# Predictions of new ABO_{3} perovskite compounds by combining machine learning and density functional theory

Prasanna V. Balachandran, Antoine A. Emery, James E. Gubernatis, Turab Lookman, Chris Wolverton, and Alex Zunger

Phys. Rev. Materials **2**, 043802 — Published 11 April 2018

DOI: 10.1103/PhysRevMaterials.2.043802

# Predictions of New $ABO_3$ Perovskite Compounds by Combining Machine Learning and Density Functional Theory

P. V. Balachandran,[1] A. A. Emery,[2] J. E. Gubernatis,[1]

T. Lookman,[1,*] C. Wolverton,[2] and A. Zunger[3]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545 U.S.A.*

[2] *Department of Materials Science and Engineering,*

*Northwestern University, Evanston, IL 60208 U.S.A.*

[3] *Renewable and Sustainable Energy Institute,*

*University of Colorado, Boulder, CO 80309 U.S.A.*

(Dated: February 28, 2018)

## Abstract

We apply machine learning (ML) methods to a database of 390 experimentally reported $ABO_3$ compounds to construct two statistical models that predict possible new perovskite materials and possible new cubic perovskites. The first ML model classified the 390 compounds into 254 perovskites and 136 that are not perovskites with a 90% average cross-validation (CV) accuracy; the perovskites are further classified by a second ML model into 22 known cubic perovskites and 232 known non-cubic perovskites with a 94% average CV accuracy. We find that the most effective chemical descriptors affecting our classification include largely geometric constructs such as the A and B Shannon ionic radii, the tolerance and octahedral factors, the A-O and B-O bond length, and the A and B Villars' Mendeleev numbers. We then construct an additional list of 625 $ABO_3$ compounds assembled from charge conserving combinations of A and B atoms absent from our list of known compounds. Then, using the two ML models constructed on the known compounds we predict that 235 of the 625 exist in a perovskite structure with a confidence greater than 50% and among them that 20 exist in the cubic structure (albeit, the latter with only $\sim$50% confidence). We find that the new perovskites are most likely to occur when the A and B atoms are a lanthanide or actinide, when the A atom is an alkali, alkali earth, or late transition metal atom, or when the B atom is a *p*-block atom. We also compare the ML findings with the density functional theory calculations and convex hull analyses in the Open Quantum Materials Database (OQMD), which predicts the T=0 K ground state stability of all the $ABO_3$ compounds. We find that OQMD predicts 186 of 254 of the perovskites in the experimental database to be thermodynamically stable within 100 meV/atom of the convex hull and predicts 87 of the 235 ML-predicted perovskite compounds to be thermodynamically stable within 100 meV/atom of the convex hull, including 6 of these to be in cubic structures. We suggest these 87 as the most promising candidates for future experimental synthesis of novel perovskites.

## I. INTRODUCTION

The $ABO_3$ compounds, particularly those with a perovskite structure hold a special place in the arena of design and discovery of new materials with interesting and technology-enabling target functionalities[1–3] because of the range of elements involved, the variety of crystal structures possible, and the breadth of important physical properties they exhibit. The perovskite crystal structures can exist not only in the cubic "undistorted" structure, but also in a wide variety of structural distortions.[4] Common to all is the $ABO_3$ chemical formula where the A atom is 9- to 12-fold coordinated by oxygen, whereas the B atom is 6-fold coordinated by oxygen, and most importantly, the $BO_6$ octahedra are corner-connected in all three directions. Perovskite structures include space groups that are subclasses of cubic, orthorhombic, tetragonal, rhombohedral, monoclinic, and triclinic crystals. We adopt the definition of Zhang *et al.*[5] who limit these space groups to the 15 unique structures, noted by Lufaso and Woodward.[4] There are about 254 experimentally synthesized inorganic perovskites, of which 22 are cubic and 232 non-cubic (Tables I and II, Supplemental Material[6]). A number of interesting physical properties of $ABO_3$ compounds depend on whether they have a perovskite structure or not, and also if they have a perovskite structure, whether they are cubic perovskites or non-cubic perovskites. The perovskite structure, for example, is characteristic of high temperature superconductors,[7] colossal magneto-resistors,[8] and multiferroic materials.[9] Cubic perovskites are important as ionic conductors[10] and as a new class of topological insulators.[11,12] Accordingly, there is both an interest and need to design and discover new perovskites.

In this paper we apply two machine learning (ML) methods to (a) train a ML model to classify the $ABO_3$ compounds, which were assembled from the experimental literature as "perovskites or not," (b) train another ML model to classify experimentally known perovskites as "cubic or not" and then (c) use these models to predict whether members of a proposed set of $ABO_3$ compounds (not included in the training set) are perovskites. Finally (d), we compare our ML predictions for the proposed $ABO_3$ compounds with density functional theory (DFT) and a convex hull (CH) analysis of stability with respect to decomposition into other phases, using the Open Quantum Materials Database (OQMD).[13–15] Items (c) and (d) are the new findings presented in this paper where we report predictions of possible new formable and stable $ABO_3$ perovskite compounds. ML and DFT-based convex

hull (DFT-CH) methods provide complementary insights into whether an $ABO_3$ compound can exist in a perovskite structure.

We also consider a list of 625 $ABO_3$ compounds (given in Tables III and IV, Supplemental Material[6]) that are not present in the training set and assembled from charge conserving combinations of A and B atoms, and predict from ML that 235 compounds might be perovskites, including 20 in cubic structure, the latter albeit with only ∼50% confidence (only as good as random guessing). We predict from ML that new perovskites are most likely to occur when the A and B atoms are a lanthanide or actinide, or when the A atom is an alkali, alkali earth, or late transition metal atom, and when the B atom is a *p*-block atom. The DFT-CH description, as implemented in the OQMD, predicts 87 of the 235 ML-predicted perovskite compounds to be thermodynamically stable within 100 meV/atom of the CH, (which is our threshold for either decomposition or comparison to other $ABO_3$ structures with the same composition), including 6 of these to be in cubic structures. The predictions of a number of possible interlanthanides and interactinides are noteworthy as compounds in these classes are generally difficult to calculate accurately with *ab initio* methods but present no computational challenge for the ML methods.

The OQMD database was built from DFT calculations at zero temperature and pressure and can automatically construct the ground state CH[13–15] of a given A-B-O system, including all possible combinations of competing phases. An $ABO_3$ compound is stable in a given structure if it lies *on* the CH. However, metastable phases can often be synthesized, and hence we consider also a reasonable range of "degree of metastability" (DOM), whereby a structure predicted to lie near, but somewhat above the CH ($<$ 100 meV/atom) will also be a candidate for metastable formability (See Section III C).[16] While the question of formability can be experimentally validated by synthesizing the compounds in the laboratory, validation of stability is expected to be non-trivial due to the existence of potentially many energetically competing metastable structures and complex kinetic pathways before reaching the final stable state. Our hypothesis, supported by other analyses,[16] is that if a compound is predicted to be thermodynamically stable then it should be synthesizeable. If a compound has a small, but positive, DOM (0-100 meV/atom), then it could possibly be formable. We acknowledge that whether a given stoichiometry will form a compound or not will depend on synthesis conditions, which we cannot predict. Out of the 235 compounds predicted by ML to be in a perovskite structure, 87 are found to be OQMD stable or nearly stable in these

structures. We thus suggest that these 87 are most likely formable (see Appendix). Some of the notable $ABO_3$ perovskites that we predict by superimposing the ML predictions with the DFT-based CH analysis are $EuIrO_3$, $BiVO_3$, $EuVO_3$, $NdRuO_3$, and $EuMnO_3$. Validating our suggestions is a challenge we propose to experimentalists.

## II. BACKGROUND

### A. Classification of materials by ML

ML is a statistical approach which can be applied to the identification of classes of new materials based on existing knowledge of already formed members. For decades, researchers have used diagrammatic classification of the previously measured structure types of broad chemical groups of synthesized compounds in terms of phenomenological chemical coordinates describing the constituent elements (such as the atomic radii, quantum orbital radii, electronegativities, etc). The historical precursors to ML were "structure plots," i.e., scatter plots of two chemical coordinates of a material class having one or more targeted properties (generally, the crystal structure type). Straight lines were often drawn on the plot by hand to group materials with similar characteristics. Such plots often capture trends in materials behavior reflecting trends in the periodic table. Possible new materials are defined by the vacant regions of the plot. One looks for materials to occupy these regions in the vicinity of those whose properties one is seeking to enhance. Likely, the best known of the earliest structure plots are those of Mooser and Pearson, Phillips and Chelikowsky for binary AB octet compounds, Zunger for octet and non-octet structures and those of Villars and Pettifor for classifying ternary phases.[17–20]

As the databases of compounds subject to such classification queries has increased in size [for example, the growth of the Inorganic Crystal Structure Database (ICSD)][21,22] and the number of pertinent chemical coordinates diversified, ML methods have become a popular extension of the two-dimensional structure plot. Such methods have been applied to AB, $AB_2$ and $ABO_3$ materials with improved accuracy in predicted capabilities relative to the traditional structure plots[23–29] as well as to compounds with arbitrary stoichiometries.[30,31] Many of these ML applications start with materials known to exist and construct statistical models that predict expected properties of materials not yet known to exist. Since the

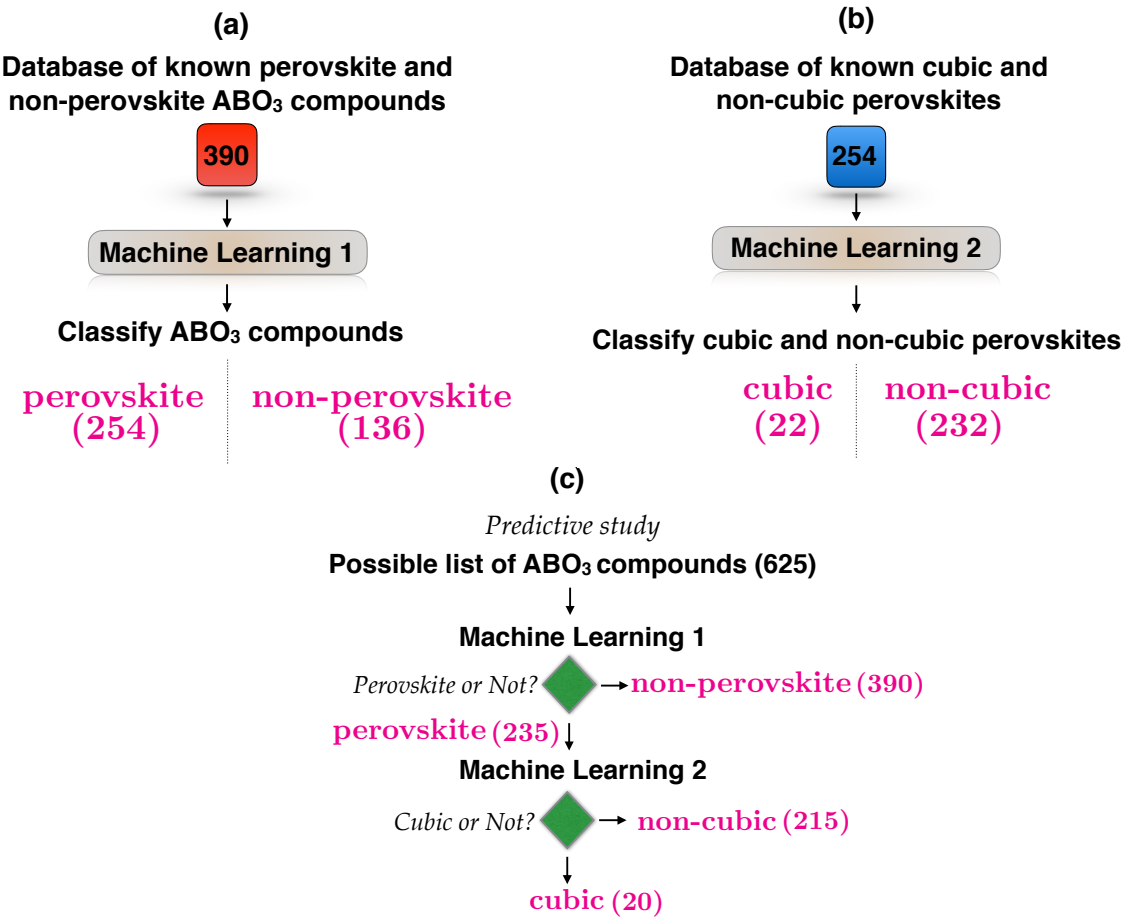


FIG. 1. Our ML workflow for the prediction of new $ABO_3$ cubic perovskites. We build two independent ML models for (a) the classification of $ABO_3$ perovskite or not (Machine Learning 1). We assembled a dataset of 390 $ABO_3$ compounds from surveying the literature, which included 254 perovskites and 136 non-perovskites. (b) Classification of cubic and non-cubic $ABO_3$ perovskites (Machine Learning 2). Out of the 254 perovskites, there were 232 in the non-cubic structures (e.g., orthorhombic, monoclinic, tetragonal, rhombohedral etc.) and 22 in the cubic structure (space group, $Pm\bar{3}m$). To predict whether a new $ABO_3$ compound will have a cubic perovskite structure or not, we utilize these two ML models in a hierarchical manner [as shown in (c)]. We assembled a list of 625 possible $ABO_3$ compounds that were not present in the training set. Machine Learning 1 predicted 235 as possible in the perovskite structure. Machine Learning 2 further predicted 215 $ABO_3$ compounds in the non-cubic perovskite structures and 20 in the cubic perovskite structure.

input to the learning process consists of previously synthesized compounds, the ensuing ML predictions provide insights into chemistries in a given stoichiometry that are probable in a crystal structure type without commenting on whether they are thermodynamically stable or not.

### B. Predicting new $ABO_3$ perovskites by ML

Our overarching ML strategy is shown in Figure 1. We first demonstrate that the ML models can classify the known 390 $ABO_3$ compounds into the 254 perovskites and 136 that are not perovskites with 90% average cross-validation (CV) accuracy determined by the stratified CV procedure (Figure 1a). This success in part requires identifying effective chemical descriptors (features) that enable such classification. Then, we build another set of ML models to classify all formable 254 perovskite structures into the 22 known cubic perovskites and 232 known non-cubic perovskites finding similarly 94% CV accuracy (Figure 1b). We interpret that the misclassification of our models are more a source of important physical information than they are failings of the model. For example, $KTaO_3$ and $SrVO_3$ were classified as non-cubic perovskite by ML whereas they were reported to be formed as cubic perovskite. They are likely poised to undergo a structural transition from the experimentally observed cubic to non-cubic perovskite. This possibility remains to be experimentally validated (discussed in subsection IV C).

We next apply the trained ML models constructed from these classifications of known compounds to predict whether they would be in cubic or non-cubic perovskite crystal structures for a list of 625 possible $ABO_3$ stoichiometries (Figure 1c), which are not in our current compilations of perovskite materials and are generated from charge conserving combinations of A and B cations. It is unknown to the ML models whether any of the 625 assembled stoichiometries is chemically stable at this stage because stability data was not included in the ML training. Stability will be assessed by DFT below. We formulate this prediction problem as a two-step task. In the first step, we use our trained ML models (Figure 1a) to screen for compounds and classify them into perovskites and those that are not perovskites. Only those $ABO_3$ compounds that are classified as perovskites reach the second step, where new ML models are trained (Figure 1b) then used to classify cubic from non-cubic perovskites. *Using this strategy, we predict a total of 235 $ABO_3$ compounds (not present in the*

*training set) in the perovskite structure, including 20 in the cubic perovskite structure.* We also identify chemical trends in the A-B element pairs that are predicted to be perovskites. They are: (i) When the A and B atoms are interlanthanides and interactinides, then the resulting $ABO_3$ compound is predicted to be a perovskite, *i.e.,* lanthanides and actinides appear to substitute for each other relatively easily. (ii) When the A atom is an alkali, alkaline earth, or late transition metal atom then there is a strong likelihood for the compound to be in a perovskite structure. (iii) Similarly, when the B atom is a *p*-block atom then it favors the perovskite structure. These specific trends are consistent with previous ML,[26] to some extent the HT-DFT studies of Emery *et al.*,[15] and with the recent analyses of known compounds of various structures and element sets.[32,33]

### C. Predicting stability of compounds by DFT convex hull construction

Our ML models can predict whether a particular $ABO_3$ compound will have a perovskite crystal structure or not and if it is a perovskite then whether it is cubic or not. However, the models lacks the thermodynamical stability insights, *i.e.,* they cannot indicate whether those compounds are stable or not, or whether it would readily decompose to other products. This is mainly because our training dataset lacks such detailed thermodynamic information. This question can be addressed by first principles based total energy calculations and CH analysis, which can indicate the degree to which a hypothetical compound is on or above the equilibrium CH. One can thus retain both stable compounds and those with a degree of metastability (DOM) energy not higher than a reasonable threshold value. One generally expects that compounds formed in the laboratory by near equilibrium growth methods (such as melt or solution growth, but not via artificial layer-by-layer growth methods from the gas phase such as molecular beam epitaxy, or Pulse Laser Deposition) are thermodynamically stable or weakly metastable. However, not all synthesizeable compounds are stable. Many metastable compounds that are known to form are protected from decomposition by insurmountable kinetic barriers.[34] Also, stable compounds may have thus far escaped successful synthesis (for example, because the right experimental conditions have not yet been found), and such combinations may also be used to (mis)train ML, whereas DFT will identify them as stable.

First principles calculations based on DFT provide a means to identify stable, or weakly

metastable materials. A particular implementation is based on "high-throughput calculations" in which a large series of DFT calculations are evaluated in an automatic high-throughput (HT) fashion. The total energy data for the compounds in the HT-DFT dataset can then be used to perform a CH analysis in which the lowest energy combination of phases can be identified for an arbitrary stoichiometry (say, $ABO_3$) and the structure type (or types) that leads to the lowest energy from a group of pre-selected competing phases is identified. The pre-selected competing phases include (i) alternative crystal structures of the target compound $ABO_3$ itself (*i.e.,* non-perovskite $ABO_3$) and (ii) decomposition products of the target compound (such as $A+B+3/2O_2$, or $AO+BO_2$). A target compound with an energy lower than those in the combined groups (i) and (ii) is declared a stable ground state structure. If its energy is higher, then it is declared unstable. If unstable, but close to the CH, the compound might still be potentially formable as a metastable phase with a reasonable level of DOM.

A number of open source implementations of the DFT-CH approach to predicting compound stability are available.[14,35,36] In the analysis that follows, we used the results of the predictions available from the OQMD. Instead of referring to these results as DFT-CH, we refer to them as OQMD to be specific about the source of the DFT-CH predictions. Different open sources databases can give different predictions depending on the details of the DFT and the construction of the CH.

## III. INPUTS AND METHODS

### A. Database of known $ABO_3$ compounds

Our database of $ABO_3$ compounds consists of 390 compounds and was created via an augmentation of the database of 354 $ABO_3$ compounds explored earlier by Pilania *et al.*.[26] This data included that compiled by Zhang *et al.*[5] who gathered their data from a number of resources, including the Inorganic Crystal Structure Database (ICSD) and published literature data. We added to our earlier 354 compounds 36 new $ABO_3$ compounds taken from the literature[37] and those compiled by Emery *et al.*[15] We note that in all 390 compounds the sum of the valences of A and B adds to 6 so these are "charge balanced compounds." For example, each A-B pair has nominal I-V, II-IV or III-III valences. No A-B pairs in

this set have IV-II or V-I valences. Each previously documented $ABO_3$ compound has a label signifying whether it is a perovskite or not. Compounds with the $ABO_3$ formula that satisfy the definition of "perovskite" (given in the Introduction) are labeled "Perovskite". Compounds with the $ABO_3$ formula that do not satisfy the definition, for example, ones whose octahedra are not corner-sharing and ones that decompose into constituents without forming any $ABO_3$ compound (failed synthesis[37]), are labelled "Non-perovskites." This definition is consistent with that of Zhang *et al.*[5] Of the 390 compounds, 254 are perovskites and 136 are not perovskites (Table I, Supplemental Material). We changed the label for $BaRuO_3$ from non-perovskite to perovskite (but not cubic), following the work of Jin *et al.*[38] Each perovskite is also labeled signifying whether it is cubic or not. 22 perovskites are cubic and 232 are non-cubic perovskites.

The structures for most materials were experimentally determined at ambient conditions; however, some were determined under other non-ambient conditions. Those designated as a perovskite are single-phase materials and could be either thermodynamically stable or metastable. Those designated as non-perovskite could be a single phase or mixed phase material that is thermodynamically stable or metastable. Our database lacks this more descriptive information.

For each compound, we include values of various chemical coordinates (features) associated with their A and B atoms or the chemical bonds. Initially, our database had 30 features. Most were classical chemical constructs obtained independently of structure classification such as atomic radii, orbital radii, and electronegativities. Other chemical scales, the Mendeleev numbers, were selected explicitly for structural classification without attributing intrinsic, independent chemical significance to these numbers. Our preprocessing however showed that many features in the database were not as influential in classifying the compounds as Perovskite or Not or as Cubic or Not as 4 sets of feature pairs previously used in structure plots.[5,39–41] These pairs are: The Shannon ionic radii[42] of the A and B atoms, the tolerance and octahedral factors, the bond valence theory estimates of the A-O and B-O bond lengths,[43] and the Villars' Mendeleev numbers.[44] With the exception of the Mendeleev numbers, these feature pairs provide a geometric characterization of a hard-core ionic sphere model of the crystal structure. The values of these features for each compound in our database are given in Table I of the Supplemental Material.

The Shannon radii are well-used estimates of an element's ionic hard-sphere radii ex-

tracted from experiment. With them, we computed the octahedral and Goldschmidt tolerance factors. These dimensionless numbers are commonly used metrics in studies involving perovskite or non-perovskite structures of materials. They measure ionic misfit of the B atom and the deviation of the structure from an ideal cubic geometry. The octahedral factor for an $ABO_3$ solid is

$$\mathcal{O} = r_{\mathrm{B}}/r_{\mathrm{O}},$$

where $r_{\mathrm{B}}$ and $r_{\mathrm{O}}$ are the Shannon radii for the B atom and oxygen. We used $r_{\mathrm{O}} = 1.4$Å. The tolerance factor is defined as:

$$t = \frac{r_{\mathrm{A}} + r_{\mathrm{O}}}{\sqrt{2}(r_{\mathrm{B}} + r_{\mathrm{O}})}.$$

From empirical studies of stable $ABO_3$ solids,[5,39–41] it is well known that hexagonal structures are favored if $t > 1$, cubic, if $0.9 < t < 1.0$, and orthorhombic, if $0.75 < t < 0.9$. If $t < 0.75$, the compound is generally not a perovskite. If $t = 1$, the material is perfectly cubic. Six-fold co-ordination seems to require $0.414 < \mathcal{O} < 0.732$. $\mathcal{O} = 0.435$ corresponds to the arrangement where hard sphere B and O ions are touching in a close-packed arrangement. Empirical studies have also correlated crystal structures with ranges of $r_{\mathrm{A}}$ and $r_{\mathrm{B}}$ values: Generally, we must have $r_{\mathrm{A}} > r_{\mathrm{B}}$. The A atoms are in a 12-fold coordinated site if $r_{\mathrm{A}} > 0.9$Å and in a 6-fold coordinated (octahedral) site if $r_{\mathrm{A}} < 0.8$Å, as long as $r_{\mathrm{B}} < 0.7$Å.

In bond valence theory, a valence

$$V_{\mathrm{i}} = \sum_{\mathrm{i}} \nu_{\mathrm{ij}}$$

is assigned to an ion (cation or anion) as the sum of valences

$$\nu_{\mathrm{ij}} = \exp\left(d_0 - d_{\mathrm{ij}}\right)/b$$

associated with its chemical bonds with neighboring ions of opposite charges. $d_{\mathrm{ij}}$ is the bond length, and $d_0$ and $b$ are parameters fit to experimental data. $d_0$ depends on the cation-anion pair. $b$ has a nearly universal value of 1.4Å. If $\nu_{\mathrm{ij}}$ is taken to be the nominal valence of atom $i$ divided by the number of its nearest neighbors, then the $d_{\mathrm{ij}}$ bond length is easily computed from (4). The Zhang *et al.* database has two features, the bond valence theory predictions, which accounts for a bond length increasing or decreasing to accommodate the changes in

the valence of ion pairs due to charge transfer among the ions, of the A-O and B-O bond lengths $d_{\mathrm{AO}}$ and $d_{\mathrm{BO}}$.

Whereas the best known Mendeleev numbers are due to Pettifor[20] (replacing the atomic numbers by numbers determined by their ability to fit the observed crystal structure in a structure map), we use here the Mendeleev numbers of Villars *et al.*.[44] These numbers sequence the elements in structurally similar groups. Villars labels the elements sequentially in their columns always from top to bottom. The *s*-block elements are in {1,10}, Sc=11, Y=12, the *f*-block elements (lanthanides and actinides) are in the interval {13,42}, the *d*-block elements are in {43,66}, and the *p*-block elements are in {67,10}. The lanthanides and actinides are regarded as 10 columns as opposed to two rows. The message from Pettifor, Villars, and others who proposed permutations of the atomic numbers of the elements to accentuate grouping of materials is that chemical trends in atomic co-ordination and crystal structure are best seen by grouping elements by column as opposed to grouping by rows.

The structure plots for each of the 4 feature pairs are given in Fig. 2. A black dot marks a perovskite; a red dot, a non-perovskite. These plots clearly illustrate the difficulty one would have drawing a single few-sided polygon to separate cleanly the perovskites from those that are not. These plots illustrate the challenges confronting ML.

### B. Technical details on ML Classifiers

Classification is a form of supervised learning, meaning the predictions of one variable (the class label) is based on the values of the other variables (the features). We used the random forest (RFC)[45] and gradient tree boosting (GTBC)[46] classifiers, as implemented in the open-source software package SCIKIT learn,[47] to build our ML models. These classifiers are ensemble methods, meaning they use a combination of many models, each trained on the data, to produce the final model. At the core, RFC and GTBC use a decision tree classifier. A random forest is a simple average of the sum of many classifiers, each a decision tree fit to a bootstrap sample of the training data. Gradient tree boosting is a weighted sum of many decision trees of shallow depth, which makes the model a weak classifier of the data. The individual weak classifiers are built recursively such that in constructing the new classifier from the current one, the data is reweighted. Thus, what was misclassified by the present classifier is weighted more heavily in the construction of the new classifier.

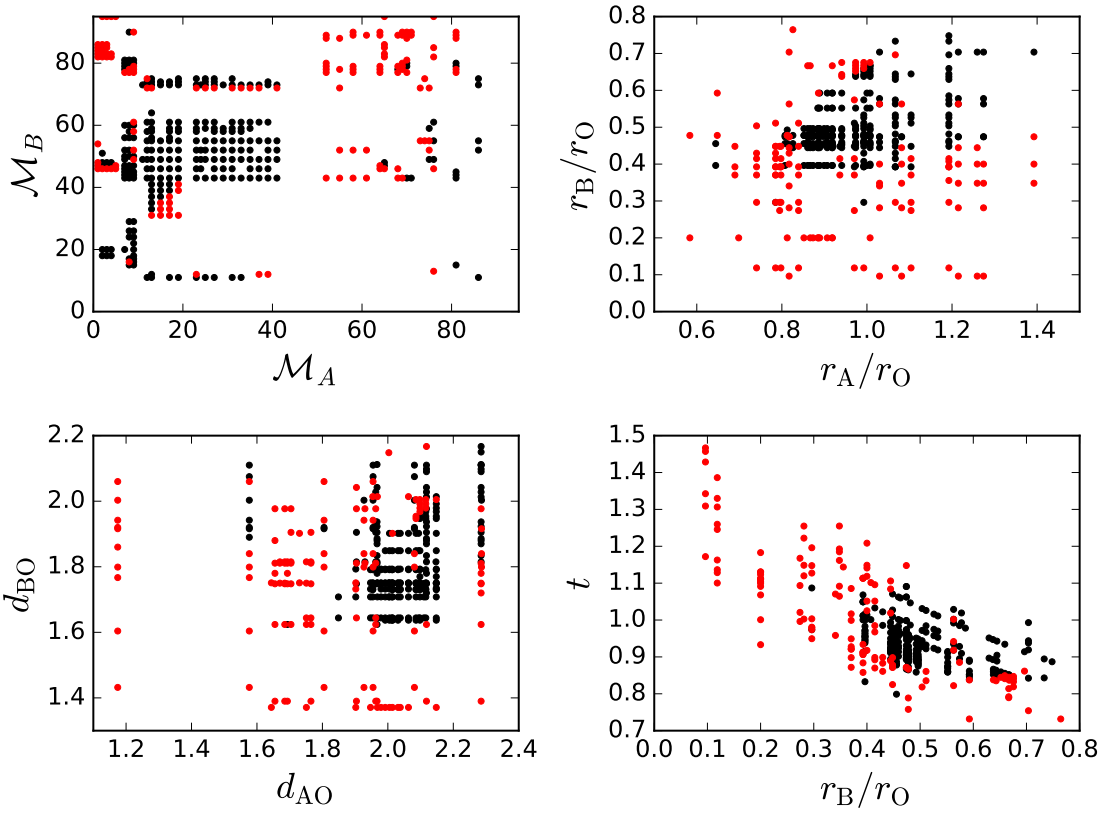


FIG. 2. The perovskite (black) or not (red) structure plots of the known $ABO_3$ compounds for the four features pairs we adopted for our analysis. The top left is the structure plot when the Villars' Mendeleev numbers are used; top right, the Shannon radii for the A and B atoms (divided by Shannon's ionic radii for oxygen); bottom left, for the bond valence theory A-O and B-O bond lengths; and bottom right, the tolerance and octahedral factors.

To set the hyperparameters in the models we used a stratified shuffle split cross-validation scheme where we created our training and test data sets on the basis of a 50/50 split, formed randomly but in such manner that the percentage of perovskites was the same in each split as it was in the entire data set, and for the cubic or not case, the percentage of cubics was the same in each split as it was in the perovskite subset of the database. In this scheme we create the model and perform its testing on sub-databases having similar populations. While a 50/50 split means building the model with a smaller database, we found the adopted scheme useful for the cubic or not case. The cubics are only 10% of the perovskite data. A more conventional 75/25 or 90/10 split has large fluctuations in the number of cubics in the

training and test sets and subsequently larger variances in the predictions. The 50/50 spilt seemed to help control the means and variances of the test set predictions.

The relevant hyperparameters for random forests were the number of bootstrap samples, which we set at 200, and the maximum tree depth, which we set at 6 for the perovskite or not case and at 4 for the cubic or not case. For gradient tree boosting, we set the subsampling of the training data at 50%, the number of ensembles at 2500 for the perovskite or not case and 2000 for the cubic or not case, and the learning rate at 0.001. In the RFC case, deep tree depths resulted in a significant overfitting of the training data, often approaching 100% accuracy but with a large variance. We simply decreased the depth, observing the accuracy of the predictions on the test data increasing and the variance decreasing. When the mean accuracy started to decrease, we stopped. We adjusted the hyperparameters for GTBC similarly but set the maximum tree depth of its trees to 3 to make the classifier weak. We made these adjustments for the octahedral and tolerance factor case, whose model gave the initial highest accuracy and hence had the greatest likelihood being overfit, and applied them to the classifiers for the other feature pair cases.

In Table I, we give the mean and standard deviations of the different model predictions for the test data. We computed these on the basis of 100 runs for each classifier (model) built. The histograms of the predicted accuracy were reasonably symmetric with the median of the predictions nearly equal to the mean. By accuracy of the predictions, we mean the number of times the model predicted correctly the entries in the training or the test data divided by the number of data in the given sub-dataset.

### C. Technical details on DFT calculations of stability of different $ABO_3$ compounds

The OQMD[13,14] calculations were performed using the Vienna *ab initio* simulation package (VASP)[48,49] using projector-augmented wave method potentials (PAW)[50] and the Perdew-Burke-Ernzerhof (PBE)[51] generalized gradient approximation to the exchange-correlation functional. DFT+$U$[52,53] was used for some elements (V, Cr, Mn, Fe, Co, Ni, Cu, Th, U, Np, Pu, see Table II) and calculations containing $3d$ transition elements (Sc-Cu) or actinides are spin-polarized with ferromagnetic alignment of spins.

For all calculations, Γ-centred $k$-point meshes are used. The electronic self-consistency (for a given set of ion positions) is converged to within $10^{-4}$ eV/atom. Any calculation

TABLE I. The mean accuracy and the standard deviation (STD) of the predictions of our RFC and GTBC models on the test data. These quantities were computed on the basis of 100 repetitions of the machine learning fits. We give these statistical quantities for each pair of features used in this paper.

| | Perovskite or Not | | | | Cubic or Not | | | |
|---|---|---|---|---|---|---|---|---|
| | RFC | | GTBC | | RFC | | GTBC | |
| Feature Pair | Mean | STD | Mean | STD | Mean | STD | Mean | STD |
| $\mathcal{M}_{\rm A}, \mathcal{M}_{\rm B}$ | 0.860 | 0.028 | 0.841 | 0.024 | 0.913 | 0.016 | 0.914 | 0.016 |
| $d_{\rm AO}, d_{\rm BO}$ | 0.859 | 0.022 | 0.848 | 0.024 | 0.918 | 0.016 | 0.928 | 0.016 |
| $r_{\rm A}/r_{\rm O}, r_{\rm B}/r_{\rm O}$ | 0.903 | 0.017 | 0.899 | 0.020 | 0.933 | 0.015 | 0.937 | 0.016 |
| $\mathcal{O}, t$ | 0.898 | 0.017 | 0.900 | 0.018 | 0.933 | 0.016 | 0.933 | 0.015 |

containing $d$-block or actinide elements are spin polarized with a ferromagnetic alignment of spins to capture possible magnetism, with initial magnetic moments of 5 and 7 $\mu_B$ for the $d$-block and actinide elements, respectively. It should be noted that this approach will not capture more complex magnetic ordering, such as antiferromagnetism. For several $3d$- and $f$-block elements, the GGA+$U$ approach is implemented to improve the exchange and correlation description of the localized charge density when these elements are in compounds with oxygen.

All calculations were completed in a two-step scheme. First, the structures were fully relaxed, followed by a static calculation. The relaxation calculations are performed at a plane-wave basis-set energy cutoff at the energy recommended in the VASP potentials of the elements in the structure, and 6,000 $k$-points per reciprocal atom. The quasi-Newton scheme is used to optimize the structure to within $10^{-3}$ eV/atom. The final static calculation of the structure is performed at an energy cutoff of 520 eV using tetrahedral $k$-point integration. The 520 eV cutoff is chosen because it is 25% higher than the highest recommended energy cutoff over all of the potentials used. This constant cutoff for all calculations ensures that all the energies calculated in OQMD are compatible, and can be used to evaluate the formation energies of compounds and T=0 K ground-state phase diagrams. More details on the OQMD DFT framework can be found in Kirklin *et al.*[14]

TABLE II. $U$-values for 11 elements used in the high-throughput density functional theory calculations.

| Element | $U$-value (eV) |
|---|---|
| V | 3.1 |
| Cr | 3.5 |
| Mn | 3.8 |
| Fe | 4.0 |
| Co | 3.3 |
| Ni | 6.4 |
| Cu | 4.0 |
| Th | 4.0 |
| U | 4.0 |
| Np | 4.0 |
| Pu | 4.0 |

The OQMD currently contains over 470,000 DFT calculations consisting of ∼40,000 experimentally observed compounds from the ICSD[21,22] and ∼430,000 hypothetical structures. Among those hypothetical structures, the OQMD contains 5329 $ABO_3$ cubic perovskite and 2162 rhombohedral, tetragonal and orthorhombic perovskites that were calculated in a previous HT-DFT study.[15,54] Those 3 distortions, in addition to the cubic phase, are the most common perovskite structures found in the ICSD and in the literature. From OQMD we extract two quantities, ΔH (in eV/atom) and ΔE (in meV/atom), which refers to formation enthalpy and distance from the CH, respectively.

Thus, OQMD contains DFT calculations of both experimentally observed compounds from ICSD plus those for hypothetical compounds that are not found in the ICSD. To construct the convex hull for a given A-B-O element set, it uses the total energy data computed by DFT of all compounds in the dataset-both existing and hypothetical. As a result, if we have an $ABO_3$ compound which is already in the OQMD (because it is in the ICSD), then the distance from the convex hull will be either a positive value or zero. On the other hand, if we have a new, hypothetical ABO3 compound (not in the OQMD), then

its distance from the convex hull will be evaluated with respect to the stable compounds in the OQMD. In this case, the distance from the convex hull can take either a positive (the hypothetical $ABO_3$ compound is metastable or unstable) or negative value (the hypothetical $ABO_3$ compound is more stable than other compounds that lie on the convex hull). The routines for extracting these data are available through a web interface on www.oqmd.org or through the `qmpy` python package (https://github.com/wolverton-research-group/qmpy).

We note an alternate way to define the convex hull distance is just the energy of the compound minus the energy of the convex hull for all compounds in question, *including* the proposed compound. Under this definition the convex hull for a stable compound is zero, and there can be no negative value of the convex hull distance. The definition we adopted gives more information about whether a compound is "barely" on the convex hull or whether it breaks through the hull significantly.

Hypothetical prototypes are generated by a combinatorial analysis of possible combinations of atoms satisfying certain constraints with respect to valence, compatibility with certain crystal structures, etc. Both the OQMD and ICSD databases are constantly evolving as new compounds are added to ICSD and new prototypes are added to OQMD. Out of the 390 compounds in our database, the stability of 387 were computed in OQMD, and out of the 625 compounds in our list of possibilities, 598 were computed in OQMD.

### D. Phonon Calculations

We also performed spot DFT calculations to provide additional information to address the discrepancy in the predictions of new cubic perovskites between ML and OQMD. To accomplish this, we could have compared the total energy of $ABO_3$ in cubic structure to that of the other 14 perovskite structures. Another approach is to examine if at T=0 K an assumed structure has dynamically unstable phonons. If it does, then another structure will need to replace the cubic structure at some temperature if the compound is to be stable. We note that these calculations are not part of the data stored in the OQMD database.

For these phonon calculations, we used the plane wave pseudopotential code, `Quantum ESPRESSO` (QE).[55] A plane-wave cutoff of 60 Ry was used during the ionic and electronic relaxation steps. We explored two flavors of the Generalized Gradient Approximation (GGA), namely PBE and PBE for solids (PBEsol), to calculate the total energies. Within the PBE[56]

and PBEsol functionals,[57] we used the Projector Augmented-Wave (PAW)[58] and Ultra-soft method[59] for generating the pseudopotentials, respectively.[60] We used PBE PAW with QE for establishing a direct comparison with OQMD, which also uses PBE PAW functionals but with VASP. We also performed calculations using PBE Ultra-soft functional with QE and found negligible difference in the results (lattice constant of the cubic $ABO_3$ perovskites and the phonon spectra) between PBE Ultra-soft and PBE PAW functionals with QE. In addition, we explored the PBEsol Ultra-soft functional because it provides an improved description of the crystal structure for solids.[61–70] Having learned that PAW and Ultra-soft produce comparable results with negligible differences, we chose the PBEsol Ultra-soft for our calculations. Our spot calculations involving vanadium (e.g., $BaVO_3$) were performed within the DFT+$U$ formalism[71] with a ferromagnetic spin-order imposed on the V-atom. An effective Hubbard-$U$ of 2 eV was chosen. In calculations involving Rhenium (Re) atom, we explored DFT, DFT+$U$ ($U$=1.5 eV), spin-polarized (ferromagnetic spin order) and non spin-polarized calculations. All Re-containing calculations converged to a non-magnetic ground state and therefore, we only report the results from non-spin-polarized DFT calculations. The DFT optimized lattice constants for the $ABO_3$ compounds in the $Pm\bar{3}m$ cubic structure is given in Table V in the Supplemental Material.[6]

To determine the dynamical stability, we performed frozen phonon calculations using `PHONOPY` code[72] that uses the DFT forces from QE as input for calculating the dynamical matrices and interatomic force constants. We employed a supercell of size 2×2×2 with 40 atoms for the frozen phonon calculations. Our DFT phonon calculations were performed by assuming a ferromagnetic spin-order imposed on the V atom.

## IV. RESULTS

Using the two classifiers (random forest and gradient tree boosting), built from the four different feature pairs (Mendeleev numbers of the A and B atoms, the bond valence theory A-O and B-O bond lengths, the Shannon ionic radii of the A and B atoms, and the tolerance and octahedral factors), and applying them to the case of perovskite or not and cubic perovskite or not yields 16 sets of results. In Figs. 1 through 8 of the Supplemental Material,[6] we give results for all 16 sets. Here we present two sets of representative results, both obtained by the gradient tree boosting method.

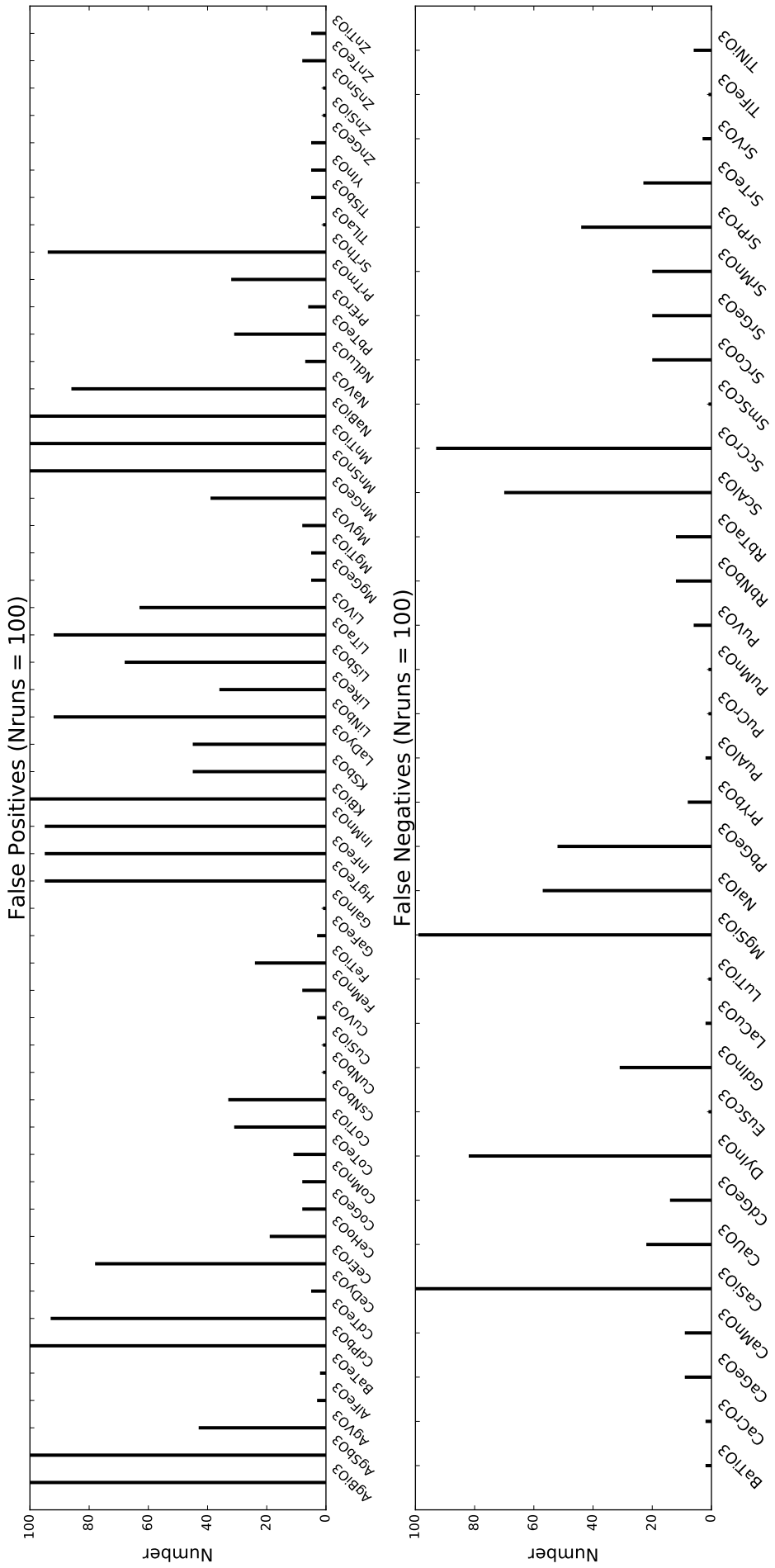


FIG. 3. For the octahedral and tolerance factor feature pair case, plots of the number of times a $ABO_3$ compound was classified false positive or false negative as a perovskite or not for 100 trials at classification of the entire database. Compounds from the database not listed in either figure were always classified correctly.

Because of their length, we provide comprehensive tables of the ML classification and the OQMD DFT-CH description of the 390 known $ABO_3$ perovskites, cubic perovskites and non-perovskites as well as the 625 corresponding cases for compounds not included in the learning set in the Supplementary Materials (Tables II, III, and IV). All these Tables represent the state-of-the art understanding of the capabilities and possible shortcomings of the two leading approaches to predictive theories of structures – ML based on learning from experiment and *ab initio* DFT-CH T=0 K thermodynamics.

### A. ML classification of experimentally synthesized Perovskite vs Non-perovskites

We first consider the ML classification of the known formed $ABO_3$ compounds into a perovskite or non-perovskite. In Fig. 3, we present for octahedral and tolerance factor feature pair the compounds that GTBC misclassifies and for each the number of times the misclassification was a false positive or false negative. For the false positives, the solid was classified as a perovskite but is listed in the data as not being a perovskite. For the false negatives, the solid was classified as not being a perovskite but is listed in the data as being a perovskite. Compounds not listed on these figures were always classified correctly as a perovskite or non-perovskite. A total of 100 classification attempts were made, each with a different random stratified 50/50 spilt into training and test data. Because of the stochastic nature of the analysis, some of the $ABO_3$ compounds are generally weakly misclassified as a consequence of statistical fluctuations producing an outlier. Ones that are strongly misclassified point to a possible mislabeling of the data or a material that had or is about to have a structural transition if the temperature or pressure is varied from ambient conditions. Alternatively, it could also be a ML error due to either insufficient representative samples in the training set or the lack of meaningful features in representing that specific chemical space. Within statistical fluctuations the misclassification plots for the other feature pairs are very similar to Fig. 3. In general, the same false positives and false negatives reoccur. The frequency at which a particular misclassification occurs is what principally changes.

The number of compounds in the training set that are classified as non-perovskites by ML is 125 (out of which 118 are in agreement with the experimentally determined label). The $ABO_3$ compounds that are classified as non-perovskites by ML, but that are experimentally known to be perovskites are the I-V valence compound $NaIO_3$, the II-IV compounds $CaSiO_3$,

$MgSiO_3$, $PbGeO_3$, and the III-III compounds $ScAlO_3$, $ScCrO_3$ and $DyInO_3$. On the other hand, compounds classified by ML as perovskites that are actually non-perovskites, are the I-V members $AgBiO_3$, $AgSbO_3$, $LiNbO_3$, $LiSbO_3$, $LiTaO_3$, $LiVO_3$, $NaBiO_3$, $NaVO_3$, $KBiO_3$, the II-IV members $CdPbO_3$, $CdTeO_3$, $HgTeO_3$, $MnSnO_3$, $MnTiO_3$, and the III-III members $SrThO_3$, $CeErO_3$, $InFeO_3$, and $InMnO_3$.

### B. ML predictions Perovskite vs Non-perovskites of new compounds

Next, we use our ML models created from the data of known $ABO_3$ compounds, to predict possible new perovskite compounds. For the experimentally observed compounds, 41 elements occur as A atoms and 54 as B atoms. Using these elements, we created 625 possible combinations of $ABO_3$ compounds consistent with the requirements that the sum of the valences of the A and B atoms was 6 and the pair was not already in our database. By imposing the constraint that the sum of the valences equal 6, we are omitting cases that are potentially stable with other valences. However, our initial data set of 390 compounds does follow this constraint with no exception. For several A atoms, such as Fe, Eu, and Tl,

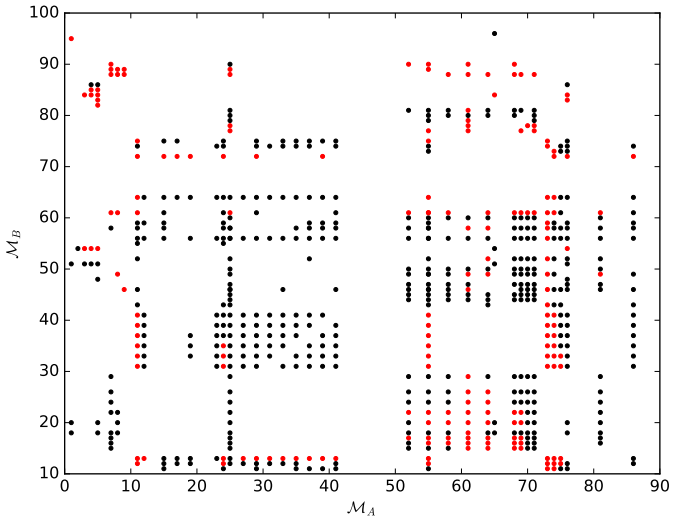


FIG. 4. For octahedral and tolerance factor feature pair case, the structure plot for the predictions plotted as a function of the Mendeleev number of the A and B atoms. Black dots mark predicted perovskites; red dots, predicted non-perovskites.

we admitted multiple valence states of +2/+3, +2/+3, and +1/+2, meaning these atoms occurred as A atoms with two different sets of B atoms. This combinatorial exercise yielded our list of possible $ABO_3$ solids not present in our training set. We next generated the feature set for each proposed compound. These solids had no label assignment. Predicting their labels is the task of the ML model constructed from the labeled data.

Out of 625 possible $ABO_3$ compounds, ML classifies 235 as perovskite and 390 as non-perovskites. Figure 4 is the predicted structure plot for the new compounds for each octahedral and tolerance factor feature pair. Although predicted for a model constructed for this feature pair, we chose to present the result as a function of the Mendeleev numbers ($\mathcal{M}$) for easier chemical identification and to promote the grouping of chemically similar solids. Apart from statistical fluctuations the predictions for the other three feature pairs are quite similar and are given in the Supplemental Material. The octahedral and tolerance factor feature pair model predicts more perovskites for $\mathcal{M}$ in the interval $\{51, 71\}$ than other models that are built from other feature pairs.

Figure 4 places possible perovskites into vacant spaces near other perovskites in the Mendeleev structure plot of Fig. 2. Very generally, perovskites exist or are predicted to exist for A atoms being an $s$-block or lanthanide atom. Some additional perovskites are predicted when the A atom is from the $d$-block and the B atom is from the $d$- or $f$-blocks. These generalizations are consistent with our past ML analysis[26] and the substitution probability analysis[32,33] as well as some of the prior HT-DFT[15] results. With respect to Fig. 2, for ease of convenient reference, $\mathcal{M}$ = 25, 55, 76, 81, and 86 are Eu, Fe, Tl, Pb, and Bi. With the exception of Fe, the other 4 are predicted to form a perovskite with a variety of B atoms spread across the periodic table.

Our ML methods also estimate empirically the probabilities of the predictions. Figure 5 shows these estimated probabilities for the octahedral and tolerance factor pair case. The classifiers label their predictions as a perovskite if the probability is 0.5 or greater (black dots) and a non-perovskite if the probability is less than 0.5 (red dots). From the plot, one sees most of the predictions have a probability well below or well above 0.5. The others are basically "coin flip" cases. The octahedral and tolerance factor feature pair model is more optimistic in some of its predictions than the other three models which tend to be more consistent with each other.

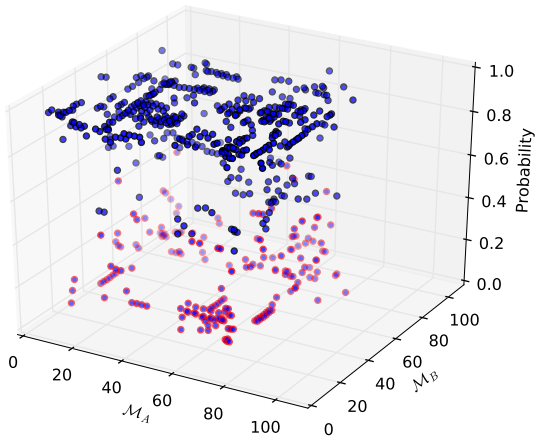


FIG. 5. For the octahedral ratio and tolerance factor feature pair case, the probabilities for the predictions in Fig. 4 plotted as a function of the Mendeleev numbers. Black dots marked materials predicted to be a perovskite with probability greater than 0.5; red dots for those predicted to be a perovskite with probability less than 0.5.

### C. ML classification of Cubic vs Non-Cubic perovskites of known compounds

The cubic or not case is a more difficult ML problem than the perovskite or not case because only 10% of the perovskite data are cubic and hence there are just a few from which to learn. Furthermore, cubic phases are often high-T phases, and some are classified as cubic only because the synthesis was performed at high-T following by quenching to low-T. Figure 6 shows the structure plots for the known data as function of our four sets of feature pairs. In each case a black dot marks a cubic perovskite; a red dot, a non-cubic perovskite.

In Fig. 7, we present for the Shannon radii of the A and B atoms feature pair the $ABO_3$ solids that are misclassified by the GTBC method. In addition, we also show for each misclassified $ABO_3$ compound, the number of times the misclassification was as a false positive or false negative. For the false positives, the compound was classified as cubic but is listed in the data as not being cubic, and vice versa for the false negatives. Compounds not listed on these figures were always classified correctly as a cubic perovskite or not. As in the perovskite or not case, the models built with the different feature pairs showed similar

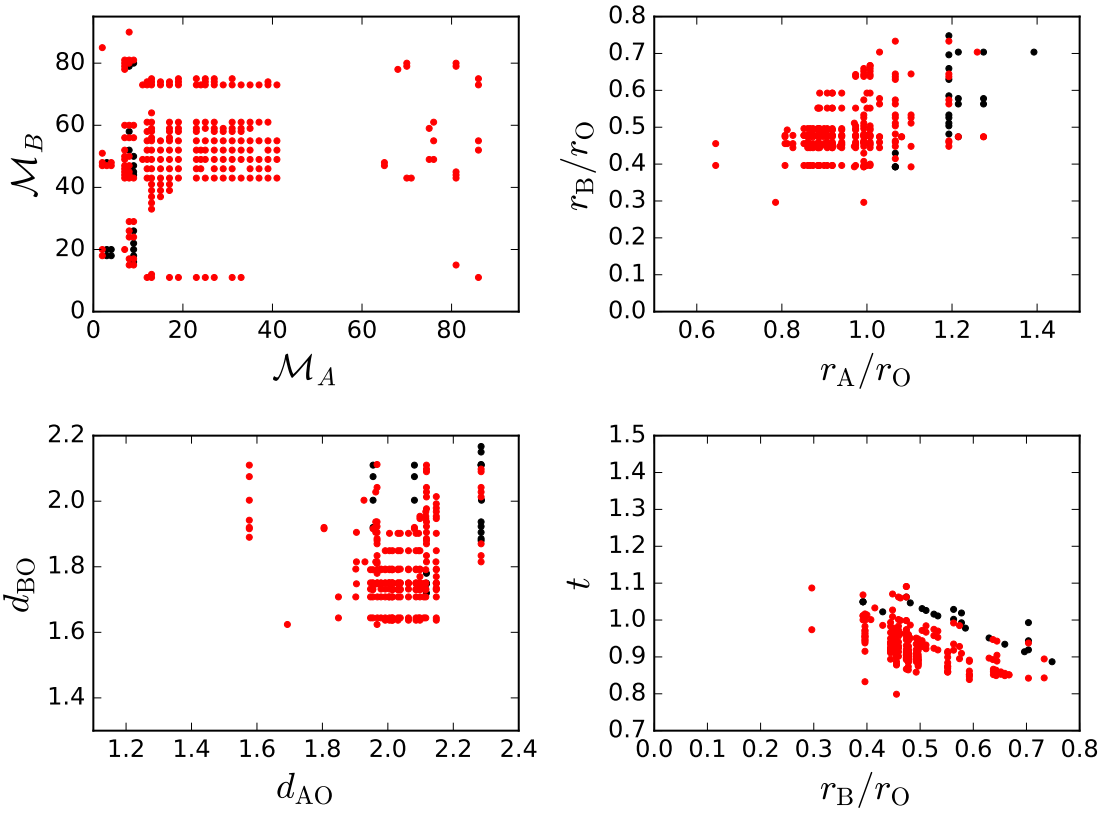


FIG. 6. The cubic perovskite (black) or non-cubic perovskite (red) structure plots of the known perovskite compounds for the four features pairs we adopted for our analysis. The top left is the structure plot when the Villars' Mendeleev numbers are used; top right, the Shannon radii for the A and B atoms (divided by Shannon's ionic radii for oxygen); bottom left, for the bond valence theory A-O and B-O bond lengths; and bottom right, the tolerance and octahedral factors.

misclassification.

What distinguishes this case from the perovskite or not case is the misclassification tends to have mainly the A atom being Sr or Ba ($\mathcal{M} = 8$ or 9) and with less frequently being K or Rb ($\mathcal{M} = 3$ or 4). This points to the fact that the feature pairs used for the classifiers not capturing a trend that would enable a more accurate classification. Other features, such as Born charge and Villars elemental property parameters,[73] were explored without any significant improvement in the predictions.

The number of compounds in the training set that are classified as non-cubic perovskites by ML is 246, out of which 225 are correct. There are 232 non-cubic perovskites in the

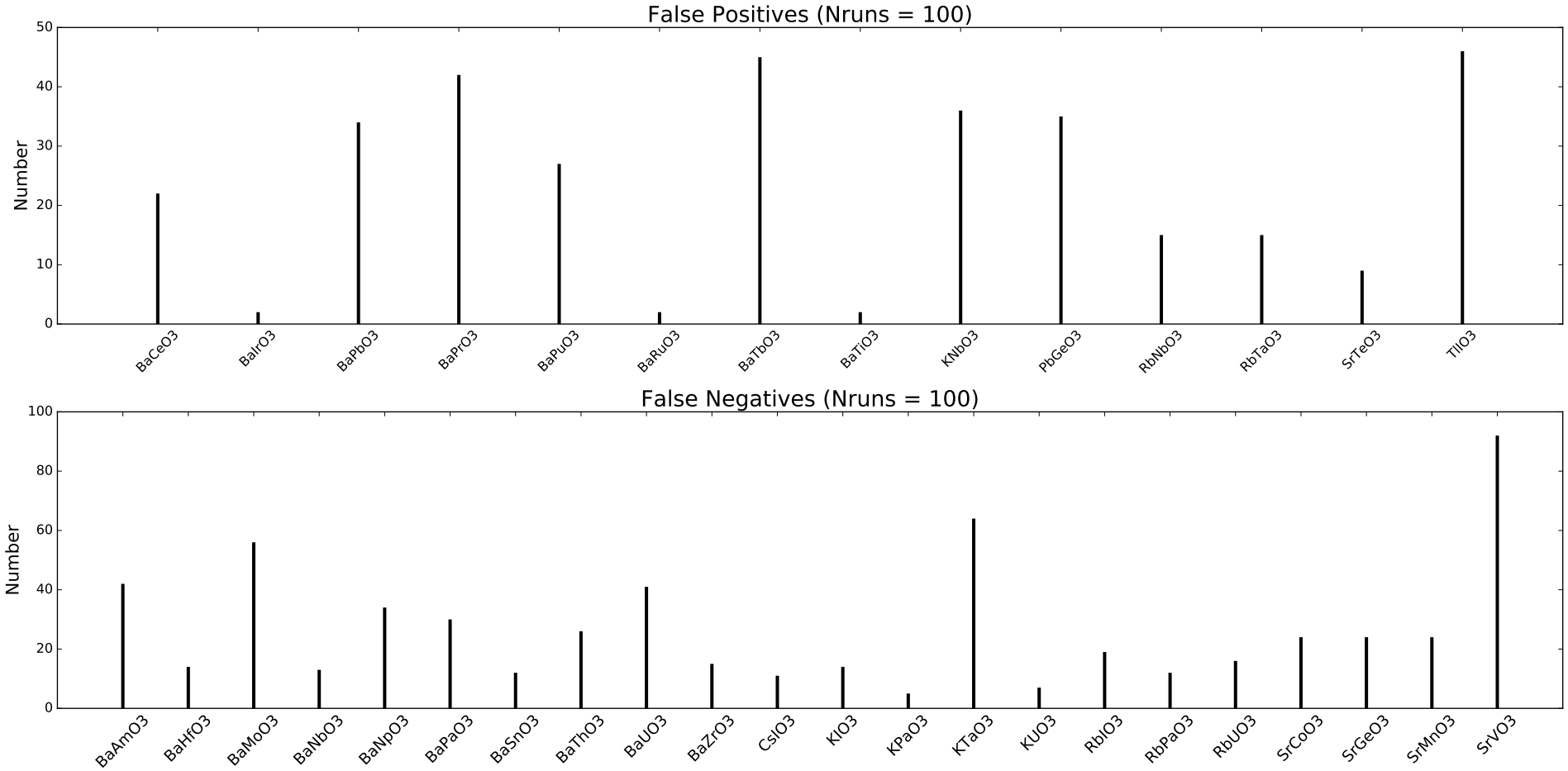


FIG. 7. For the Shannon radii feature pair case, plots of the number of times a perovskite compound was classified false positive or false negative as a cubic or not for 100 trials at classification of the entire database. Compounds from the database not listed in either figure were always classified correctly.

database. The number of compounds classified by ML as cubic perovskites is 19 (out of which 19 are correct). There are 22 cubic perovskites in the database. There were no notable $ABO_3$ compounds that are classified as cubic perovskites by ML, but are actually determined as non-cubic perovskites or non-perovskites from experiments. Notable compounds classified as non-cubic perovskites by ML that are actually cubic perovskites include only three cases: $KTaO_3$, $BaMoO_3$, and $SrVO_3$. We note that the cubic phase is stable at high temperatures and often transforms at low temperature into other non-cubic phases. So, it is possible that the above misclassified compounds pertain to phases that may transform at lower temperature to non-cubic. For example, consider $KTaO_3$. Feng *et al.*[40] note $KTaO_3$ as formable in cubic perovskite structure. Phonon calculations from density functional theory (DFT) also find no phonon instability in the bulk cubic $KTaO_3$ structure at T= 0 K.[74] Therefore, we assign a cubic label to $KTaO_3$ in our dataset of 254 formable perovskites. After training our ML models using this dataset, we find that ML "misclassifies" $KTaO_3$ as non-cubic with about 65% confidence. Experimentally, $KTaO_3$ has been shown as an incipient ferroelectric with anomalous dielectric behavior,[75] indicating that it is poised to undergo a ferroelectric phase transition below a critical temperature, so the ML classification as non-cubic could pertain to a low-T phase. Intriguingly, HT-DFT data in OQMD predicts $KTaO_3$ as a cubic perovskite. Similarly, our trained ML models also classify $SrVO_3$ as a non-cubic perovskite with greater than 90% confidence. The HT-DFT data in OQMD also predicts $SrVO_3$ as stable in the orthorhombic ($Pnma$) perovskite structure. But, the observation based on high temperature synthesis conditions (1000° C) under reduction atmosphere is cubic.[76] We recommend low temperature X-ray diffraction studies to find if this compound is non-cubic at low-T and resolve the discrepancy. In Table III we summarize the training performance of the ML (for the $t$ and O features), where the sum of the entries across a row equals the total number of non-cubic perovskites, non-perovskites, and cubic perovskites in the experimental dataset. Similarly, sum of the entries down a column equals the total number of non-cubic perovskites, non-perovskites, and cubic perovskites as classified by ML after training. Diagonal entries are the number of cases where ML exactly captures the experimental data. For example, down the ML: non-cubic column ML classifies 225 compounds in agreement with the experimental data, and only 18 of its non-cubic predictions are experimentally non-perovskites and 3 are cubic perovskites in the data. The experimental data had a total of 232 non-cubic perovskites.

TABLE III. Comparison of the classifications of experimentally synthesized $ABO_3$ compounds (referred to as "DATA" in the table) with those predicted from ML.predicted from the OQMD database. Out of the 390 compounds in our database, the stability of 387 were computed in OQMD. The word "cubics" refers to a perovskite in the cubic structure ($Pm\bar{3}m$). Similarly, "non-cubics" refers to a perovskite in a structure other than cubic. Finally, "non-perovskites" refers to all other cases.

| | ML: cubics | ML: non-cubics | ML: non-perovskites | |
|---|---|---|---|---|
| DATA: cubics | 19 | 3 | 0 | 22 |
| DATA: non-cubics | 0 | 225 | 7 | 232 |
| DATA: non-perovskites | 0 | 18 | 118 | 136 |
| | 19 | 246 | 125 | |

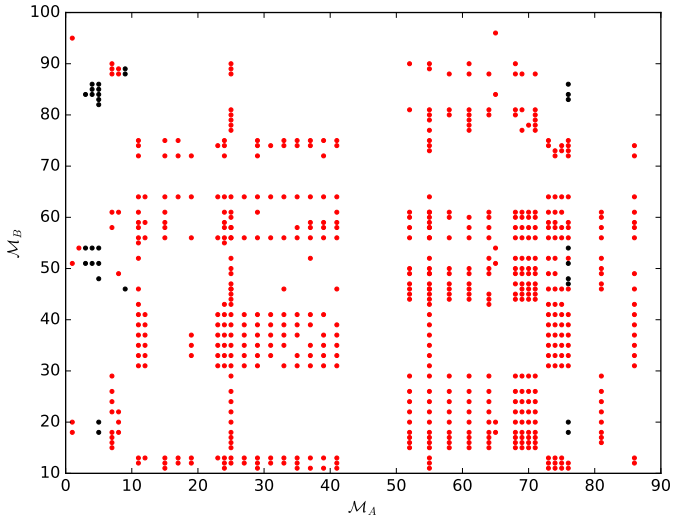


FIG. 8. For the Shannon radii feature pair case, the structure plot for the cubic perovskite predictions plotted as a function of the Mendeleev number of the A and B atoms. Black dots mark predicted cubic perovskites; red dots, predicted non-cubic perovskites.

### D. ML predictions of new cubic perovskites

Recall that of 625 possibilities, 235 are predicted to be in the perovskite structure by ML. Our ML models here used the Shannon's ionic radii as features and predicts a total of

20 new cubic perovskites. They are: $BaVO_3$, $CsBiO_3$, $CsPaO_3$, $CsReO_3$, $CsSbO_3$, $CsTaO_3$, $CsUO_3$, $CsWO_3$, $KReO_3$, $KWO_3$, $RbBiO_3$, $RbReO_3$, $RbSbO_3$, $RbWO_3$, $TlBiO_3$, $TlNbO_3$, $TlPaO_3$, $TlReO_3$, $TlTaO_3$, and $TlUO_3$. Most of the new cubics predicted by ML have the A atom being an alkali atom (K, Rb, Cs). About a third are $TlXO_3$ solids with element X sprinkled across the periodic table. Figure 8 is the structure plot for the predicted new cubic perovskites. Although predicted for models constructed for the selected feature pair variables, we again chose to present the results as a function of the Mendeleev numbers for easier chemical identification and better similar solid grouping. The model for the Shannon radii is the most optimistic one as it predicts about 5 times the number of cubics as the model for the Mendeleev number feature pair and twice the number as the model for the bond length pair (Fig. 7 of the Supplemental Material). The model for the octahedral and tolerance factor feature pair predicts no new cubic perovskites. Figure 9 shows the ML estimated probabilities for the cubic or not predictions. About half the predicted cubics are in the coin-flip range (around 50% confidence); however, the probabilities of the cubics predicted by the models constructed from the other feature pairs (Fig. 8 of the Supplemental Material) are almost all coin-flips and generally involve an alkali A atom; that is, the other feature pairs predict few if any cubic zinc-based perovskites. Because of this disagreement among the predictions of the different feature pair models, we believe the probabilities of our predicted new cubics should be regarded as coin-flips.

Thus, there is a striking contrast whereby ML predicts 20 and OQMD predicts 6 cubic perovskite compounds, which are also included in the 20 as predicted by ML. The 6 compounds include $CsPaO_3$, $CsUO_3$, $KReO_3$, $KWO_3$, $TlPaO_3$, and $TlUO_3$. We checked the dynamical stability of the remaining 14 compounds predicted by ML to be in cubic perovskite structure by using phonon calculations. The results are given in Table IV. We find that 5 to 6 compounds (depending on the pseudopotential and functional used) have imaginary frequencies at one or more high-symmetry points in the irreducible Brillouin zones indicating that they are dynamically unstable as a cubic (in disagreement with ML that says they should be cubic) whereas 7 are locally dynamically stable as cubic (in agreement with ML, which predicts them as cubic). However, even if they are dynamically stable, the DFT-CH calculations in OQMD predict that they are statically unstable, *i.e.,* in the OQMD database, these 14 compounds are either predicted as non-perovskite or they are predicted to decompose (data given in Table III in the Supplemental Material).

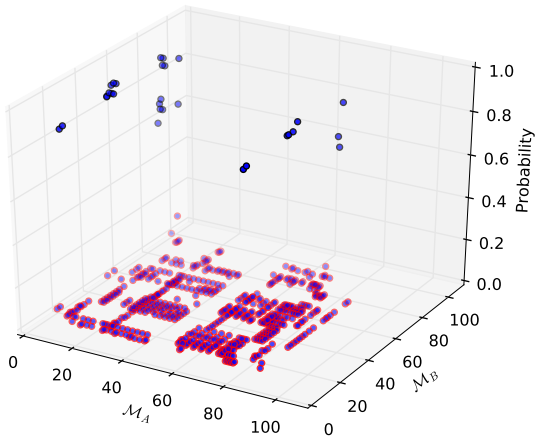


FIG. 9. For the Shannon radii feature pair case, the probabilities for the predictions in Fig. 8 plotted as a function of the Mendeleev numbers. Black dots marked materials predicted to be a cubic perovskite with probability greater than 0.5; red dots for those predicted to be a cubic perovskite with a probability less than 0.5.

### E. Comparison of ML and OQMD classifications and predictions

The stability predictions of OQMD offer a complementary means to evaluate the ML prediction of new materials. We first compare OQMD predictions of stability in the non-cubic perovskites, cubic perovskites and non-perovskites with the classifications of the experimental data for the synthesized compounds. A summary of these comparisons is given in Table V (Tables II, III and IV in the Supplemental Material give the full details). In Table V of the manuscript, the sum of the entries across a row equals the total number of non-cubic perovskites, non-perovskites, and cubic perovskites in the experimental dataset. The sum of the entries down a column equals the total number of non-cubic perovskites, non-perovskites, and cubic perovskites predicted stable by OQMD. Diagonal entries are the number of cases where DFT T=0 K stability agrees with experimental data. The sum of the diagonal entries divided by the total number of entries is the estimated fraction of experimentally synthesized phases that are thermodynamically stable. For example, down the OQMD: non-cubic column OQMD predicts 163 stable, non-cubics in agreement with the experimental data,

TABLE IV. Examination of the cubic versus not cubic ML predictions by DFT phonon calculations for 14 $ABO_3$ compounds that are classified as cubic perovskite by ML. We explored two flavors of GGA (PBEsol and PBE) and two different pseudopotentials (ultra-soft and PAW). "Not cubic" indicates that the phonon calculations found the corresponding $ABO_3$ compound to be dynamically unstable in the cubic perovskite crystal structure. "Cubic" means that the phonons of the cubic phase at $T = 0$ are normal.

| Compound | Phonon calculations (QE) | |
|---|---|---|
| | PBEsol ultra-soft | PBE PAW |
| $BaVO_3$ | Cubic | Not cubic |
| $CsBiO_3$ | Cubic | Cubic |
| $CsReO_3$ | Cubic | Cubic |
| $CsSbO_3$ | Cubic | Cubic |
| $CsTaO_3$ | Not cubic | Not cubic |
| $CsWO_3$ | Cubic | Cubic |
| $RbBiO_3$ | Not cubic | Cubic |
| $RbReO_3$ | Cubic | Cubic |
| $RbSbO_3$ | Not cubic | Cubic |
| $RbWO_3$ | Cubic | Cubic |
| $TlBiO_3$ | Not cubic | Not cubic |
| $TlNbO_3$ | Not cubic | Not cubic |
| $TlReO_3$ | Cubic | Cubic |
| $TlTaO_3$ | Not cubic | Not cubic |

and only 14 of its non-cubic predictions are experimentally non-perovskites and 6 are cubic perovskites in the data. The experimental data had a total of 232 non-cubic perovskites.

Interesting cases comparing OQMD with experiments include,

(a) OQMD predicts non-cubic perovskite as stable whereas synthesized as non-perovskite: These include $NaPO_3$, $BaCO_3$, $HgSeO_3$, $MgCO_3$, $PbCO_3$, and $PbSO_3$, where the B atom is an element that is either too small to occupy the octahedral site or acting as a cation whereas it belongs to normally anionic species. This could reflect either an insufficient number of non-perovskite structural candidates, experimental error or the compounds synthesized at

TABLE V. Comparison of the classifications of experimentally synthesized $ABO_3$ compounds (referred to as "DATA" in the table) with the stability predicted from the OQMD database. Out of the 390 compounds in our database, the stability of 387 were computed in OQMD. The word "cubics" refers to a perovskite in the cubic structure ($Pm\bar{3}m$). Similarly, "non-cubics" refers to a perovskite in a structure other than cubic. Finally, "non-perovskites" refers to all other cases.

| | OQMD: cubics | OQMD: non-cubics | OQMD: non-perovskites | |
|---|---|---|---|---|
| DATA: cubics | 10 | 6 | 5 | 21 |
| DATA: non-cubics | 7 | 163 | 61 | 231 |
| DATA: non-perovskites | 0 | 14 | 121 | 135 |
| | 17 | 183 | 187 | |

non-ambient conditions (*i.e.,* high-T or high-P).

(b) OQMD finds non-perovskite as stable but non-cubic perovskite synthesized: Examples for the I-V cases include $KNbO_3$, $RbNbO_3$, $RbTaO_3$, and $AgNbO_3$, the II-IV cases include $CaIrO_3$, $CaMnO_3$, $CaPbO_3$, $CaSiO_3$, $BaIrO_3$, $BaRuO_3$, $BaTiO_3$, $MgSiO_3$, $PbGeO_3$, $PbZrO_3$, $SrMnO_3$, and $HgTiO_3$, the III-III cases include $BiAlO_3$, $BiFeO_3$, $DyMnO_3$, and $YGaO_3$. These cases reflect examples of experimentally synthesized metastable perovskite phases rather than a failure of the DFT calculations.

(c) OQMD predicts cubic perovskite as stable and non-cubic perovskite or non-perovskite synthesized: Examples are $EuAlO_3$, $EuNiO_3$, $SrMoO_3$, $SrTiO_3$, and $YbAlO_3$. These could correspond to non-cubic structural distortions missing from the OQMD database.

(d) OQMD predicts non-cubic perovskite as stable and cubic perovskite synthesized: These cases are $NaWO_3$, $BaNpO_3$, $BaPaO_3$, $BaThO_3$, and $BaUO_3$. This could reflect the fact that OQMD determines the ground state T=0 K stability, whereas cubic perovskite is often a high temperature phase.

(e) OQMD perovskite compounds that are predicted solidly unstable (>100 meV above the CH), that is, predicted to decompose to other phases, yet they are experimentally formable: Notable cases include compounds with a single actinide or lanthanide element in the A or B positions: $CaUO_3$, $CeYbO_3$, $PbCeO_3$, $SmYO_3$, $TmYO_3$, $YbYO_3$, or inter lanthanides $NdDyO_3$, $NdYbO_3$, $PrDyO_3$, $PrHoO_3$ and $PrYbO_3$. In addition, a few other examples include $InCrO_3$, $InRhO_3$, $CoSiO_3$, and $CoTeO_3$. These cases might reflect either

TABLE VI. Comparison of the classifications of possible new $ABO_3$ compounds as predicted by ML and from the OQMD database. Out of the 625 compounds in our list of possibilities, 598 were computed in OQMD. predicted from the OQMD database. Out of the 390 compounds in our database, the stability of 387 were computed in OQMD. The word "cubics" refers to a perovskite in the cubic structure ($Pm\bar{3}m$). Similarly, "non-cubics" refers to a perovskite in a structure other than cubic. Finally, "non-perovskites" refers to all other cases.

| | OQMD: cubics | OQMD: non-cubics | OQMD: non-perovskites | |
|---|---|---|---|---|
| ML: cubics | 6 | 0 | 14 | 11 |
| ML: non-cubics. | 3 | 78 | 129 | 210 |
| ML: non-perovskites | 2 | 22 | 344 | 368 |
| | 11 | 100 | 487 | |

intrinsic DFT errors in exchange-correlation functional, or CH errors (insufficient number of trial $ABO_3$ structures especially for the rare cases of lanthanides), or that some metastable compounds can form despite having an energy much above the stability limit.

From Table V, we see that in many cases the OQMD predicts synthesized perovskites to be metastable (*i.e.,* stable as non-perovskite). For instance, the 61 compositions, which are experimentally observed to form perovskites (albeit non-cubic), are predicted by OQMD to be stable but in different structures, *i.e.,* (non-cubic) non-perovskites. The DFT calculations are performed at T=0 K, where a lower-energy non-perovskite structure can exist, whereas at the T>0 K synthesis temperature a perovskite phase can be stabilized relative to a non-perovskite phase. More details on to the stability of each compound are given in Table III in the Supplementary Material.[6] On the other hand, if OQMD predicts a compound to be stable in a perovskite crystal structure (cubic or non-cubic), then it is very likely it can also be experimentally synthesized. This is consistent with our hypothesis that stable compounds can often be synthesized.

Table VI compares the predictions of ML with OQMD for our list of possible new compounds. Out of 235 compounds that are predicted by ML as perovskite, 87 are also OQMD stable. Examples of OQMD predicted unstable compounds, which ML predicts as perovskites include compounds where the B position is a rare earth (actinide or lanthanide).

In the Appendix we list all 87 compounds predicted to be stable by DFT and as perovskite by ML. These compounds we regard as the most likely candidates for experimental synthesis of new perovskites. Those with $\Delta E \leq 0$ are more likely to be stable than those with $\Delta E > 0$.

Out of 390 new compounds that are predicted by ML as non-perovskites, only 24 are OQMD stable (DOM=100 meV/atom) in one of the perovskite structures. The agreements are $CuMoO_3$, $CuPaO_3$, $CuUO_3$, $ErCuO_3$, $EuNpO_3$, $EuThO_3$, $HoCuO_3$, $LiPaO_3$, $LiUO_3$, $LiWO_3$, $LuCuO_3$, $MnPaO_3$, $PbCrO_3$, $PbMnO_3$, $PuFeO_3$, $PuNiO_3$, $PuScO_3$, $TlAlO_3$, $TlCuO_3$, $TmCuO_3$, $YCuO_3$, $ZnPaO_3$, $Eu^{II}Ni^{IV}O_3$ and $EuSiO_3$. Amongst these 24 compounds, 11 have an actinide element (Pa, U, Np, Th or Pu) in either A or B site of the perovskite lattice. Even with respect to known materials the tendency of OQMD is to predict the compound to be other than a perovskite.

We also note that from OQMD we obtain a more detailed description of its classifications and predictions than implied by Tables V and VI. OQMD predicts whether the compound is in a stable perovskite structure and if so, whether that structure is cubic, is in a stable structure but one that is not a perovskite structure, or decomposes into some mixed phase of other compounds or possibly a single phase with a chemistry other than $ABO_3$ (Table III and IV, Supplemental Material). In Figure 10, we summarize the key outcomes from both ML and OQMD in terms of their relative performances with respect to the experimental data (training set for ML) and those that were not included in training the ML models.

### F. Validation of ML and OQMD predictions for recently synthesized $ABO_3$ perovskites

We now directly evaluate the predictive capabilities of ML and OQMD stability using $BaVO_3$, $PbMoO_3$, $KWO_3$, and $CaCoO_3$ compounds. We note that $BaVO_3$, $PbMoO_3$, $KWO_3$, and $CaCoO_3$ were not part of our ML training dataset. Nishimura *et al.*[77] recently experimentally synthesized pure $BaVO_3$ perovskite in the cubic structure using high-pressure synthesis conditions. This experimental synthesis is in agreement with our ML predictions. Further, our phonon calculations using PBEsol predict a locally stable cubic phase for $BaVO_3$ (Table IV). OQMD, on the other hand, gives a $\Delta E$ of +105 meV/atom for $BaVO_3$ in the cubic ($Pm\bar{3}m$) perovskite structure, but it predicts the ground state of $BaVO_3$ as a non-

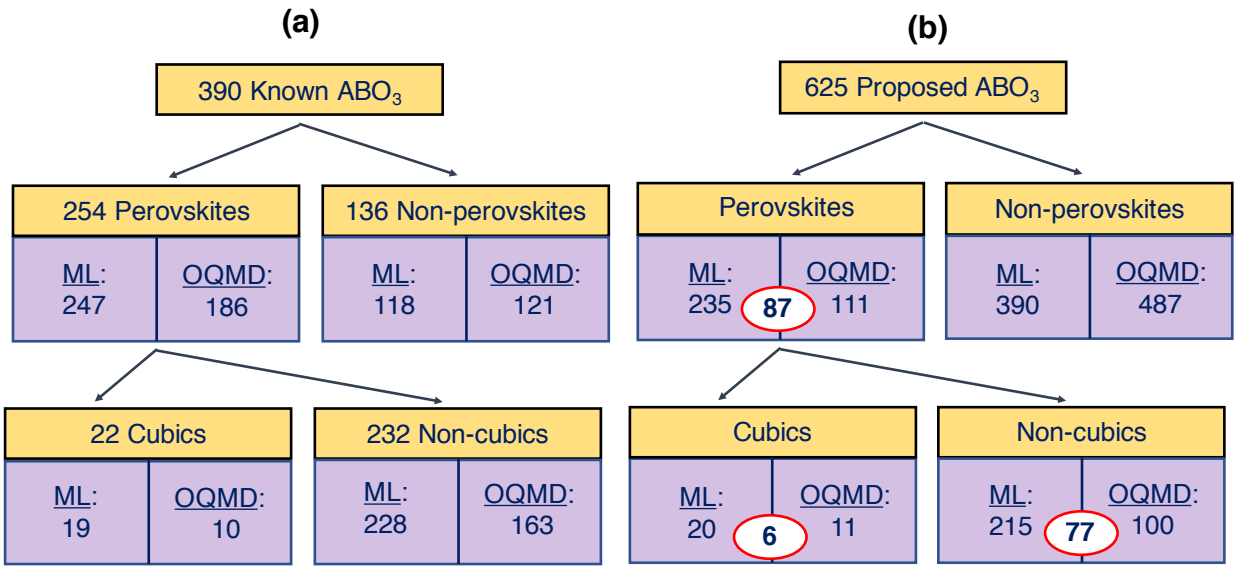


FIG. 10. Summary of the key outcome related to data from this work. (a) Out of the known 390 $ABO_3$ compounds, there are 254 Perovskites and 136 Non-perovskites. Our ML models, that were trained using the 390 $ABO_3$ compounds, classified 247 and 118 as Perovskites and Non-perovskites, respectively. On the other hand, OQMD predicts 186 and 121 as Perovskites and Non-perovskites, respectively. Similarly, among the 254 $ABO_3$ Perovskites, there are 22 compounds in the Cubic structure and 232 in the Non-cubic structures. ML classified 19 in the Cubic and 228 in the Non-cubic structures. OQMD, on the other hand, predicted 10 and 163 in the Cubic and Non-cubic structures, respectively. Not all known compounds were accurately captured by both ML and OQMD, which we discuss in detail in the Results section. (b) We enumerated a total of 625 compounds that were not present in the training set and then used our trained ML models and DFT-CH data in the OQMD to predict if there are potential perovskite compounds among them for synthesis. ML predicts 235 compounds in the perovskite structure, which in turn can be further subdivided into 20 and 215 Cubic and Non-cubic structures, respectively. On the other hand, OQMD predicts 111 $ABO_3$ compounds in stable perovskite structure, including 11 and 100 in the Cubic and Non-cubic structures, respectively. In total, we have 87 $ABO_3$ compounds that are predicted to be perovskite by both ML and OQMD of which both methods predict that 6 are Cubic and 77 are Non-cubic (details given in the Appendix). We identify these 87 compounds as promising for synthesis. We note that because of different methodologies involved, a direct comparison between ML and OQMD may not be appropriate.

perovskite. Similarly, Takatsu *et al.*[78] very recently synthesized $PbMoO_3$ in cubic perovskite structure by high-pressure and high temperature synthesis methods. Our ML models predict $PbMoO_3$ as perovskite, but with non-cubic crystal structure. The $\Delta$E from OQMD for $PbMoO_3$ in the cubic $Pm\bar{3}m$ perovskite structure is +137 meV/atom, indicating that the compound will likely decompose at P=0 and T=0 K. In both cases, the instability predicted by T=0 K DFT-CH is entirely consistent with the need for non-ambient conditions in experimental synthesis.

Ikeuchi *et al.*[79] recently synthesized $KWO_3$ in cubic perovskite structure by using a high pressure (7 GPa) and high temperature (1600° C) synthesis route. Both ML models and DFT-CH calculations in OQMD also predict $KWO_3$ in the cubic perovskite structure. More recently, Osaka *et al.*[80] synthesized $CaCoO_3$ in non-cubic perovskite structure using high-pressure oxygen annealing. ML and OQMD predict that this compound is a weakly metastable perovskite. Both also predict that this will be non-cubic, in agreement with the experimental work. More such comparisons are warranted to fully understand the advantages and limitations of ML and DFT-CH stability analysis.

## V. CONCLUSIONS

We performed a ML analysis of experimental data of $ABO_3$ solids known to be a perovskite or not and known to be a cubic perovskite or a non-cubic perovskite. From a list of possible new $ABO_3$ solids, we obtained similar perovskite or not and cubic or not predictions from two different ML methods. For additional consistency, we used the same cross-validation procedure for both the perovskite or not and the cubic or not cases. In choosing the cross-validation procedure, we were mainly targeting consistent performance for the cubic or not case which was difficult to achieve because so few cubics are in the data. In particular, we were finding that other cross-validation techniques were giving predictions that sometimes had large variances most likely due to overfitting the training data.

We emphasize our ML analyses and predictions are statistical in nature and hence are always subject to changes caused by fluctuations. Further, other ML approaches might produce results with higher accuracy if they were to use more features and optimize their hyperparameters for each case considered as opposed to our selecting just two features and using a one-size-fits-all setting of the hyperparameters. In another paper,[26] for example,

we demonstrated that using more than just pairs from the set of four feature pairs we could increase the accuracy of the predictions to nearly 95%. However, similar predictions of possible new perovskites were still made. In part, this improved accuracy is likely a consequence of using more parameters to fit the data as opposed to using features that delineate trends in the data better. Increasing the number of cubics in the database should improve the probability estimate of a predicted new cubic.

The feature pairs we considered were known beforehand to predict the formability of perovskites well. We did observe that in our training of the ML models for the cubic or not case, known perovskites with the A atom being Sr or Ba were misclassified as false positives or false negatives frequently. This systematic misclassification occurred to a lesser extent for compounds with the A atom being K or Rb. As already mentioned, all these misclassifications point to the need for at least one more feature with a chemical trend correlated with the labels in the database to improve classification accuracy with respect to this sub-class of compounds.

In closing, we note our use of OQMD and ML has produced a relatively large list of possible new perovskite compounds. While DFT-CH and ML are not predicting the same thing, we have made a hypothesis that when their predictions agree, the suggested compounds merit experimental study. While we are asserting that for many DFT-CH predictions T=0 K stability is a sufficient indicator for synthesizeability (*i.e.,* most stable phases can be synthesized), we remark that it is not a necessary condition (*i.e.,* a compound does not need to be stable in order to be synthesized). Now, we comment on the issues that warrant a better understanding and shed light on the plausible reasons that may have caused the DFT-CH and ML predictions to disagree on a relatively large number of known and possible new perovskites.

We first emphasize that what is known experimentally certainly includes a number of metastable compounds, something that DFT-CH can only address through including as possibly synthesizeable compounds those with a positive distance [a degree of metastability (DOM)] from our definition of the convex hull. Indeed, increasing the DOM from 0 to 100 meV/atom increased the OQMD agreement with known compounds from 60% to 70%.

Other reasons for the discrepancies between the experimental observations and OQMD predictions are common to all DFT-CH analyses and have at least two types of origins: physical and computational: (i) *Physical origins* mean that the synthesized compound could

correspond to a metastable structure of the element set A + B + O, *i.e.,* what forms in synthesis is not the lowest energy compound for these elements, but instead is a composition and structure kinetically trapped in a particular reaction path. Different synthesis methods and even different reaction protocols for the same synthesis method often produce different final products. (ii) *Computational origins* pertain to imperfections in the prediction engine. For example, the outcome of high throughput calculations depends on how versatile is the set of prototype structures used to gauge the stability of the target structure (here, $ABO_3$). If the set of competing constituent phases (components into which $ABO_3$ can decompose) is restricted, the calculations might predict a false positive stability of $ABO_3$, or if the set of candidate structures of the target phase $ABO_3$ is restricted, they might predict a false negative instability. An imperfect exchange correlation functional or an inappropriate assignment of a magnetic configuration (such as ferromagnetic, anti-ferromagnetic or paramagnetic) may sway a stability prediction to reactants instead of products or vice versa. Although one could research type (ii) discrepancies methodically by studying different approximations in a low-throughput manner, at this time we cannot determine how many of the present discrepancies are due to metastable synthesis conditions versus imperfect theoretical predictions.

For the present work, there might be a more material specific reason for some of the discrepancies. What we learned from the ML, relative to both the known and possible compounds, is OQMD and ML have systematic disagreements with respect to known and possible compounds involving lanthanide and actinide elements. Accurate energy calculations via DFT+U for such materials, which are typically strongly correlated, can be difficult. Clearly, with a material class as broad as the perovskites, we suggest that it is important to consider computational variations within a DFT-CH scheme to address the specific classes of chemical and structural complexity.

There is another aspect for reconciliation that is more difficult to address computationally. A DFT-CH analysis is performed at zero temperature with the expectation that the CH analysis with a non-zero DOM will adequately embrace what is happening at finite temperatures. In essence, the CH analysis is providing the likely positions of minima in the *internal energy* for different compounds, crystal structures, *etc.* and finding the one that is global. What is missing is how the entropic contributions ($-$TS term) to the *free energy* at finite temperatures will shift these positions and change their relative importance. Thus,

for a number of materials, the DFT-CH approach suffices but for others that are rich in structural and other phase transitions, such as the perovskites, it might not be so.

As we were concluding our manuscript, we became aware of a just released manuscript by Legrain *et al.*[81] that has a similar intent and conclusions as ours with respect to prediction of new materials using ML on data for known compounds versus using HT-DFT and CH analysis on lists of compounds without exploiting what is known experimentally. These authors compared the effectiveness of ML and DFT-CH calculations for the discovery of new half-heuslers compounds instead of perovskites. The predictions of ML and DFT-CH for compounds not yet known to be formed also showed significant inconsistencies. In these regards, their experience for the half-heuslers is similar to our experience for the perovskites.

We note that there are some important differences between our and their studies. One is the number of possible half-heusler structures is three and hence considerably smaller than the number of possible perovskite structures. In principle, the CH analysis and construction are less complex. Another difference is the known and predicted new half-heuslers are a small fraction of the total known and the total number of new possibilities. In this regard, the prediction task difficulties are similar to our cubic or not case. Nevertheless, Legrain *et al.* offer similar reasons for the inconsistencies between ML and HT-DFT, such as too few structures in the CH analysis and the inherent inaccuracies of DFT calculations. Their comparisons observed inconsistencies in the predictions of three different HT-DFT studies and underscores the importance of controlling these issues. Our phonon calculations, yielding different predictions for DFT calculations using different functionals and pseudopotentials, also reinforces the importance of controlling these issues.

# VI. APPENDIX

List of 87 promising $ABO_3$ compounds predicted as potentially formable in the perovskite structure by ML and as thermodynamically stable or nearly stable (DOM threshold set at 100 meV/atom) in the perovskite structures by HT-DFT in OQMD. Labels CP and P indicate cubic perovskite and non-cubic perovskite, respectively. $\Delta$H (in eV/atom) and $\Delta$E (in meV/atom) refer to formation enthalpy and distance from the CH, respectively; SG stands for space group number in the International symbol for which the OQMD energetic data ($\Delta$H and $\Delta$E) is reported. Predictions of compounds on the CH are "stronger predictions" and ones that are near the CH are metastable phases, which could also be synthesized. In OQMD, the CH distance for each of the $ABO_3$ compound is calculated in the same manner and thus these distances are comparable. Additional details can be found in the Supplemental Tables I–IV.

| Formula | ML prediction | OQMD prediction | $\Delta$H | $\Delta$E | SG |
|---|---|---|---|---|---|
| $CsPaO_3$ | CP | CP | -2.87 | 0 | 221 |
| $CsUO_3$ | CP | CP | -3.04 | 10 | 221 |
| $KReO_3$ | CP | CP | -1.92 | 8 | 221 |
| $KWO_3$ | CP | CP | -2.37 | 0 | 221 |
| $TlPaO_3$ | CP | CP | -2.56 | -358 | 221 |
| $TlUO_3$ | CP | CP | -2.75 | -108 | 221 |
| $EuCuO_3$ | P | CP | -1.78 | 60 | 221 |
| $HgPaO_3$ | P | CP | -2.17 | 1 | 221 |
| $NaReO_3$ | P | CP | -1.93 | 7 | 221 |
| $AgPaO_3$ | P | P | -2.33 | -361 | 167 |
| $AgUO_3$ | P | P | -2.51 | -9 | 167 |
| $BiCrO_3$ | P | P | -1.96 | 19 | 62 |
| $BiCuO_3$ | P | P | -1.06 | 30 | 62 |
| $BiLuO_3$ | P | P | -2.69 | 32 | 62 |
| $BiRhO_3$ | P | P | -1.18 | 60 | 62 |
| $BiVO_3$ | P | P | -2.05 | 7 | 62 |
| $CaCoO_3$ | P | P | -1.89 | 37 | 167 |
| $CaPuO_3$ | P | P | -3.34 | -109 | 62 |
| $CdPaO_3$ | P | P | -2.37 | -13 | 167 |
| $CdPuO_3$ | P | P | -2.49 | 4 | 167 |
| $CeCoO_3$ | P | P | -2.42 | -36 | 62 |
| $CeCuO_3$ | P | P | -2.15 | 4 | 167 |
| $CeInO_3$ | P | P | -2.69 | -4 | 62 |
| $CeNiO_3$ | P | P | -2.29 | 2 | 167 |
| $CeRhO_3$ | P | P | -2.29 | -61 | 62 |
| $CeRuO_3$ | P | P | -2.29 | 37 | 62 |
| $CeScO_3$ | P | P | -3.7 | -38 | 62 |
| $DyCuO_3$ | P | P | -2.28 | 47 | 62 |
| $DyGaO_3$ | P | P | -2.97 | 17 | 62 |
| $ErCoO_3$ | P | P | -2.55 | 24 | 62 |
| $ErGaO_3$ | P | P | -2.99 | 21 | 62 |
| $Eu^{II}Co^{IV}O_3$ | P | P | -2.14 | -38 | 62 |
| $Eu^{III}Co^{III}O_3$ | P | P | -2.14 | -38 | 62 |
| $EuCrO_3$ | P | P | -2.8 | -51 | 62 |
| $EuGeO_3$ | P | P | -2.7 | -212 | 167 |
| $EuHfO_3$ | P | P | -3.84 | -182 | 62 |
| $EuIrO_3$ | P | P | -2.1 | -43 | 62 |
| $EuMnO_3$ | P | P | -2.61 | 3 | 167 |
| $EuMoO_3$ | P | P | -2.81 | -101 | 62 |
| $EuNbO_3$ | P | P | -3.21 | -53 | 62 |
| $EuPaO_3$ | P | P | -3.34 | -123 | 62 |
| $EuPbO_3$ | P | P | -2.18 | -150 | 62 |
| $EuPuO_3$ | P | P | -3.5 | -187 | 62 |
| $Eu^{II}Ru^{IV}O_3$ | P | P | -2.26 | -112 | 62 |
| $Eu^{III}Ru^{III}O_3$ | P | P | -2.26 | -112 | 62 |
| $EuSnO_3$ | P | P | -2.7 | -171 | 62 |
| $EuTiO_3$ | P | P | -3.56 | 3 | 167 |
| $Eu^{II}V^{IV}O_3$ | P | P | -3.1 | -93 | 62 |
| $Eu^{III}V^{III}O_3$ | P | P | -3.1 | -93 | 62 |
| $EuZrO_3$ | P | P | -3.69 | -152 | 62 |
| $GdCuO_3$ | P | P | -2.26 | 43 | 62 |
| $HgHfO_3$ | P | P | -2.42 | 67 | 167 |
| $HgPuO_3$ | P | P | -2.17 | -28 | 167 |
| $HgZrO_3$ | P | P | -2.28 | 86 | 167 |
| $HoGaO_3$ | P | P | -2.98 | 20 | 62 |
| $HoVO_3$ | P | P | -3.27 | 1 | 62 |
| $LuCoO_3$ | P | P | -2.57 | 35 | 62 |
| $LuGaO_3$ | P | P | -3 | 30 | 62 |
| $LuNiO_3$ | P | P | -2.44 | 37 | 62 |
| $NdCuO_3$ | P | P | -2.18 | 74 | 62 |
| $NdRuO_3$ | P | P | -2.34 | 41 | 62 |
| $PbPaO_3$ | P | P | -2.41 | -22 | 62 |
| $PbPuO_3$ | P | P | -2.54 | -55 | 62 |
| $PrCuO_3$ | P | P | -2.17 | 3 | 167 |

| | | | | | |
|---|---|---|---|---|---|
| $PrInO_3$ | P | P | -2.71 | 5 | 62 |
| $PuGaO_3$ | P | P | -2.9 | -28 | 62 |
| $SmCuO_3$ | P | P | -2.22 | 47 | 62 |
| $SmGaO_3$ | P | P | -2.92 | 6 | 62 |
| $SmRuO_3$ | P | P | -2.37 | 49 | 62 |
| $SrCrO_3$ | P | P | -2.56 | 41 | 62 |
| $SrNpO_3$ | P | P | -3.42 | -14 | 62 |
| $SrPaO_3$ | P | P | -3.18 | -144 | 62 |
| $SrUO_3$ | P | P | -3.49 | -18 | 62 |
| $TbCuO_3$ | P | P | -2.14 | -40 | 62 |
| $TbGaO_3$ | P | P | -2.83 | 15 | 62 |
| $TbNiO_3$ | P | P | -2.27 | 8 | 62 |
| $TbScO_3$ | P | P | -3.66 | 19 | 62 |
| $TlMnO_3$ | P | P | -1.43 | 51 | 62 |
| $TmCoO_3$ | P | P | -2.57 | 27 | 62 |
| $TmGaO_3$ | P | P | -3 | 19 | 62 |
| $YbCoO_3$ | P | P | -2.11 | -79 | 62 |
| $YbRhO_3$ | P | P | -2.11 | -89 | 62 |
| $YbRuO_3$ | P | P | -2.25 | -83 | 62 |
| $YbScO_3$ | P | P | -3.21 | 98 | 62 |
| $EuErO_3$ | P | P | -3.21 | 98 | 62 |
| $EuLuO_3$ | P | P | -3.26 | 94 | 62 |
| $EuTmO_3$ | P | P | -3.23 | 90 | 62 |


## ACKNOWLEDGMENTS

This work of PVB, JEG, and TL was supported in part by the Laboratory Directed Research and Development program of the Los Alamos National Laboratory. AAE and CW were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Grant DE-FG02-07ER46433. AZ is supported by the US Department of Energy, Office of Science, Basic Energy Science, MSE Division under Grant No. DE-FG02-13ER46959. We thank J. Lashley for a helpful conversation. The machine learning calculations were performed by PVB, JEG and TL. The phonon calculations were performed by PVB. CW and AAE extracted the information from OQMD, and AAE performed additional DFT calculations. All authors participated in the analysis and interpretation of the results.


---


* txl@lanl.gov



[1] A. S. Bhalla, R. Guo, and R. Roy, Material Research Innovations **4**, 3 (2000).

[2] J. B. Goodenough, Reports on Progress in Physics **67**, 1915 (2004).

[3] M. A. Peña and J. L. G. Fierro, Chemical Reviews **101**, 1981 (2001).

[4] M. W. Lufaso and P. M. Woodward, Acta Crystallographica Section B **57**, 725 (2001).

[5] H. Zhang, N. Li, K. Li, and D. Xue, Acta Crystallographica Section B **63**, 812 (2007).

[6] See supplementary material at `http://link.aps.org/supplemental/XYZ` for details.

[7] J. G. Bednorz and K. A. Müller, Rev. Mod. Phys. **60**, 585 (1988).

[8] S. Jin, T. H. Tiefel, M. McCormack, R. A. Fastnacht, R. Ramesh, and L. H. Chen, Science **264**, 413 (1994).

[9] S.-W. Cheong and M. Mostovoy, Nat. Mater. **6**, 13 (2007).

[10] J. C. Boivin and G. Mairesse, Chemistry of Materials **10**, 2870 (1998).

[11] H. Jin, S. H. Rhim, J. Im, and A. J. Freeman, Scientific Reports **3**, 1651 (2013).

[12] G. Trimarchi, X. Zhang, A. J. Freeman, and A. Zunger, Phys. Rev. B **90**, 161111 (2014).

[13] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).

[14] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Comput. Mater. **1**, 15010 (2015).

[15] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, Chemistry of Materials **28**, 5621 (2016).

[16] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. persson, and G. Ceder, Sci. Adv. **2**, e1600225 (2016).

[17] E. Mooser and W. B. Pearson, Acta Crystallographica **12**, 1015 (1959).

[18] J. R. Chelikowsky and J. C. Phillips, Phys. Rev. B **17**, 2453 (1978).

[19] A. Zunger, Phys. Rev. B **22**, 5839 (1980).

[20] D. Pettifor, Journal of the Less Common Metals **114**, 7 (1985).

[21] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown, J. Chem. Inf. Model. **23**, 66 (1983).

[22] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallogr. Sect. B Struct. Sci. **58**, 364 (2002).

[23] P. V. Balachandran, S. R. Broderick, and K. Rajan, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science **467**, 2271 (2011).

[24] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).

[25] G. Pilania, J. Gubernatis, and T. Lookman, Phys. Rev. B **91**, 124301 (2015).

[26] G. Pilania, P. V. Balachandran, J. E. Gubernatis, and T. Lookman, Acta Cryst. B **71**, 507 (2015).

[27] P. V. Balachandran, J. Theiler, J. M. Rondinelli, and T. Lookman, Scientific Reports **5**, 13285 (2015).

[28] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).

[29] P. V. Balachandran, J. Young, T. Lookman, and J. M. Rondinelli, Nature Communications **8**, 14282 (2017).

[30] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).

[31] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Computational Materials **2**, 16028 (2016).

[32] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorganic Chemistry **50**, 656 (2011).

[33] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, New Journal of Physics **18**, 093011 (2016).

[34] A. Zakutayev, A. J. Allen, X. Zhang, J. Vidal, Z. Cui, S. Lany, M. Yang, F. J. DiSalvo, and D. S. Ginley, Chemistry of Materials **26**, 4970 (2014).

[35] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Mater. **1**, 011002 (2013), http://dx.doi.org/10.1063/1.4812323.

[36] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Computational Materials Science **58**, 227 (2012).

[37] K. Ito, K. Tezuka, and Y. Hinatsu, Journal of Solid State Chemistry **157**, 173 (2001).

[38] C.-Q. Jin, J.-S. Zhou, J. B. Goodenough, Q. Q. Liu, J. G. Zhao, L. X. Yang, Y. Yu, R. C. Yu, T. Katsura, A. Shatskiy, and E. Ito, Proceedings of the National Academy of Sciences **105**, 7115 (2008).

[39] D. M. Giaquinta and H.-C. zur Loye, Chemistry of Materials **6**, 365 (1994).

[40] L. M. Feng, L. Q. Jiang, M. Zhu, H. B. Liu, X. Zhou, and C. H. Li, J. Phys. Chem. Solids **69**, 967 (2008).

[41] C. Li, K. C. K. Soh, and P. Wu, Journal of Alloys and Compounds **372**, 40 (2004).

[42] R. D. Shannon, Acta. Cryst. A **32**, 751 (1976).

[43] I. D. Brown, Chemical Reviews **109**, 6858 (2009).

[44] P. Villars, K. Cenzual, J. Daams, Y. Chen, and S. Iwata, Journal of Alloys and Compounds **367**, 167 (2004).

[45] L. Breiman, Machine Learning **45**, 5 (2001).

[46] J. H. Friedman, The Annals of Statistics **29**, 1189 (2001).

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret-

tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[48] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[49] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[50] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[51] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[52] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton, Phys. Rev. B **57**, 1505 (1998).

[53] L. Wang, T. Maxisch, and G. Ceder, Phys. Rev. B **73**, 195107 (2006).

[54] A. A. Emery and C. Wolverton, Scientific Data **4**, 170153 (2017).

[55] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, Journal of Physics: Condensed Matter **21**, 395502 (2009).

[56] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[57] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).

[58] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[59] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[60] A. D. Corso, Computational Materials Science **95**, 337 (2014).

[61] M. Retuerto, S. Skiadopoulou, M.-R. Li, A. M. Abakumov, M. Croft, A. Ignatov, T. Sarkar, B. M. Abbett, J. Pokorný, M. Savinov, D. Nuzhnyy, J. Prokleška, M. Abeykoon, P. W. Stephens, J. P. Hodges, P. Vaněk, C. J. Fennie, K. M. Rabe, S. Kamba, and M. Greenblatt, Inorganic Chemistry **55**, 4320 (2016).

[62] M. De La Pierre, R. Orlando, L. Maschio, K. Doll, P. Ugliengo, and R. Dovesi, Journal of Computational Chemistry **32**, 1775 (2011).

[63] C. Eames, J. M. Frost, P. R. F. Barnes, B. C. O'Regan, A. Walsh, and M. S. Islam, Nature Communications **6**, 7497 (2015).

[64] L. He, F. Liu, G. Hautier, M. J. T. Oliveira, M. A. L. Marques, F. D. Vila, J. J. Rehr, G.-M.

Rignanese, and A. Zhou, Phys. Rev. B **89**, 064305 (2014).

[65] M. Ye and D. Vanderbilt, Phys. Rev. B **95**, 014105 (2017).

[66] N. Charles and J. M. Rondinelli, Phys. Rev. B **94**, 174108 (2016).

[67] N. A. Benedek and C. J. Fennie, The Journal of Physical Chemistry C **117**, 13339 (2013).

[68] F. Brivio, A. B. Walker, and A. Walsh, APL Materials **1**, 042111 (2013).

[69] U. Aschauer and N. A. Spaldin, Journal of Physics: Condensed Matter **26**, 122203 (2014).

[70] S. Liu, I. Grinberg, and A. M. Rappe, Nature **534**, 360 (2016).

[71] S. L. Dudarev, L.-M. Peng, S. Y. Savrasov, and J.-M. Zuo, Phys. Rev. B **61**, 2506 (2000).

[72] A. Togo, F. Oba, and I. Tanaka, Phys. Rev. B **78**, 134106 (2008).

[73] K. M. Rabe, J. C. Phillips, P. Villars, and I. D. Brown, Phys. Rev. B **45**, 7650 (1992).

[74] D. J. Singh, Phys. Rev. B **53**, 176 (1996).

[75] Y. Ichikawa, M. Nagai, and K. Tanaka, Phys. Rev. B **71**, 092106 (2005).

[76] M. Rey, P. Dehaudt, J. Joubert, B. Lambert-Andron, M. Cyrot, and F. Cyrot-Lackmann, Journal of Solid State Chemistry **86**, 101 (1990).

[77] K. Nishimura, I. Yamada, K. Oka, Y. Shimakawa, and M. Azuma, Journal of Physics and Chemistry of Solids **75**, 710 (2014).

[78] H. Takatsu, O. Hernandez, W. Yoshimune, C. Prestipino, T. Yamamoto, C. Tassel, Y. Kobayashi, D. Batuk, Y. Shibata, A. M. Abakumov, C. M. Brown, and H. Kageyama, Phys. Rev. B, **95**, 155105 (2017).

[79] Y. Ikeuchi, H. Takatsu, C. Tassel, Y. Goto, T. Murakami, and H. Kageyama, Angewandte Chemie International Edition **56**, 5770 (2017).

[80] T. Osaka, H. Takahashi, H. Sagayama, Y. Yamasaki, and S. Ishiwata, Phys. Rev. B **95**, 224440 (2017).

[81] F. Legrain, J. Carrete, A. van Roekeghem, G. K. Madsen, and N. Mingo, The Journal of Physical Chemistry B **122**, 625 (2018).