

Machine learning substitutional defect formation energies in ABO_3 perovskites

Cite as: J. Appl. Phys. **128**, 034902 (2020); doi: [10.1063/5.0015538](https://doi.org/10.1063/5.0015538)

Submitted: 28 May 2020 · Accepted: 24 June 2020 ·

Published Online: 15 July 2020



Vinit Sharma,^{1,2,a)}  Pankaj Kumar,³  Pratibha Dev,³  and Chanshyam Pilania⁴ 

AFFILIATIONS


¹Joint Institute for Computational Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA

²National Institute for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

³Department of Physics and Astronomy, Howard University, Washington, D.C. 20059, USA

⁴Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Note: This paper is part of the special collection on Machine Learning for Materials Design and Discovery

a) Author to whom correspondence should be addressed: vinit.sharma@utk.edu 

ABSTRACT

Perovskite oxides are a promising material platform for use in a wide range of technological applications including electronics, sensors, fuel cells, and catalysis. This is owing to the extraordinary tunability of their physical and chemical properties via defect engineering. The feasibility and the stability of a defect, such as a substitutional dopant, in the host lattice is usually obtained via experiments and/or through detailed quantum mechanical calculations. Both of these conventional routes are expensive and time consuming. An alternative is a data-driven machine learning (ML)-based approach. In this work, we have applied ML techniques to identify the factors that influence defect formation energy, which is an important measure of the stability of the defects, in perovskite oxides. Using 13 elemental properties as features and random forest regression, we demonstrate a systematic approach to down-selecting from the larger set of features to those that are important, establishing a framework for accurate predictions of the defect formation energy. We quantitatively show that the most important factors that control the dopant stability are the dopant ionic size, heat of formation, effective tolerance factor, and oxidation state. Our work reveals previously unknown correlations, chemical trends, and the interplay between stability and underlying chemistries. Hence, these results showcase the efficacy of ML tools in identifying and quantifying different feature-dependencies and provide a promising route toward dopant selection in the perovskites. We have developed a framework that itself is general and can be potentially applied to other material classes.

Published under license by AIP Publishing. <https://doi.org/10.1063/5.0015538>

I. INTRODUCTION

Perovskite oxides have a chemical formula of ABO_3 , where the larger A-site cation is 12-fold coordinated and the smaller B-site cation is sixfold coordinated to oxygen. The perovskites offer material solutions for an ever-increasing range of technological applications including electronics, sensors, fuel cells, and catalysis.^{1–8} Different functionalities of the perovskites originate from the large number of cations of different sizes that can take up positions A and B in the lattice. The properties can be further tuned by substitutional doping in either of the cationic sites, opening up avenues for further tailoring physical and chemical properties in this crystal structure.^{8,9} In fact, the chemical doping of ABO_3 compounds is essential in many applications where, for instance, one needs to tune the bandgap of the host material,¹⁰ enhance or suppress ionic

conductivity,^{11,12} manipulate and control the electrical and electro-active behavior,^{13,14} tune magnetic properties such as the Curie temperature and domain switching,^{15,16} or optimize the electrochemical or catalytic activity of the host material.^{7,17,18}

Although chemical doping in perovskites has been attempted in many experimental and theoretical studies,^{3–9,19–24} the underlying strategies have largely been based on costly and time consuming trial and error methods guided by chemical intuition. While the traditional approaches have had a considerable impact, a quantitative understanding of factors that dictate thermodynamical stability and substitutional site preference of various chemical dopant in a perovskite oxide has not been well established.^{25–32} A systematic study that aims to understand physiochemical factors governing substitutional dopant stability trends over a large

chemical space can potentially enable Hume–Rothery-like rules³³—well established for substitutional solid solutions in metals and alloys—for perovskite oxides and beyond. Such a generalization will be useful in developing advanced dopant selection strategies for materials design and optimization efforts for a target application. To this end, machine learning (ML) techniques can be used to predict and understand trends in defect energetics and their implications on the underlying property landscape. However, there are only limited numbers of such studies.^{34,35} In this work, we use ML techniques that are combined with high-throughput first principles computations, to identify key factors that dictate substitutional dopant stability in perovskite oxides.

As an important step toward learning chemical trends in defect energetics, we employ an extensive dataset of cation substitutional defects in ABO₃ perovskite oxides in order to (i) develop a predictive ML model for relative defect formation energy and (ii) identify the key elemental descriptors that govern the underlying chemical trends in the energetics. More specifically, as host materials, we have considered BaTiO₃ in the cubic (*c*-BTO) and the rhombohedral (*R*-BTO) phases as shown in Figs. 1(a) and 1(b), respectively, and LaMnO₃ in the cubic (*c*-LMO) phase, shown in Fig. 1(c). For each of the host chemistries, a set of substitutional dopants are considered for both cationic sites. The dopant atoms cover a large part of periodic table, ranging from alkali and alkaline earth metals, through 3*d*, 4*d*, and 5*d* transition metals up to group VA of the periodic table (K–As, Rb–Sb, and Cs–Bi). The target dopant chemical space is depicted in Fig. 1(d). The database generated from accurate first principles-based density functional theory (DFT) computations was subsequently employed to identify a set of most relevant elemental descriptors to learn, validate, and test a predictive random forest-based ML regression model. While the feature selection was limited to training data comprised of defect formation energies with *R*-BTO phase as a host, it is shown that the set of identified features can be generalized beyond the specific host chemistry and are capable of reliably predicting the defect formation energies for the *c*-BTO and *c*-LMO host phases as well.

II. TECHNICAL DETAILS

A. First principles computations

The substitutional defect formation energy training dataset was generated with *ab initio* DFT calculations, performed within a plane-wave basis set formalism using the projector augmented wave (PAW) method as implemented in the Vienna *ab initio* Simulation Package (VASP) code.^{36–38} A plane-wave energy cutoff of 520 eV was consistently used throughout this study. The exchange correlation effects were included through the Perdew, Burke, and Ernzerhof (PBE) generalized gradient approximation (GGA).³⁹

The total energies of hosts containing a substitutional defect at either a 12-fold or a sixfold coordinated cation site were calculated using a 40-atom, 2 × 2 × 2 supercell containing a total of eight formula units of the host chemistries, namely, *c*-BTO, *R*-BTO, and *c*-LMO. For each of the host lattice, substitutional dopants ranging from K to As, Rb to Sb, and Cs to Bi in the three periods across the periodic table were considered. A Monkhorst–Pack *k*-point mesh⁴⁰ of 8 × 8 × 8 was employed for the reciprocal space integration to produce total energies converged within 0.1 meV per formula unit. After introducing substitutional dopants in the cationic sites, we performed a full atomic relaxation of internal coordinates and optimization of lattice parameters. The atomic positions were relaxed until the forces on all atoms were less than 0.02 eV/Å. Additional details on specific choices of VASP pseudopotentials used to describe valence electrons and core ion interactions for each of the elemental species are provided in Table SI of the [Supplementary material](#) accompanying the paper.

The dopant formation energies E_f^D in an ABO₃ host lattice were calculated as

$$E_f^D = E_{\text{ABO}_3}^D - E_{\text{ABO}_3}^H - (\mu^D - \mu^H), \quad (1)$$

where $E_{\text{ABO}_3}^H$ and $E_{\text{ABO}_3}^D$ are the total energies of the pure and doped ABO₃ supercells, respectively. The formation energy for a dopant *D*, occupying a host site *H* (i.e., the A- or B-site cation),

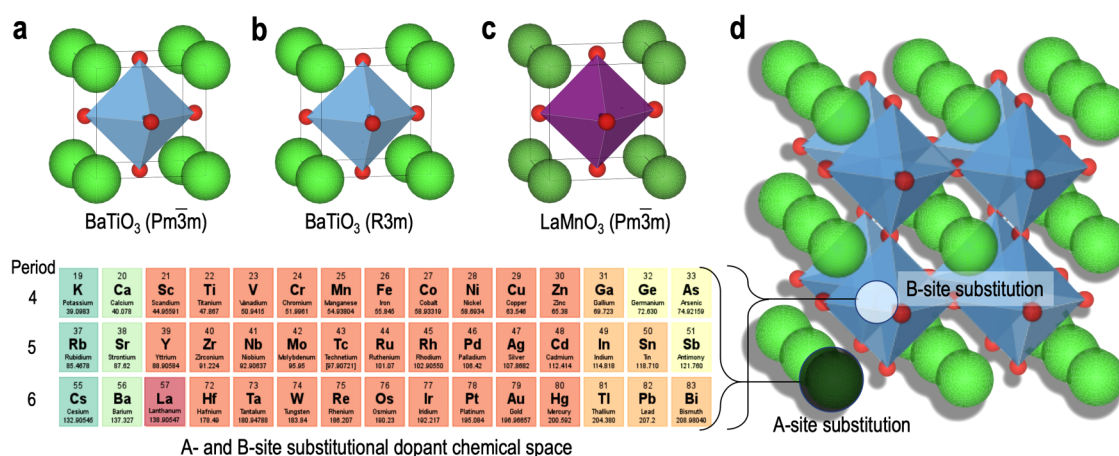


FIG. 1. Perovskite host chemistries (a) *c*-BTO, (b) *R*-BTO, (c) *c*-LMO, and (d) the entire target substitutional dopant chemical space considered in this study is shown. Furthermore, the 12-fold and sixfold coordinated A and B dopant substitutional sites in a perovskite crystal structure are explicitly identified in (d).

depends on the particular choice of the reference atomic chemical potentials μ^D and μ^H . In the present study, the chemical potential of either the host or the dopant atoms, μ^M , is defined with respect to the most stable binary oxides of the respective elemental species, as

$$\mu^M = \frac{1}{y} \left(E_{M_yO_x} - \frac{x}{2} \mu_{O_2} \right), \quad (2)$$

where $E_{M_yO_x}$ is the DFT energy per formula unit of the most stable oxide M_yO_x , M represents either the host or the dopant cation, x and y denote the stoichiometry of the oxide, and μ_{O_2} denotes the oxygen chemical potential, taken here as the computed total energy of an isolated oxygen molecule with the following caveat. Since GGA is well known for overbinding an O_2 molecule,^{8,41–43} we have corrected this artificial stabilization by destabilizing the O_2 molecule by 0.77 eV/ O_2 molecule,⁸ in order to correctly reproduce experimentally measured formation energies for a range of binary oxides.^{42,43}

Lastly, we note that the GGA-PBE functional is known to underestimate bandgaps in transition metal compounds.^{44,45} This can be fixed by incorporating an effective on-site Coulomb interaction for the transition metal's d states using the GGA+U method.^{46,47} However, in this work, we are interested in total energy differences, and the GGA-PBE functional gives reasonably accurate relative energy differences for a wide range of chemistries, including those containing transition metal ions.^{8,42,43,48} We also explicitly confirmed the accuracy of calculations using GGA-PBE functional by directly comparing the DFT-computed formation energies against the corresponding experimentally measured values for a set of simple binary oxides spanning the entire set of 45 cation chemistries considered in the present work (see Fig. S1 in the [Supplementary material](#)).

B. Details of feature dataset

Our DFT-computed defect formation energies are presented in [Fig. 2](#). While further details for the chemical trends displayed by the computed energetics will be discussed in the [Sec. III A](#), here, we note that the formation energies can span over a wide range, varying from 0 to ~ 15 eV, depending on host chemistry, chemical

nature of the substitutional dopant, and local coordination environment of the dopant site. Given the target property dataset, our primary aim at this stage is to identify easily accessible attributes or features that not only allow us to uniquely represent a dopant-host perovskite system but also enable a physically meaningful and predictive mapping via employing a suitable ML model. Toward this goal, we start by hand-picking a set of 13 features, representing properties of isolated atoms or bulk elemental species of the corresponding substitutional dopant computed relative to the substituted host cation species. In other words, each of the features used here represents an atomic or bulk elemental property of the dopant relative to the substituted host cation species. More specifically, the atomic and bulk elemental features considered here include coordination-dependent Shannon's ionic radius (IR), atomic radius (AR), metallic radius (MR), Pauling's electronegativity (EN), first ionization potential (IP), electron affinity (EA), bulk heat of formation (HF), boiling point (BP), melting point (MP), modified (compositionally averaged) tolerance factor (TF), oxidation state (OS) displayed in its most thermodynamical stable binary oxide, substitutional defect's most commonly exhibited valence state (VS), and dipole polarizability (DP).^{49–51} Further details of the primary features along with entire feature set used in this study, along with the DFT-computed dopant formation energy dataset, is also made available in the [Supplementary material](#) provided with this manuscript.

Starting with the primary feature set, next, we perform a pairwise feature correlation analysis to remove redundant information from the initial feature set. To quantify correlation between any two features, our feature correlation analysis employs the well-known Pearson correlation coefficient, \mathcal{P} , defined for two N dimensional column vectors, \mathbf{v}_1 and \mathbf{v}_2 , pair-wise, as

$$\mathcal{P}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\sum_{i=1}^N [v_1(i) - \bar{v}_1][v_2(i) - \bar{v}_2]}{\sqrt{\sum_{i=1}^N [v_1(i) - \bar{v}_1]^2} \sqrt{\sum_{i=1}^N [v_2(i) - \bar{v}_2]^2}}, \quad (3)$$

where \bar{v}_1 and \bar{v}_2 denote the corresponding means and $v(i)$

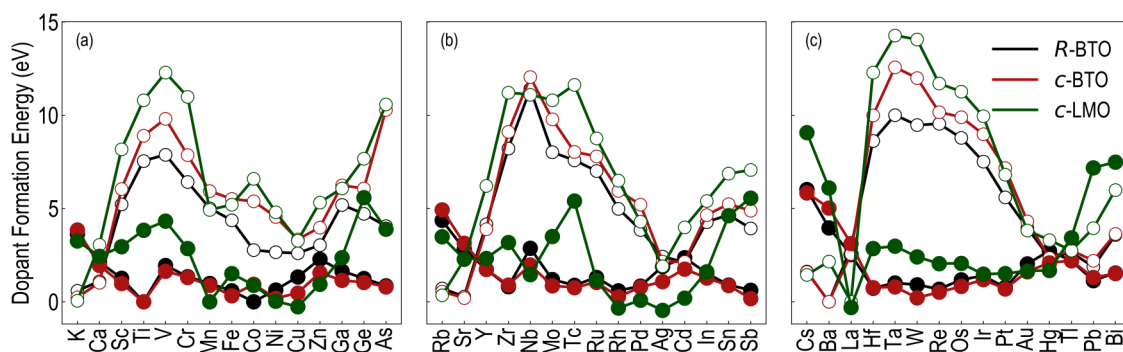


FIG. 2. Calculated defect formation energies for different substitutional dopants at either A- or B-site in R-BTO, c-BTO, c-LMO. Dopants considered here are from the (a) 3d, (b) 4d, and (c) 5d series, along with their neighboring Group IA–VA elements in the periodic table. Open and filled circles correspond to substitutional dopants at the A and B sites, respectively.

represents the i th component of the vector \mathbf{v} . In our case, i labels a host perovskite chemistry doped at a specific cation site with a particular substitutional dopant chemistry in the dataset comprised of N training examples. Note that the correlation coefficient \mathcal{P} always lies between -1 and $+1$, representing allowed limiting values for a perfect negative or positive correlation, respectively.

C. Random forest regression model

Based on the correlation analysis, a down-selected subset of the initial feature set is next used for the training and validation of a ML regression model to predict the dopant formation energies. For ML, we employ a highly efficient and effective learning algorithm known as the random forest (RF) regression⁵² proposed by Breiman in 2001. The RF algorithm is built upon an ensemble of decision trees^{53,54} and is closely related to *bagging*-based ensemble-learning methods.⁵⁵ In bagging based decision tree methods, successive trees are constructed independently using a bootstrap sample of the dataset while considering all variables/features at each node. In contrast, RF presents an additional layer of randomness, where in addition to constructing each tree using a different bootstrap sample of the data, each node is split by employing only a subset of features randomly chosen at that node. In other words, the decisions trees in the forest are constructed by selecting not only a subset of training samples but also a fraction of features. This strategy leads to a highly effective (i.e., both efficient and predictive) ML model that is naturally robust against overfitting—an undesirable situation where a model's prediction performance on unseen data is much worse than the training performance.⁵⁶

Common to any ML algorithm, RF regression also has model parameters that need to be explicitly optimized as a part of the model training procedure. The two major parameters are the maximum allowable tree depth in a decision tree and the number of trees (or estimators) that make the forest. Although RFs are not very sensitive to the specific values chosen for these parameters (if sufficient model flexibility is provided via selecting overestimated values for both the parameters), it is generally desirable to optimize values of these two parameters during the model training stages. Finally, we note that throughout this paper, we have consistently used a randomly selected 90%/10% training/test split and a fivefold cross-validation. Random forest regression model implementation available within scikit-learn was used in the present work.⁵⁷

III. RESULTS AND DISCUSSION

A. Substitutional defect formation energies

In this section, we analyze qualitative chemical trends exhibited by the computed substitutional dopant formation energies as a function of the dopant's chemical nature, substitutional site, and host chemistry. We also make a connection between our computational results with previously reported experimental findings wherever possible.

Form the results presented in Figs. 2(a)–2(c), it can be seen that the dopant formation energies portray qualitatively similar trends across the three periods of the periodic table as well as for the three host chemistries. The computed energetics are closely aligned for *R*-BTO and *c*-BTO hosts. Owing to the distinct charge

states exhibited by the A- and B-site chemistries in LMO (both $+3$ as opposed to $+2$ and $+4$ in BTO), in some cases dopant energetics computed for this host chemistry exhibit deviations with respect to those for the two BTO host chemistries. Overall, the results suggest that while alkali and alkaline earth metal ions favor the A-site substitution, a strong preference for the B-site is present among the transition and post-transition metal ions.

The trends observed in our computed defect formation energies are corroborated by previously reported experimental measurements in the literature.^{7,16,25–27,58–75} The finding that Group IA (K, Rb, Cs) and Group IIA (Ca, Sr, Ba) cations favor the A-site substitution is consistent with past reports showing that K,^{58–60} Ca,^{16,61,62} Ba,⁶² and Sr^{7,62,63} prefer the A-site. On the other hand, the transition metal ions from the 3d (Cr, Mn, Fe, Co, Ni, Cu, Zn),^{16,66–69} 4d (Zr, Nb, Mo, Rh, Cd),^{16,66–69} and 5d series (Hf, Ta, Re, Pb),^{69–71} along with Group IIIA elements (Ga, In)^{62,76} tend to prefer the B-site substitution. Another important feature observed in the trends is a crossover between the preference for A-site to B-site that can be seen going from Ca to Sc, Sr to Y, and Ba to La in the third, fourth and fifth periods, respectively. A similar, albeit less pronounced, tendency for a crossover is also observed in the post-transition metal ions. Consistent with these crossover trends, past experimental reports from the literature have shown that Ca,⁷² Y,⁷³ Cd,⁷⁴ and La⁷⁵ can indeed occupy the A- or B- sites in BTO.

At a qualitative level, the results presented in Fig. 2 reveal that local strain and electrostatic effects—caused, respectively, by the substitutional dopant and the corresponding host element's size and oxidation state mismatches—tend to govern chemical trends in the computed substitutional defect formation energies. As a natural next step, we use correlation-based feature down-selection analysis and RF ML regression to convert these observed qualitative trends in the computed energetics into a quantitative and predictive model.

B. Feature set down-selection

The primary goal of any ML-based materials property prediction approach is to establish an interpolative and predictive mapping between a set of easily accessible material attributes or features and the property of interest. As the number of features involved in the development of the model increases not only does the model complexity increase but also the span of the underlying interpolative space increases exponentially. Therefore, to avoid the curse of dimensionality and to conform to the well established conventional wisdom in the field of statistical learning that *simpler models are better generalizable*, it is highly desirable to first identify a reduced set of most relevant features which are subsequently employed to develop a predictive model of the defect formation energies.

Toward this aim, starting with the initial set of features described in Sec. II B, in this step, we use pairwise feature correlation analysis to (i) remove redundant information in the initial feature set and (ii) down-select a subset of features most relevant toward predicting the target property. For the correlation-based feature down-selection, we proceed as follows: if any two features in the initial dataset show an absolute Pearson correlation coefficient $\mathcal{P} > 85\%$, the one that is poorly correlated with the target property is eliminated. At the same time, we also require that any

down-selected feature should exhibit at least $|\mathcal{P}| > 10\%$ with the defect formation energy. The Pearson correlation matrix, presented in Fig. 3(a) for the *R*-BTO dataset, shows that indeed several features in the initial feature set are highly correlated. Pairwise correlations for the reduced feature set obtained after the aforementioned screening are shown by the correlogram presented in Fig. 3(b). Further, the top row and left column show the correlations of the down-selected features with the defect formation energy; the screened features are arranged in an increasing order of their importance (i.e., the absolute correlation coefficients with the property) going from left to right (or top to bottom). In particular, the correlation-based analysis identifies TF, OS, HF, and IR as the top four most important features in this case, which are highlighted in Fig. 3(b).

In addition to correlation-based feature selection, the random forest ML algorithm itself provides an in-built measure for estimating relative feature importance. A random forest consists of a number of decision trees where a node in the decision trees is typically represented by a condition on a single feature, designed to split the dataset into two subsets following a pre-specified criterion. In classification problems, it is typical to use an information gain or entropy reduction as the criterion, while variance reduction at the node is generally used for a regression problem. Thus, to compute the relative feature importance in a random forest regression model, first variance reduction due to each feature is computed for each of the individual trees that make the forest. Subsequently, this value for each feature is averaged over the entire forest to generate a relative ranking measure among the features. Note that while the correlation-based feature analysis is solely built on feature–feature correlations and correlations of the individual features with the target property, the random forest-based relative feature importance analysis also implicitly accounts for feature interaction effects. The latter is a phenomenon where inclusion or exclusion of a feature in the

feature set affects the performance of a different feature toward prediction of the target property.

In Fig. 3(c), we present the relative feature importances computed for the initial feature set using a random forest regression model, targeted to predict the defect formation energies over the *R*-BTO dataset. The bars and the error bars in the plot represent the means and standard deviations of the relative feature importances, respectively, computed over five different randomly selected 90% training sets that were separately used to train the random forest regression model. Sufficiently large values for the tree depth and the number of trees in the forest were used to give converged performance on the test set before computing the reported feature importances. Note that the top four features identified by our feature correlation analysis also unambiguously appear as the top four most important features in the random forest feature importance analysis.

Before moving to the ML model development and validation exercise using the down-selected feature set, in Fig. 4, we visually analyze the correlations of the screened features with the defect formation energies for the *R*-BTO dataset. Each of the down-selected features exhibits strong substitutional site dependent trends with the dopant formation energy. These relations are also quite intuitive and physically meaningful. Overall, Fig. 4 clearly demonstrates that dopant size, charge state, and chemical nature largely govern the DFT-computed dopant formation energies. For a particular substitutional site, as the size (quantified by the relative ionic radius) and/or charge state (quantified by the relative oxidation state) deviate from those of the host chemistry, the dopant formation energy generally rises. Unlike the relative ionic radius feature, the effective (or compositionally averaged) tolerance factor accounts for both the host cations in quantifying the local strain effects. Lastly, the relative heat of formation feature captures a measure of chemical and thermodynamic similarity between the dopants and the host chemistries. After rationalizing the underlying physically meaningful causal relationships, next, we turn to ML for

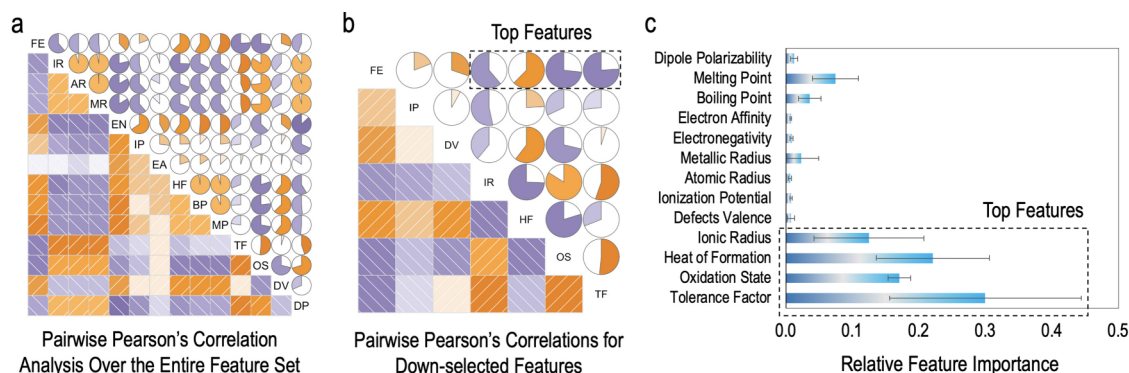


FIG. 3. (a) Pearson correlation matrix capturing pairwise feature–feature and property–feature correlations for the *R*-BTO dataset. The target property (FE) and the features in the initial dataset are listed in the diagonal; positive and negative correlations are shown using orange and purple colors, respectively. The upper and lower triangular regions of the plot convey the same information in two different visualization schemes. The filled fraction of the pie charts in the upper triangle represents the absolute value of the associated Pearson correlation coefficient, while the lighter and darker shades of the colors also directly correspond to the strength of the correlation. (b) The pairwise correlations for the down-selected feature sets. The top row identifies correlation of the down-selected features with the target. The four most correlated features with the computed defect formations energies in *R*-BTO are highlighted. (c) RF-computed relative feature importance of all features considered in the present work.

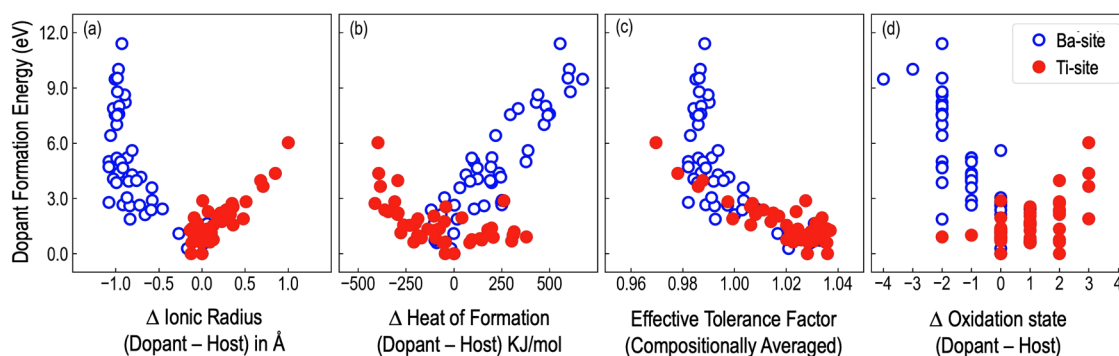


FIG. 4. Trends in the computed substitutional defect formation energies at the A or B substitutional sites in *R*-BTO with respect to the four most important features identified via our feature down-selection approach, namely, (a) Shannon's ionic radius of the dopants relative to the host atom, (b) elemental bulk heat of formation of the dopants relative to the host chemistry, (c) compositionally averaged tolerance factor (see the [Supplementary material](#) for details), and (d) the oxidation state of the dopants relative to the host atom computed in the respective most stable binary oxide chemistries.

crystallizing this qualitative understanding into a quantitative and predictive model.

C. Predictive ML model development

As mentioned in Sec. II C, an RF model has two model hyperparameters: namely, the maximum allowable tree depth and the number of trees (or estimators) in the forest. If trees in the forest are too shallow then they might have a high bias, leading to an underfitting. On the other hand, if the trees are too deep, then they will have a low bias or high variance, translating to an overfitted model. Although going from decision trees to an ensemble (or the forest) generally alleviates the latter issue and RFs are inherently robust against overfitting. Therefore, any sufficiently large values for these two parameters should, in principle, lead to a reasonably accurate prediction performance. However, a shallow choice of maximum allowed

tree depth and insufficient number of estimators in the ensemble can potentially lead to a sub-optimal performance. This bias-variance trade-off is handled by tuning hyperparameters using cross-validation. Therefore, to start with, we systematically evaluate the model's predictive performance with respect to the two parameters.

To choose an optimal value for each of the two hyperparameters, we train different RF models with systematically varying choices of the maximum allowable tree depth and the number of estimators employed. For each set of hyperparameters, we then evaluate the trained model's performance on a randomly selected and left-out 10% test set. Furthermore, to account for the sensitivity of the predictions with respect to different randomly selected training/test splits, each model is trained on 20 different training/test splits. The average test set RMSE in dopant formation energy and its variance computed over these 20 different test set selections [presented in Figs. 5(a) and 5(b), respectively], then

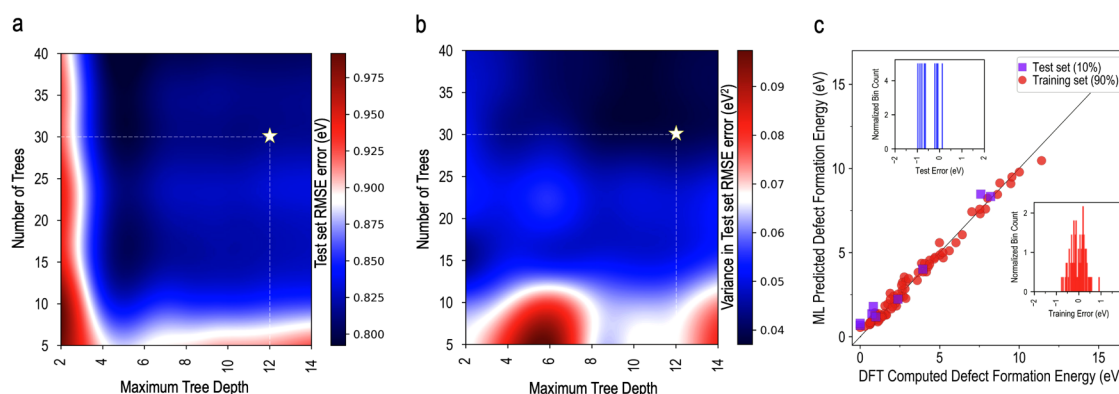


FIG. 5. The test set average root mean squared error (RMSE) (a) and variance (b) in the predictive performances (computed over 20 different randomly selected training/test splits) as a function of the RF model hyperparameters, namely, the maximum allowable tree depth and number of trees in the forest. (c) A representative parity plot comparing the DFT-computed defect formation energies against those predicted via the developed ML model. The insets show histograms of the test and training set errors.

provide a measure of the model's predictive performance on unseen data and the associated uncertainties for those predictions.

Results presented in Fig. 5(a) show that the model's performance depends less sensitively upon the number of estimators as compared to the maximum tree depth. Furthermore, beyond a maximum tree depth of 4 and number of estimators above 10, the achieved prediction performance is rather flat, reconfirming the well-known overfitting-resistant behavior of the RF model. Finally, we note that as the maximum allowed tree depth and the number of estimators increase to 12 and 30, respectively, the average RMSE error on unseen data remains largely unchanged but the model's predictive uncertainties shrink [Fig. 5(b)]. Based on these results, we set the values of the number of estimators and the maximum tree depth to 30 and 12, respectively, as indicated by a "★" in Figs. 5(a)–5(b). With these parameters, the average training and test RMSE computed over five different randomly selected 90%-10% training–test splits for the *R*-BTO dataset are 0.43 eV and 0.87 eV, respectively. A representative parity plot directly comparing the DFT-computed dopant formation energies with those predicted using the developed ML model for the training and test data are shown in Fig. 5(c).

Thus far, our analysis has focused only on the dopant formation energies computed for the *R*-BTO host. While the excellent performance of the developed four-feature RF regression model in predicting the substitutional dopant formation energies over a vast chemical space is already exciting, next, we perform a more stringent test for the predictive power of the selected features. To test how the predictive performance generalizes to other materials going beyond the *R*-BTO host, we augment the formation energy dataset with the computed substitutional dopant formation energies for the *c*-BTO and *c*-LMO hosts. Considering these two host chemistries allows us to probe the effect of changing the host crystal structure (while keeping the chemistry constant, in *c*-BTO) as well as changing both the chemistry and crystal structure simultaneously *c*-LMO.

An important point to consider here is that our adopted feature engineering scheme, which accounts for atomic or bulk elemental attributes of various dopants relative to the host elemental chemistries, can naturally address different host materials with varying chemistry. However, there is no feature that makes a distinction among different hosts with the same chemistry but varying crystal structures (e.g., *c*-BTO and *R*-BTO). Therefore, at this stage, we explicitly augment our feature set with a categorical feature distinguishing the three host chemistries. Subsequently, we retain the RF ML regression model with these five features on the entire dopant formation energy dataset again using a 90%–10% training–test split. The model's performance is presented in Fig. 6, in the form of a parity plot. More quantitatively, the RMSE errors on the training and test datasets were found to be 0.59 eV and 0.98 eV, respectively. Note that in this case, the RMSE error obtained on the unseen data is quite comparable to 0.87 eV obtained for the *R*-BTO dataset, indicating a good generalizability and high relevance of the identified features toward explaining relative trends in the computed dopant formation energies.

Finally, we note that while the entire possible range of defect/dopant formation energies varies over the range of 0–15 eV, the test set prediction performance or the error on unseen data is of

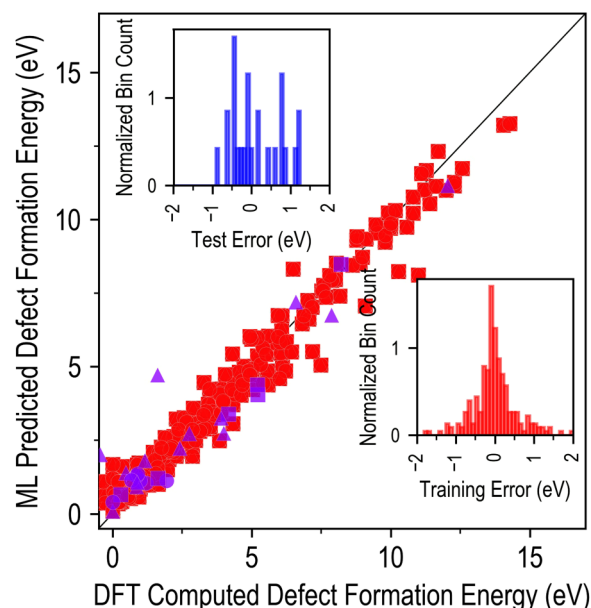


FIG. 6. Parity plot comparing the DFT-computed and ML-predicted defect formation energies for entire dataset that includes *R*-BTO (squares), *c*-BTO (circles), *c*-LMO (triangles) host chemistries. Training (90%) and test (10%) set data points are shown in red and purple, respectively. In the inset, histograms of the test and training set errors are explicitly shown.

the order of <1 eV, indicating that the trained model can allow for a down-selection of relevant substitutional dopants for a given host with an accuracy falling within this energy range. This down-selected subset might have a chemically diverse set of dopant chemistries and the impact of the doping on a property would eventually be governed by both final selection of the substitutional dopant chemistry and its doping concentration. In this sense, the true utility of the developed model lies in its use as a screening filter for a subsequent in-depth domain-knowledge-based analysis. Furthermore, the developed model and the adopted informatics-based framework can also be considered as an extension of classical Hume–Rothery rules for substitutional solid solutions in metals and alloys to more complex oxide chemistries, providing (semi-) quantitative trends in relative substitutional defect formation energetics and also highlighting various physically relevant features dictating these trends.

IV. CONCLUSIONS

Combining high-throughput computation and ML techniques, we developed a thorough and comprehensive understanding of the behavior of dopants within the perovskites lattices and identified the governing factors that control the dopant formation energies in ABO₃ perovskite-type materials. We considered *R*-BTO, *c*-BTO, and *c*-LMO as host perovskites. Across the periodic table, 44 dopants were systematically substituted in either the A- or the B-site in the host lattices. Using ML techniques, we show that the

most important factors that govern the dopant stability in ABO₃ perovskite-type materials are (i) dopant's ionic size, (ii) heat of formation, (iii) effective tolerance factor, and (iv) the oxidation state of the dopant. We also show that our framework for determining defect formation energy, and hence, the feasibility of doping with a particular element is general. In order to do so, our ML model was developed by initially fitting to the training data comprised of the defect formation energies for *R*-BTO. We then used this model to predict the stability of the dopants in *c*-BTO and *c*-LMO successfully. Hence, the present work not only showcases a promising route toward dopant selection in ABO₃ perovskites but also establishes a framework that is general and can be potentially applied to other material classes.

SUPPLEMENTARY MATERIAL

See the [Supplementary material](#) for details on (1) the VASP pseudopotentials used in the DFT computations for the substitutional defect formation energies and (2) the employed feature set. The entire feature–property dataset employed to develop the ML model in this study is also provided.

ACKNOWLEDGMENTS

V.S. acknowledges the XSEDE computational resource allocation (Grant No. TG-DMR200008) and the Advanced Computer Facility (ACF) of the University of Tennessee for computational resources. P.D. and P.K. acknowledge the computational support provided by XSEDE under Project No. PHY180014. P.D. also thanks XSEDE Extended Collaborative Support Service (ECSS) program (ECSS3-5672). P.K. was supported by the National Science Foundation (Grant No. ECCS-1831954). G.P. would like to acknowledge support from the Los Alamos National Laboratory's Laboratory Directed Research and Development (LDRD) program's Directed Research (DR) (Project No. 20190043DR) and computational support from Los Alamos National Laboratory's high performance computing clusters. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

The authors declare that they have no competing interests.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

REFERENCES

- ¹J. S. Capurro, A. B. Alles, and W. A. Schulze, "Processing of laminated barium titanate structures for stress-sensing applications," *J. Am. Ceram. Soc.* **78**(9), 2476–2480 (1995).
- ²G. H. Haertling, "Ferroelectric ceramics: History and technology," *J. Am. Ceram. Soc.* **82**(4), 797–818 (1999).
- ³X. Ren, "Large electric-field-induced strain in ferroelectric crystals by point-defect-mediated reversible domain switching," *Nat. Mater.* **3**(2), 91–94 (2004).
- ⁴J. Zhu, H. Li, L. Zhong, P. Xiao, X. Xu, X. Yang, Z. Zhao, and J. Li, "Perovskite oxides: Preparation, characterizations, and applications in heterogeneous catalysis," *ACS Catal.* **4**(9), 2917–2940 (2014).
- ⁵S. Samira, X.-K. Gu, and E. Nikolla, "Design strategies for efficient nonstoichiometric mixed metal oxide electrocatalysts: Correlating measurable oxide properties to electrocatalytic performance," *ACS Catal.* **9**(11), 10575–10586 (2019).
- ⁶V. Sharma, M. K. Mahapatra, P. Singh, and R. Ramprasad, "Cationic surface segregation in doped LaMnO₃," *J. Mater. Sci.* **50**(8), 3051–3056 (2015).
- ⁷V. Sharma, M. K. Mahapatra, S. Krishnan, Z. Thatcher, B. D. Huey, P. Singh, and R. Ramprasad, "Effects of moisture on (La, A)MnO₃ (A = Ca, Sr, and Ba) solid oxide fuel cell cathodes: A first-principles and experimental study," *J. Mater. Chem. A* **4**, 5605–5615 (2016).
- ⁸V. Sharma, G. Pilania, G. A. Rossetti, K. Slenes, and R. Ramprasad, "Comprehensive examination of dopants and defects in BaTiO₃ from first principles," *Phys. Rev. B* **87**, 134109 (2013).
- ⁹S. Krishnan, V. Sharma, P. Singh, and R. Ramprasad, "Dopants in lanthanum manganite: Insights from first-principles chemical space exploration," *J. Phys. Chem. C* **120**(39), 22126–22133 (2016).
- ¹⁰S. Upadhyay, J. Shrivastava, A. Solanki, S. Choudhary, V. Sharma, P. Kumar, N. Singh, V. R. Satsangi, R. Shrivastav, U. V. Waghmare, and S. Dass, "Enhanced photoelectrochemical response of BaTiO₃ with Fe doping: Experiments and first-principles analysis," *J. Phys. Chem. C* **115**(49), 24373–24380 (2011).
- ¹¹C. Metzmacher and K. Albertsen, "Microstructural investigations of barium titanate-based material for base metal electrode ceramic multilayer capacitor," *J. Am. Ceram. Soc.* **84**(4), 821–826 (2001).
- ¹²A. Honda, S. Higai, Y. Motoyoshi, N. Wada, and H. Takagi, "Theoretical study on interactions between oxygen vacancy and doped rare-earth elements in barium titanate," *Jpn. J. Appl. Phys.* **50**(9), 09NE01 (2011).
- ¹³A. D. Caviglia, S. Gariglio, N. Reyren, D. Jaccard, T. Schneider, M. Gabay, S. Thiel, G. Hammerl, J. Mannhart, and J. M. Triscone, "Electric field control of the LaAlO₃/SrTiO₃ interface ground state," *Nature* **456**(7222), 624–627 (2008).
- ¹⁴H. Ohta, S. W. Kim, Y. Mune, T. Mizoguchi, K. Nomura, S. Ohta, T. Nomura, Y. Nakanishi, Y. Ikuhara, M. Hirano, H. Hosono, and K. Koumoto, "Giant thermoelectric Seebeck coefficient of a two-dimensional electron gas in SrTiO₃," *Nat. Mater.* **6**(2), 129–134 (2007).
- ¹⁵J.-H. Jeon, "Effect of SrTiO₃ concentration and sintering temperature on microstructure and dielectric constant of Ba_{1-x}Sr_xTiO₃," *J. Eur. Ceram. Soc.* **24**(6), 1045–1048 (2004).
- ¹⁶H.-J. Hagemann and D. Hennings, "Reversible weight change of acceptor-doped BaTiO₃," *J. Am. Ceram. Soc.* **64**(10), 590–594 (1981).
- ¹⁷J. Suntivich, K. J. May, H. A. Gasteiger, J. B. Goodenough, and Y. Shao-Horn, "A perovskite oxide optimized for oxygen evolution catalysis from molecular orbital principles," *Science* **334**(6061), 1383–1385 (2011).
- ¹⁸J. Suntivich, H. A. Gasteiger, N. Yabuuchi, H. Nakanishi, J. B. Goodenough, and Y. Shao-Horn, "Design principles for oxygen-reduction activity on perovskite oxide catalysts for fuel cells and metal–air batteries," *Nat. Chem.* **3**(7), 546–550 (2011).
- ¹⁹L. P. Putilov and V. I. Tsidilkovski, "The role of deep acceptor centers in the oxidation of acceptor-doped wide-band-gap perovskites ABO₃," *J. Solid State Chem.* **247**, 147–155 (2017).
- ²⁰H. Mo, H. Nan, X. Lang, S. Liu, L. Qiao, X. Hu, and H. Tian, "Influence of calcium doping on performance of LaMnO₃ supercapacitors," *Ceram. Int.* **44**(8), 9733–9741 (2018).
- ²¹J. Varignon, M. Bibes, and A. Zunger, "Mott gapping in 3d ABO₃ perovskites without Mott–Hubbard interelectronic repulsion energy *U*," *Phys. Rev. B* **100**, 035119 (2019).
- ²²L. Wang, T. Ma, S. Dai, T. Ren, Z. Chang, L. Dou, M. Fu, and X. Li, "Experimental study on the high performance of Zr doped LaCoO₃ for solar thermochemical CO production," *Chem. Eng. J.* **389**, 124426 (2020).
- ²³Y. Sim, D. Kwon, S. An, J.-M. Ha, T.-S. Oh, and J. Chul Jung, "Catalytic behavior of ABO₃ perovskites in the oxidative coupling of methane," *Mol. Catal.* **489**, 110925 (2020).

- ²⁴D. Triyono, B. Betria, and H. Laysandra, "Effect of Fe doping on the electrical properties of BaTiO₃ crystalline materials at room temperature," *J. Phys. Conf. Ser.* **1442**, 12006 (2020).
- ²⁵H. Ihrig, "PTC effect in BaTiO₃ as a function of doping with 3d elements," *J. Am. Ceram. Soc.* **64**(10), 617–620 (1981).
- ²⁶H. Ihrig, "The phase stability of BaTiO₃ as a function of doped 3d elements: An experimental study," *J. Phys. C Solid State Phys.* **11**(4), 819–827 (1978).
- ²⁷J. R. Sambrano, E. Orhan, M. F. C. Gurgel, A. B. Campos, M. S. Góes, C. O. Paiva-Santos, J. A. Varela, and E. Longo, "Theoretical analysis of the structural deformation in Mn-doped BaTiO₃," *Chem. Phys. Lett.* **402**(4), 491–496 (2005).
- ²⁸M. T. Buscaglia, V. Buscaglia, M. Viviani, and P. Nanni, "Atomistic simulation of dopant incorporation in barium titanate," *J. Am. Ceram. Soc.* **84**(2), 376–84 (2001).
- ²⁹H. Nakayama and H. Katayama-Yoshida, "Theoretical prediction of magnetic properties of Ba(Ti_{1-x}M_x)O₃ (M=Sc,V,Cr,Mn,Fe,Co,Ni,Cu)," *Jpn. J. Appl. Phys.* **40**(Part 2, No. 12B), L1355–L1358 (2001).
- ³⁰G. V. Lewis and C. R. A. Catlow, "Defect studies of doped and undoped barium titanate using computer simulation techniques," *J. Phys. Chem. Solids* **47**(1), 89–97 (1986).
- ³¹M. J. Akhtar, Z.-U.-N. Akhtar, R. A. Jackson, and C. R. A. Catlow, "Computer simulation studies of strontium titanate," *J. Am. Ceram. Soc.* **78**(2), 421–428 (1995).
- ³²D. M. Smyth, "The defect chemistry of donor-doped BaTiO₃: A rebuttal," *J. Electroceram.* **9**(3), 179–186 (2002).
- ³³W. Hume-Rothery and G. V. Raynor, "The equilibrium and lattice-spacing relations in the system magnesium-cadmium," *Proc. R. Soc. Lond. A Math. Phys. Sci.* **174**(959), 471–486 (1940).
- ³⁴R. Batra, G. Pilania, B. P. Uberuaga, and R. Ramprasad, "Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia," *ACS Appl. Mater. Interfaces* **11**(28), 24906–24918 (2019), PMID: 30990303.
- ³⁵A. Mannodi-Kanakkithodi, M. Y. Toriyama, F. G. Sen, M. J. Davis, R. F. Klie, and M. K. Y. Chan, "Machine-learned impurity level prediction for semiconductors: The example of Cd-based chalcogenides," *npj Comput. Mater.* **6**(1), 39 (2020).
- ³⁶G. Kresse and J. Hafner, "Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium," *Phys. Rev. B* **49**, 14251–14269 (1994).
- ³⁷G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Comput. Mater. Sci.* **6**(1), 15–50 (1996).
- ³⁸G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Phys. Rev. B* **54**, 11169–11186 (1996).
- ³⁹J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
- ⁴⁰H. J. Monkhorst and J. D. Pack, "Special points for brillouin-zone integrations," *Phys. Rev. B* **13**, 5188–5192 (1976).
- ⁴¹C. Franchini, R. Podloucky, J. Paier, M. Marsman, and G. Kresse, "Ground-state properties of multivalent manganese oxides: Density functional and hybrid density functional calculations," *Phys. Rev. B* **75**, 195128 (2007).
- ⁴²A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, "A high-throughput infrastructure for density functional theory calculations," *Comput. Mater. Sci.* **50**(8), 2295–2310 (2011).
- ⁴³L. Wang, T. Maxisch, and G. Ceder, "Oxidation energies of transition metal oxides within the GGA + U framework," *Phys. Rev. B* **73**, 195107 (2006).
- ⁴⁴J. P. Perdew, "Density functional theory and the band gap problem," *Int. J. Quantum. Chem.* **28**(S19), 497–523 (1985).
- ⁴⁵J. P. Perdew, W. Yang, K. Burke, Z. Yang, E. K. U. Gross, M. Scheffler, G. E. Scuseria, T. M. Henderson, I. Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, and A. Görling, "Understanding band gaps of solids in generalized Kohn–Sham theory," *Proc. Natl. Acad. Sci. U.S.A.* **114**(11), 2801–2806 (2017).
- ⁴⁶V. Sharma, A. McDannald, M. Staruch, R. Ramprasad, and M. Jain, "Dopant-mediated structural and magnetic properties of TbMnO₃," *Appl. Phys. Lett.* **107**(1), 012901 (2015).
- ⁴⁷M. Staruch, V. Sharma, C. dela Cruz, R. Ramprasad, and M. Jain, "Magnetic ordering in TbMn_{0.5}Cr_{0.5}O₃ studied by neutron diffraction and first-principles calculations," *J. Appl. Phys.* **116**(3), 033919 (2014).
- ⁴⁸S. Curtarolo, D. Morgan, and G. Ceder, "Accuracy of ab initio methods in predicting the crystal structures of metals: A review of 80 binary alloys," *Calphad* **29**(3), 163–211 (2005).
- ⁴⁹S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," *Comput. Mater. Sci.* **68**, 314–319 (2013).
- ⁵⁰R. D. Shannon, "Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides," *Acta Crystallogr.* **A32**(3), 751–767 (1976).
- ⁵¹L. Mentel, Mendeleev—A python resource for properties of chemical elements, ions and isotopes, 2018.
- ⁵²L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).
- ⁵³J. R. Quinlan, "Induction of decision trees," *Mach. Learn.* **1**(1), 81–106 (1986).
- ⁵⁴J. R. Quinlan, "Simplifying decision trees," *Int. J. Man Mach. Stud.* **27**(3), 221–234 (1987).
- ⁵⁵C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications* (Springer, 2012).
- ⁵⁶A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News* **2**(3), 18–22 (2002).
- ⁵⁷F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- ⁵⁸S. Das and T. K. Dey, "Magnetic entropy change in polycrystalline La_{1-x}K_xMnO₃ perovskites," *J. Alloys Compd.* **440**(1), 30–35 (2007).
- ⁵⁹G. Huo, Z. Gu, S. Liu, Y. Wang, and Y. Wang, "Structure and magnetic properties of substitution of K for La in LaMnO₃ perovskite," *J. Alloys Compd.* **433**(1), 41–45 (2007).
- ⁶⁰D. Wu, B. Fang, Q. Du, and J. Ding, "Preparation and properties of La and K co-doped BaTiO₃ lead-free piezoelectric ceramics," *Ferroelectrics* **432**(1), 81–91 (2012).
- ⁶¹M. H. Khedhri, N. Abdelmoula, H. Khemakhem, R. Douali, and F. Dubois, "Structural, spectroscopic and dielectric properties of Ca-doped BaTiO₃," *Appl. Phys. A* **125**(3), 193 (2019).
- ⁶²L. Wang, M. Al-Mamun, Y. L. Zhong, L. Jiang, P. Liu, Y. Wang, H. G. Yang, and H. Zhao, "Ca²⁺ and Ga³⁺ doped LaMnO₃ perovskite as a highly efficient and stable catalyst for two-step thermochemical water splitting," *Sustain. Energy Fuels* **1**, 1013–1017 (2017).
- ⁶³A. Kumari and B. D. Ghosh, "Effect of strontium doping on structural and dielectric behaviour of barium titanate nanoceramics," *Adv. Appl. Ceram.* **117**(7), 427–435 (2018).
- ⁶⁴H. Liu, B. Cao, and C. O'Connor, "Intrinsic magnetism in BaTiO₃ with magnetic transition element dopants (Co, Cr, Fe) synthesized by sol-precipitation method," *J. Appl. Phys.* **109**(7), 07B516 (2011).
- ⁶⁵M. T. Buscaglia, V. Buscaglia, M. Viviani, P. Nanni, and M. Hanuskova, "Influence of foreign ions on the crystal structure of BaTiO₃," *J. Eur. Ceram. Soc.* **20**(12), 1997–2007 (2000).
- ⁶⁶N.-H. Chan and D. M. Smyth, "Defect chemistry of donor-doped BaTiO₃," *J. Am. Ceram. Soc.* **67**(4), 285–288 (1984).
- ⁶⁷R. N. Schwartz, B. A. Wechsler, and L. West, "Spectroscopic and photorefractive properties of molybdenum-doped barium titanate," *Appl. Phys. Lett.* **67**(10), 1352–1354 (1995).
- ⁶⁸L. Liu, H. Guo, H. Lü, S. Dai, B. Cheng, and Z. Chen, "Effects of donor concentration on the electrical properties of Nb-doped BaTiO₃ thin films," *J. Appl. Phys.* **97**(5), 054102 (2005).
- ⁶⁹V. Tura and L. Mitoseriu, "Ageing of low field dielectric constant and losses in (Hf, Zr)-doped BaTiO₃ ceramics," *Europhys. Lett.* **50**(6), 810–815 (2000).

- ⁷⁰G. L. Catchen, W. E. Evenson, and D. Allred, "Structural phase transition and T_c distribution in Hf-doped LaMnO_3 investigated using perturbed-angular-correlation spectroscopy," *Phys. Rev. B* **54**, R3679–R3682 (1996).
- ⁷¹A. Kowalczyk, J. Baszynski, A. Szajek, A. Slebarski, and T. Tolinski, "Electronic structure of doped LaMnO_3 perovskite studied by x-ray photoemission spectroscopy," *J. Phys. Condens. Matter* **13**(23), 5519–5525 (2001).
- ⁷²M. C. Chang, and S.-C. Yu, *The Electronic States in Ca-doped BaTiO_3 Ceramics*, Advances in Quantum Chemistry (Academic Press, 2000), Vol. 37, pp. 179–191.
- ⁷³J. Zhi, A. Chen, Y. Zhi, P. M. Vilarinho, and J. L. Baptista, "Incorporation of yttrium in barium titanate ceramics," *J. Am. Ceram. Soc.* **82**(5), 1345–1348 (1999).
- ⁷⁴J. Qi, Z. Gui, W. Li, Y. Wang, Y. Wu, and L. Li, "Temperature stable $\text{Ba}_{1-x}\text{Cd}_x\text{TiO}_3$ dielectrics," *Mater. Lett.* **56**(4), 507–511 (2002).
- ⁷⁵T. Takeda and A. Watanabe, "Substituting sites of rare earth ions in BaTiO_3 ," *Jpn. J. Appl. Phys.* **7**(3), 232–235 (1968).
- ⁷⁶L. A. Xue, Y. Chen, and R. J. Brook, "The influence of ionic radii on the incorporation of trivalent dopants into BaTiO_3 ," *Mater. Sci. Eng. B* **1**(2), 193–201 (1988).