to explore ternary compounds with the AB2C2 composition. We chose the two most common intermetallic prototypes for this composition, namely, the t110-CeAl2Ga2 and the tP10-FeMo2B2 structures. Our results suggest that there may be ~10 times more stable compounds in these phases than previously known. These are mostly metallic and non-magnetic. While the use of machine learning reduces the overall calculation cost by around 75%, some limitations of its predictive power still exist, in particular, for compounds involving the second-row of the periodic table or magnetic elements. Published by AIP Publishing. https://doi.org/10.1063/1.5020223

I. INTRODUCTION

In recent years, the increased availability of faster computers, reliable computer codes, and accurate theories led to a revolution in materials science. It is by now possible to perform computational studies of large classes of materials and to calculate many of their properties in a systematic manner. This has led not only to a better understanding of the underlying physics and of trends, but also to a wealth of predictions that can then be submitted to experimental verification.

One of the most exciting aspects of this revolution is the possibility to discover completely new materials. In this context, an important concept is thermodynamic stability.1 A thermodynamically stable material does not decompose (possibly in an infinite time) to any other phase or compound. Clearly, we can also synthesize meta-stable materials (diamond is perhaps the best known example), but thermodynamically stable compounds are generally expected to be easier to produce and to use in technological applications. Moreover, this concept is trivial to express from a theoretical perspective: a thermodynamically stable compound is the one that forms a vertex of the convex hull of stability.

The path for the theoretical prediction of new materials is therefore clear. Calculate the energy of all possible new compounds and obtain the distance $E_{\text{hull}}$ between the formation energy and the convex hull of stability. If the new compound is above the hull, it is unstable (or meta-stable), and $E_{\text{hull}}$ is the energy released by decomposing to the more stable compounds. If this energy is negative, this compound is stable, and the hull should be recalculated to include it as a new vertex. Of course, as our knowledge of stable materials is rather incomplete, the convex hull is a moving target that evolves as new stable materials are found and added to our databases.

Several systematic investigations following this or a similar path have already appeared in the literature (for example, Refs. 2–4), and a wealth of calculated data can already be found in online databases.5–7 These calculations are in general based on density functional theory (DFT), as this is the only theory that combines a reasonable precision with relatively light computational requirements. These studies also often concentrate in a few given structure prototypes that are known experimentally. The motivation for this is that chemical elements can often be interchanged in a solid to form new stable materials.

Systematic approaches have the undeniable advantage of being exhaustive. Unfortunately, the search space of possible new compounds is rather large. Let us take ternary compounds as an example, for which we can find more than $1.3 \times 10^6$ entries in aflowlib8 (although mostly for simple stoichiometries and small unit cells). Unfortunately, this number is dwarfed by the total number of possibilities. Just for ternary intermetallics, there are 1391 structure-types known experimentally9 and there are ~500,000 possibilities of combining three metallic elements for each of these prototypes. Moreover, ternary structures can be rather complex: the most likely number of atoms in the unit cell turns out to be 14, but the majority of intermetallic ternary prototypes is considerably larger.9 Considering that a DFT calculation scales with the cube of the number of atoms in the unit cell, we are quickly led to conclude that an exhaustive search will be out of reach for the foreseeable future.

It is therefore easy to understand the need for methods to accelerate the exploration of new compounds. The search of such methods to predict material properties starting from atomic properties traces back well before the beginning of the machine-learning era;9–11 however, new ideas have recently flourished, thanks to the introduction of machine learning in materials science.12–15 In this context, machine learning has already proved to be a powerful method to

https://doi.org/10.1063/1.5020223
predict the properties of molecules or polymers\textsuperscript{16–19} as well as solids.\textsuperscript{13,20–28}

In this article, we employ machine learning\textsuperscript{29} to explore ternary intermetallic compounds. We use an approach similar to the one of Ref. 4, where it was shown how the use of machine learning techniques can accelerate the discovery of new stable perovskite phases by as much as 80%.

It turns out that the simplest ABC and ABC\textsubscript{2} compositions have been rather well studied by systematic high-throughput investigations.\textsuperscript{5} We decided therefore to concentrate on AB\textsubscript{2}C\textsubscript{2} compounds and look at the two most common intermetallic prototypes for this composition, namely, the tI\textsubscript{10}-CeAl\textsubscript{2}Ga\textsubscript{2} and the tP10-FeMo\textsubscript{2}B\textsubscript{2} structures.

The tI\textsubscript{10}-CeAl\textsubscript{2}Ga\textsubscript{2} prototype is an ordered derivation of the binary tI\textsubscript{10}-BaAl\textsubscript{4} type, where the two nonequivalent Al sites are replaced by Al or Ga. This is the 5th most common ternary intermetallic prototype, and, with 289 entries\textsuperscript{8} in the Pearson’s Crystal Data\textsuperscript{30} database (2.2% of all intermetallic compounds), the most common with AB\textsubscript{2}C\textsubscript{2} stoichiometry. It is composed of B\textsubscript{2}C\textsubscript{2} layers separated by A atoms (see the left panel of Fig. 1), usually including a rare-earth or a transition metal together with two main group elements. The crystal structure belongs to the I 4/m 2/m 2/m spacegroup (#139), with the A atom in the 2a Wyckoff position, the B atom in the 4d, and the C atom in the 4e. This family includes some interesting heavy electron systems such as ThCr\textsubscript{2}Si\textsubscript{2}, CeCu\textsubscript{2}Si\textsubscript{2}, YbCu\textsubscript{2}Si\textsubscript{2}, VRu\textsubscript{2}Si\textsubscript{2}, etc.\textsuperscript{31}

The tP10-FeMo\textsubscript{2}B\textsubscript{2}\textsuperscript{32} is derived from the binary U\textsubscript{3}Si\textsubscript{2} structure by replacing the two crystallographic nonequivalent U atoms by two different metals. This is the second most common ternary intermetallic prototype with stoichiometry AB\textsubscript{2}C\textsubscript{2} with 193 entries\textsuperscript{8} (1.5% of all intermetallic compounds). It is composed of BC channels with the A atoms forming chains inside the channels (see the right panel of Fig. 1). The crystal structure belongs to the P 4/m 21/b 2/m spacegroup (#127), with the A atom in the 2a Wyckoff position, the B atom in the 4g, and the C atom in the 4h. There exist a large family of intermetallics comprising lanthanides or actinides that crystallize in this structure,\textsuperscript{33} some of them with very interesting magnetic properties. In fact, the atomic arrangement corresponds to a so-called Shastry-Sutherland lattice,\textsuperscript{34} yielding magnetic frustration for some combinations of the effective inter-site exchange parameter.\textsuperscript{35}

II. EXPLORATION OF THE COMPOSITION SPACE

In order to proceed, we must now decide what chemical elements we allow to occupy the A, B, and C sites. The noble gases do not usually bind to form solids and can be safely eliminated. DFT calculations for solids containing the f-electron lanthanides and actinides are time-consuming and notoriously difficult to converge. We therefore decided not to include these in our study, with the exception of La. We should keep in mind that, due to the chemical similarity between the elements of these families,\textsuperscript{36} there is a large probability that stable compositions that include La will also be stable when La is replaced by another lanthanide or by an actinate.

This leads to a total of 64 elements that amount to 64 × 63 × 62 = 249 984 different combinations for each of the prototypes studied here. This number could be further reduced by, e.g., restricting our study to intermetallic compounds, by using prior knowledge of the likely ternary compositions that crystallize in the tI\textsubscript{10} and the tP10 crystal structures, or by eliminating the compositions with unusual oxidation states. We have decided, however, not to bias our search by using this information, allowing therefore for the discovery of unexpected materials.

The first step in our study is to obtain a training set for the machine learning algorithm. For each of the two prototypes, we create 20 000 crystal structures with random A, B, and C elements. We then optimize the lattice constant and calculate the total energy for each structure within ab initio DFT as implemented in the computer code vasp.\textsuperscript{37,38} The tI\textsubscript{10} structure contains 5 atoms in the unit cell, while the tP10 has 10, exactly twice the number of atoms of the cubic perovskites used in our previous study.\textsuperscript{4} The symmetry of the tI\textsubscript{10} and tP10 is also lower than the symmetry of the cubic perovskites. These two facts imply that each of our DFT calculations is 2–20 times slower; i.e., building the training set for the tP10 prototype uses more computational time than running all possible cubic perovskites.

The parameters of the DFT calculations were set to guarantee compatibility with the data available in the materials project\textsuperscript{4} and open quantum materials\textsuperscript{7} databases. We used the projector augmented wave (PAW)\textsuperscript{39} datasets of version 5.2 with a cutoff of 520 eV and Γ-centered k-point grids, as dense as required to ensure an accuracy of 2 meV/atom in the total energy. All forces were converged to better than 0.005 eV/Å. All calculations were performed with spin-polarization using the Perdew-Burke-Ernzerhof\textsuperscript{40} (PBE) exchange-correlation functional, with the exception of oxides and fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, and W, where an on-site Coulomb repulsive interaction U\textsuperscript{11} with a value of 3.32, 3.7, 5.3, 3.9, 4, 38, 6.2, 3.25, and 6.2 eV, respectively, was added to correct the d-states. We remark again that our computational setup (DFT code, level of theory, convergence criteria, PAW potentials, values of U) make it possible to combine directly our total energies with the values stored in the materials project database.

FIG. 1. Prototypes used in this work. Left: tI\textsubscript{10} crystal structure of CeAl\textsubscript{2}Ga\textsubscript{2} (cerium: yellow; aluminum: blue; gallium: light green). Right: tP10 crystal structure of FeMo\textsubscript{2}B\textsubscript{2} (iron: orange; molybdenum: pink; boron: dark green). The conventional cell is also depicted.
We define a compound thermodynamically stable if its distance to the convex hull of thermodynamic stability is zero. We recall that the convex hull is a hypersurface in composition space, which connects the formation energy of compounds that are energetically more favored than all possible decomposition channels. The positive distance to the convex hull of a phase is a measure of its instability, while a negative distance indicates that the compound is more stable than those that have been used to build the convex hull. Note that we are considering phase diagrams at zero temperature and zero pressure. The inclusion of temperature effects would in fact require the calculation of the phonon modes of the different phases of all involved compounds, to evaluate the phonon contribution to the free energy. Moreover, the contribution of configuration entropy, and therefore disorder, should also be included. Such calculations are at the moment unfeasible for such a large set of compounds.

To build a convex hull, one needs the total energy of all possible reservoir compounds. We chose to use the materials project database for our reference energies and determine the distances to the convex hull of stability with pymatgen. The materials project database includes most of the experimentally known inorganic crystals that are present in the Inorganic Crystal Structure Database (ICSD) database and an increasing number of theoretically predicted phases.

An overview of the training sets is shown in Fig. 2 where we plot a histogram of the distances to the convex hull of stability. As for the case of the perovskites, these histograms can be very well fitted by a skew normal distribution given by

\[
f(x) = \frac{2N}{\sigma \sqrt{2\pi}} e^{-x^2/2} \left[ 1 + \text{erf}(x/\sqrt{2} \alpha) \right],
\]

where \(N\) is the number of structures and \(x = (E - E_0)/\sigma\). For the tI10 prototype, we get the parameters \(E_0 = 0.227 \pm 0.004\) eV/atom, \(\sigma = 0.617 \pm 0.006\) eV/atom, and \(\alpha = 3.6 \pm 0.1\), while for the tP10 structure, we obtain \(E_0 = 0.139 \pm 0.003\) eV/atom, \(\sigma = 0.630 \pm 0.006\) eV/atom, and \(\alpha = 5.6 \pm 0.3\). It is clear from Fig. 2 and from the values of \(E_0\) that the tP10 curve is shifted toward lower energies, although the low-energy and high-energy tails look very similar for both prototypes. The width of both curves is nearly identical and much smaller than the width of the histogram for perovskites, where one finds \(\sigma = 1.206 \pm 0.009\) eV/atom. These data indicate that there are likely more stable ternary compounds with the tI10 or the tP10 crystal structure than cubic perovskites.

These data were then used to train a machine learning algorithm. Following the conclusions of Ref. 4, we adopted an algorithm based on extremely randomized trees, using the library scikit-learn. We decided to use the same features that were selected in Ref. 4 by estimating their importance using out of bag errors. The input feature vector is composed of 11 features per element: the atomic number, the Pauling electronegativity, the most common oxidation state, the average ionic radius, the number of valence electrons, the period in the periodic table, the group in the periodic table, the ionization energy, the polarizability, the number of \(x + p\) valence electrons, and the number of \(d\) or \(f\) valence electrons, resulting in a total of 33 features. This machine was then used to predict the energy of the 225 000 missing compositions. In order to reduce the influence of the randomization that happens during the training process, ten machines were trained and the average energy of each structure was used in the following.

If we simply proceeded from the recipe proposed in Ref. 4, we would then perform DFT calculations for all predicted stoichiometries beneath a given threshold of around 500–700 meV/atom. However, from Fig. 2, we see that this interval contains too large a fraction of the total number of data points. We therefore chose an alternative approach that tries to minimize the amount of DFT calculations, while ensuring an acceptably small number of false positives and negatives. We proceed as follows: (i) we recalculate using DFT all predicted structures that were beneath 200 meV/atom; (ii) add these new data points to the training set and train again the extremely randomized trees; (iii) use the machine to predict the missing data points; (iv) repeat steps (i–iii) three times. The idea behind this procedure is that adding low-energy structures to the training set increases the precision of the machine in this region, allowing us to use a smaller cutoff energy. It turns out that the number of stable systems that we found at each iteration decreased rather rapidly, giving us confidence that we found the large majority of stable systems. In total, including training and validation for both prototypes, we ran \(\approx 126\ 750\) systems using DFT out of a total of \(\approx 500\ 000\), leading to a reduction of 75% of the computational time.

The low-energy crystal structures that were calculated with DFT in step (i), together with some random structures that we had calculated but not included in the training set, were used to estimate the error of our original machine. This is not, strictly speaking, the statistical error associated with our learning machines, as we do not have here a proper test set, in contrast to the previous study on perovskites. The structures calculated in step (i) are in fact a random set, as they include systems that are selected because their distance to the convex hull of stability is predicted to be smaller than 200 meV. Keeping in mind the shortcomings of this evaluation of errors, we report as an indication of the quality of the prediction a mean absolute error of 105 meV/atom for tI10 and 92 meV/atom for tP10.
III. ANALYSIS OF THE STABLE SYSTEMS

We will define as thermodynamically stable systems that are within 5 meV/atom of the convex hull of stability. With this definition, we find 937 stable systems with the tI10 prototype and 1171 stable systems with the tP10 structure. Certainly, some of these are false positives due to our incomplete knowledge of the convex hull. There will also be a series of false positives and negatives due to the inevitable error connected to the use of the PBE(+U) approximation in our DFT calculations. However, and even taking these factors into account, these numbers (937 + 1171 = 2108 potentially stable systems) are very large and indicate that our knowledge of ternary systems is quite incomplete.

Let us start by comparing our fully unbiased predictions with known experimental data in order to have an estimation of the total of false negatives. For the tI10 structure, we can find 197 systems in the materials project database that respect our choice of chemical elements. From these, 36 have positive distances to the convex hull: these are either metastable systems that can be synthesized experimentally or just reflect the theoretical error inherent to the PBE approximation of DFT. From the remaining 161 systems, 146 were found in our simulations leading to a total of 9.3% false negatives. This number is relatively high but can be easily understood by looking at the missing systems. First they all contain an element from the second row (B, N, O, F), a magnetic 3d metal (Cr, Mn, Fe, Co, Ni), or Os. For the first two cases, we already know from Ref. 4 that we have a larger error in the machine predictions. The false negative predictions for compounds that contain Os are difficult to analyze, and their behavior cannot be easily related to known factors. Second, and as we will see later, there are relatively few tI10 stable systems with these elements, justifying the added difficulty for the machine. Note that the larger error associated with magnetic systems may be related to the restrictive choice of the spin configuration of the ground state. In fact, also in this case, we followed the recipe of the materials project and started our spin-polarized calculation with a large ferromagnetic moment on all atoms. Of course, the systems can still become antiferromagnetic (or ferrimagnetic) during the self-consistent cycle. Unfortunately, it is true that the ferromagnetic (or paramagnetic) configuration that comes out from our calculation may be a local minimum. Due to the extensive nature of our computational study, we could not include other starting spin configurations.

We now turn to the tP10 prototype. In the materials project database, we can find 27 entries for this composition, while a few more systems are listed in Ref. 33 for a grand total of 55. Our procedure was able to predict all these compounds with the exception of one, namely, AlCu3Re2. However, this compound should be discarded as it is highly unstable (0.541 eV above the convex hull). Incidentally, we realized that the structure for FeMo2B2 in the database has Mo and B sublattices exchanged.

To analyze trends across the periodic table, we counted the number of stable compositions ($E_{\text{hull}} < 5 \text{ meV/atom}$) containing a certain chemical element in the positions A, B, or C. The results for tI10 compositions can be found in Fig. 3 and those for tP10 compositions can be found in Fig. 4.
The B atom (Wyckoff 4e position) is more scattered across the periodic table, and can be often occupied by H, and a series of metals. Finally, the C atom (Wyckoff 4a position) is usually occupied by Ir, Pt, Au, or a group of elements surrounding As. Note, however, that there are many stable compositions not following these general trends. This also includes many non-intermetallic compounds containing, e.g., O or F in the B position or P or Se in the C position. We should also emphasize the large number of stable compounds including La in the Wyckoff 2a position. It is of course likely that La can be replaced by another lanthanide or by an actinide, which will greatly increase the number of stable tII0 materials.

The results for tP10 are shown in Fig. 4. For the A atom (2a Wyckoff position), we can find a large number of metals concentrated around Mg and In, while the B atom (4g Wyckoff position) is often a IIIB or a IVB metal. Finally, the C atom is most often Be, B, Si, Ga, Ge, or a metal close to Pd in the periodic table. Again we find that La can appear very often (this time in the B position).

As both the tII0 and the tP10 prototype structures are typical of intermetallic compounds, it is not surprising to find that a large majority of our stable phases turn out to be metallic. There are a few systems, however, that exhibit an electronic band gap. For the tII0 system, there are 30 semiconductors, amounting to a mere 3% of all stable compounds. All of these include at least one non-metallic element, namely, H, P, As, Sb, Bi, Se, or Te. The largest PBE gap can be found for ZnCs$_2$Se$_2$ and for ZnRb$_2$Se$_2$ and reads $\sim 2.2$ eV. Note, however, that the PBE approximation tends to underestimate considerably this electronic property.

The situation is very similar for the tP10 structure. There are only 19 semiconducting systems and a large majority of these include O. There are, however, a few fully intermetallic systems that present a small electronic band gap, such as MgBe$_2$Os$_2$. The largest PBE gap is $\sim 0.8$ eV for ON$_2$Cu$_2$.

Around 50 tII0 stable systems exhibit a ferromagnetic polarization, with the largest moment of 6.9 $\mu_B$ per unit cell found for AuH$_2$Mn$_2$. Of particular interest are the compounds that contain at the same time a lanthanide and a 3d-metal, as in the well-known SmCo$_5$ and Nd$_2$Fe$_14$B super magnets. With this kind of composition, we find, e.g., LaMn$_2$Si$_2$ (3.5 $\mu_B$ per unit cell), LaMn$_2$P$_2$ (3.4 $\mu_B$ per unit cell), and LaCo$_2$As$_2$ (1.7 $\mu_B$ per unit cell). Also around 50 tP10 systems exhibit a ferromagnetic moment, with the largest moment of 4.5 $\mu_B$ per unit cell found for CrF$_2$P$_2$. Here, however, we could not find any stable phase combining 3d and 4f metals. Interesting magnetic compounds will be better characterized in future work.

The stable crystal structures will be distributed through the Materials Project database, while the whole data set will be available in the NOMAD Repository.

IV. CONCLUSIONS

We presented a thorough investigation of the tII0-CeAl$_2$Ga$_2$ and tP10-FeMo$_2$B$_2$ intermetallic structure prototypes using a combination of machine learning and high-throughput techniques based on density functional theory. Our main finding is a total of 2108 thermodynamically stable systems, from which only 215 could be found in online databases. As expected for these two prototypes, a large majority of the stable systems are metallic and non-magnetic. We estimate a total fraction of false negatives of $\sim 9\%$ for the tII0 and $\sim 0\%$ for the tP10. A fraction of 9\% is acceptable, but it would be desirable to decrease this value further by improving the prediction of our machines. From the analysis of our results, we believe that this can be achieved by treating systems containing second-row elements and magnetic elements, for which the prediction error is considerably higher, separately from the other compounds. This could be achieved, for example, by biasing the initial training set to include a larger proportion of these elements. Work along these lines is already under way.

ACKNOWLEDGMENTS

M.A.L.M. acknowledges partial support from the German Research Foundation within the Collaborative Research Centre 762 (Project No. A11) and MA-6786/1. Computational resources were provided by the Leibniz Supercomputing Centre through the SuperMuc Project Nos. p1841a and pr48je.
