*MRS Communications (2019)*, *9*, 882–890
© The Author(s) 2019
doi:10.1557/mrc.2019.103

MRS Communications

Artificial Intelligence Research Letter

# Prediction of new iodine-containing apatites using machine learning and density functional theory

**Timothy Q. Hartnett*** and **Mukil V. Ayyasamy***, Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA 22904, USA
**Prasanna V. Balachandran**, Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA 22904, USA; Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA 22904, USA

Address all correspondence to Prasanna V. Balachandran at **pvb5e@virginia.edu**

(Received 20 February 2019; accepted 26 July 2019)

## Abstract

The authors develop a computational approach that integrates machine learning (ML) and density functional theory (DFT) with experimental data to predict formable and thermodynamically stable iodine-containing apatites. This is an important problem because radioactive iodine is toxic and capturing it in solid waste forms have implications in remediation treatments. The authors train ML models using 336 compositions and screen 54 iodine-containing compounds in apatite stoichiometry. ML models predict 18 as formable and 24 as nonformable in the apatite structure; 12 compounds were identified to be uncertain. DFT convex hull predicted two to be thermodynamically stable, one as metastable, and nine as unstable.

## Introduction

Long-term storage and safe disposal of toxic wastes (nuclear, radioactive, or otherwise) is a pressing environmental concern that must be addressed for a sustainable future. Glasses and crystalline ceramic materials represent some of the common inorganic host materials that have been explored and have shown promise for the immobilization of waste materials.[1,2] One of the waste forms that is generated from nuclear power plants is the toxic and radioactive iodine (I-129)[3], and there is a growing priority to capture radioactive iodine, because unattended I-129 entering the ecosystem is a severe concern due to its long half-life, high solubility in water, and potential human ingestion.[4] More recently, materials that form an "apatite" crystal structure type have been considered as a potential host lattice for immobilizing the I-129 isotope.[3,5]

Compounds with the apatite crystal structure type have a general formula $A_{10}(BO_4)_6X_2$, where the A-site is occupied by larger monovalent ($Na^+$, $K^+$, etc.), divalent ($Ca^{2+}$, $Sr^{2+}$, $Ba^{2+}$, $Pb^{2+}$, etc.), or trivalent ($La^{3+}$, $Y^{3+}$, $Ce^{3+}$, $Sm^{3+}$, etc.) cations, the B-site is occupied by smaller metals and metalloids ($P^{5+}$, $As^{5+}$, $V^{5+}$, $Si^{4+}$, etc.), and the X-site is filled by halide, hydroxide, or oxide anions ($F^-$, $Cl^-$, $Br^-$, $I^-$, $OH^-$, $O^{2-}$, etc.). The aristotype or the parent crystal structure of an apatite has a hexagonal symmetry in the space group, $P6_3/m$ (# 176), with a one-dimensional channel parallel to the *c*-axis. Crystallographically, there are two distinct A-sites, $A^I$ and $A^{II}$, occupying the Wyckoff sites 4*f* and 6*h*, respectively. The crystal structure is made up of a network of regular $BO_4$ tetrahedra that is connected to a $A^IO_6$ metaprism and an irregular $A^{II}O_6X_{1,2}$ polyhedral unit. The crystal structure is shown in Fig. 1. Apatites also have been shown to undergo structural-phase transformation to lower symmetry structures as a function of chemical substitutions, defects (vacancies and interstitials), and temperature.[6,7] Several desirable properties such as structural flexibility and chemical diversity, durability against radiation,[8] and accessibility to straightforward synthesis techniques justify the choice of the material.[6,7,9,10]

Despite the progress in improving the properties of iodine-containing apatites (iodoapatites) such as the loading capacity, leaching rate, thermal recovery, and mechanical properties via processing routes,[11,12] most of the experimental studies have focused only on nonstoichiometric forms of lead vanadate [$Pb_{10}(VO_4)_6I_2$] iodoapatites. The quest for iodoapatites with better properties is mainly limited by the lack of exploration. With potentially many choices for chemical substitutions in both A- and B-sites in the apatite structure, experimental investigation for all possible combinations is not realistic. *A priori* knowledge of the compositional limit of a particular chemical constituent in the apatite lattice can provide important insights toward tailoring new compounds with desired properties.

Here, we develop a computational approach built on the foundations of machine learning (ML) and density functional theory (DFT) methods to predict possible new iodoapatites. The idea of combining ML methods with DFT calculations (including the high-throughput variant) to predict the

* These two authors contributed equally to this work.


https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

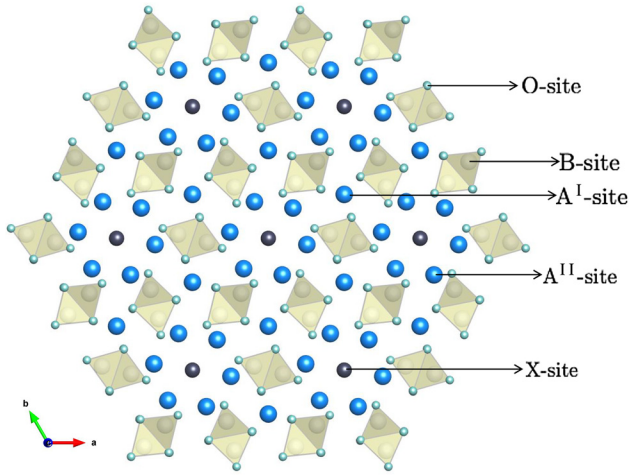



**Figure 1.** Crystal structures of $A_{10}(BO_4)_6X_2$ apatites in the aristotype $P6_3/m$ space group projected onto the (0001) plane. The structure is made of $BO_4$ tetrahedral units that are connected to a sixfold coordinated $A^IO_6$ metaprism unit. In addition, an irregular $A^{II}O_6X_{1,2}$ polyhedral unit circumscribes the one-dimensional channel parallel to the *c*-axis.

formability and thermodynamic stability is not entirely new. There is a growing body of literature in exploring the potential of ML methods for predicting compound formability. Some of the notable contributions include the prediction of formability of perovskite-based compounds, Heusler alloys, high entropy alloys, and metallic glasses (to name a few).[13–21]

In terms of prior ML studies on iodoapatites, Wang [22] built an artificial neural network model to predict the channel size of apatite structure as a function of iodine incorporation. The hypothesis was that large channels host large anions (such as $I^-$) and small channels host small anions (such as $F^-$). As a result, the goal was to identify A- and B-site elements that will lead to large channels so that iodine can be incorporated. Wang neither considered the criterion of formability nor thermodynamic stability in the published work. One of the plausible reasons could be because of the lack of databases in the literature where such formability data are compiled. We bridge the gap in this paper by constructing a database from surveying the experimental literature and demonstrate the utility of ML methods by predicting novel and formable iodoapatite compounds.

In this work, we employed ensemble ML models that were based on tree-based methods (random forests and gradient tree boosting) to establish a relationship between the crystal chemistry descriptors and apatite-phase formability. Ensemble models combine predictions from 100 or 1000 s (or more) of ML models and are shown to be more accurate than predictions from a single model.[23] Further, the apatite phase formability problem is compounded by the small nature of the training dataset sizes (336 data points). One of the challenges in the application of ML methods on a small dataset is its high variability. Ensemble methods provide a means to address the bias–variance tradeoff by reducing the effect of the variance term. In addition to ML methods, we also employed DFT calculations to calculate the thermodynamic stability of apatite compounds. We utilized the convex hull (CH) construction as implemented in the open quantum materials database (OQMD)[24,25] and calculated the decomposition energy ($\Delta E^D$) associated with each apatite compound. The $\Delta E^D$ cutoff was determined from comparing the DFT data with experimental XRD (x-ray diffraction)-phase formation data. Finally, we used our trained ML models to predict the formability of previously unexplored iodoapatites. These results were augmented with thermodynamic stability predictions from DFT, especially for those compounds where the two ML models disagreed. The key outcome is the prediction of 18 novel iodoapatites as formable, two as thermodynamically stable, and one as metastable. We recommended these promising predictions for experimental synthesis and validation.

## Methods and approach

### *Machine learning*

#### *Dataset*

The first step in the ML approach is to develop a representative dataset of known observations. For developing the classification models, we have considered a total of 336 compositions taken from the literature.[7,26–33] In Figure 2, a schematic overview of our dataset is given. The chemical compositions that contain fluoride, chloride, oxide, and hydroxide anion in the X-site are well represented in our dataset, whereas those containing bromide and iodide are relatively sparse. For the given chemical composition, we distinguished between stable apatites and nonapatites based on the peaks assigned to the apatite structure from the XRD pattern. For example, in our database, we assigned stable apatite structure to $Ca_{10}(PO_4)_6F_2$, whereas for $Zn_{10}(PO_4)_6F_2$ we assigned it to be a nonapatite, because the synthesis of fully stoichiometric $Zn_{10}(PO_4)_6F_2$ apatite was not shown to be experimentally feasible.[28] In addition, when an experimental synthesis resulted in the formation of secondary phases, we assigned it to be nonapatites. Instead of using the entire XRD pattern in our analysis, we defined a single categorical variable called "apatite formability" and assigned a categorical label "Yes" to the chemical compositions that are experimentally determined to have a single-phase apatite structure and "No" to the compositions that do not form single-phase apatite structure. Of the 336 compositions in our database, 224 had formable apatite structure and 112 were nonapatites. It should also be noted that in our database, we do not explicitly differentiate between the different crystal symmetries of apatites. Apatites are known to exist in $P6_3/m$, $P6_3/cm$, $P6_3$, $P\bar{3}$, $P2_1/b$, $P2_1$, $P\bar{6}$, and $P\bar{1}$[7,34] space groups, and the label "Yes" can correspond to any of these space groups. The dataset is given in Supplementary Material.

Past experimental and theoretical studies have indicated the importance of three descriptors: ionic radii, electronegativity, and polarizing power, in describing the apatite-phase formation.[35–39] While ionic radii can be related to the lattice energy



https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

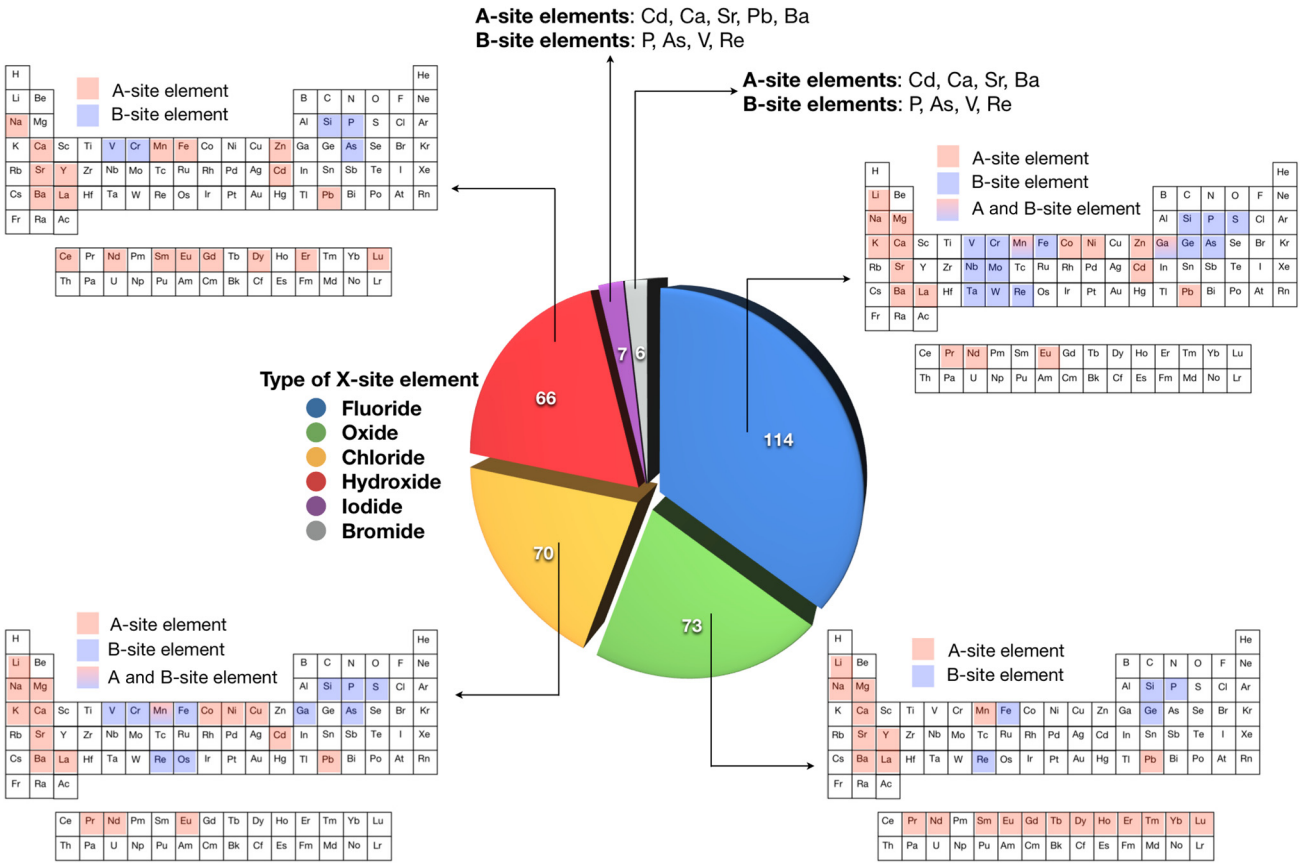



**Figure 2.** Schematic overview of various A-site, B-site, and X-site elements considered in this study. The choice of these chemical constituents was based on the available data from the past experimental studies. The numbers in the pie chart indicate the total count of chemical compositions that contains the specific X-site element. For example, in our data set, there were 114 chemical compositions that had fluoride anion in the X-site. Overall, we considered a total of 336 chemical compositions, out of which 224 had state apatite structure and 112 were nonapatites. In addition, the figure also describes the specific combination of various A-site and B-site elements used for a particular X-site element.

of solids,[40] the physical meaning of electronegativity and polarizing power (a relative measure of the power of a cation to attract an anion) [41] takes into account the bond directionality and charge density information. The polarizing power of a cation is defined as the ratio of the charge of the ion to its ionic radius. Therefore, by combining the three parameters, we hypothesize that the electronic structure and bonding of complex apatites can be well approximated. The full list of variables used in this work to quantify each stoichiometry is given in Table I. The descriptors of Shannon' ionic radii,[42] Pauling's electronegativity[43], and polarizing power of the cation were used as the predictor (or independent) variables, and the phase formability information ("apatite formability" variable) extracted from XRD analysis was used as the dependent variable. We used the group electronegativity value of 3.7 for the OH anion that was scaled to be compatible with the Pauling scale.[44] In Supplementary Fig. 1, the numerical values of the predictor variables used in the analysis are provided. Using these values, the average ionic radii, average electronegativity, and average polarizing power of the cation at each crystallographic site were calculated by multiplying each variable by its weight, adding these values and dividing the total by the total weight of A-, B-, and X-sites separately. For example, the average ionic radii of the A-site ($\langle r_A \rangle$) in $A^1_pA^2_q(BO_4)_6X_2$ is calculated using the following equation:

$$\langle rA \rangle = \frac{(r_{A^1} \times p) + (r_{A^2} \times q)}{(p+q)} \quad (1)$$

where, $r_{A^1}$ and $r_{A^2}$ are the ionic radii of $A^1$ and $A^2$ chemical elements that occupy the A-site of the lattice, respectively, and $p$ and $q$ are the relative weights of $A^1$ and $A^2$ chemistries, respectively.

The logic explained above in Eq. (1) was repeated in calculating the average values of all the predictor variables listed in Table I. The 13 crystal chemistry descriptors in Table I served as the input for ML. Since the response (or output) is a categorical, binary variable (Yes or No), the type of ML analysis is referred to as classification learning. The goal of ML is to establish a relationship between the response and the inputs. We


https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

**Table I.** List of descriptors used for training the model.

| Abbreviation | Description |
|---|---|
| $r_A$ | Average ionic radius of A-site element |
| $r_B$ | Average ionic radius of B-site element |
| $r_X$ | Average ionic radius of X-site element |
| $r_A/r_B$ | Ratio of average ionic radius of A-site to B-site element |
| $r_A/r_X$ | Ratio of average ionic radius of A-site to X-site element |
| $EN_A$ | Average electronegativity of A-site element |
| $EN_B$ | Average electronegativity of B-site element |
| $EN_X$ | Average electronegativity of X-site element |
| $EN_A - EN_B$ | Difference between average electronegativity of |
| | A-site and B-site element (absolute value) |
| $EN_A - EN_X$ | Difference between average electronegativity of |
| | A-site and X-site element (absolute value) |
| $PP_A$ | Average polarizing power of A-site element |
| $PP_B$ | Average polarizing power of B-site element |
| $PP_A/PP_B$ | Ratio of average polarizing power of A-site to B-site element |

All descriptors must be interpreted in terms of their average values as calculated based on Eq. (1).

utilized random forests and gradient tree boosting for classification learning, which are discussed next.

### Random forests

Random forest is an ensemble learning method that generates a specified number of unbiased decision trees then aggregates the decisions of all trees to make a final prediction of the class label.[45,46] These trees are generated using a bootstrap method where the training data are sampled, with replacement, to grow the forest. Most random forest algorithms typically withhold 30% of the training data during this process, although the number may vary slightly depending on the algorithm. These withheld data points are referred to as the out-of-bag (OOB) samples and are used to assess the accuracy of the trained model. This is a straight forward process since the true labels of the OOB samples are contained in the training set.

The algorithm works by first sampling a pre-specified number of bootstrapped samples (*ntree*), and larger values of *ntree* generally increase the accuracy of the model but are more complex to train and often do not increase accuracy by a substantial amount beyond a certain point. The algorithm then randomly samples, with replacement, a pre-specified number of input parameters (*mtry*) and generates a decision tree for each bootstrapped sample based on those parameters. We utilized the random forest method as implemented in the RANDOMFOREST **R**-package.[46]

### Gradient tree boosting

Gradient tree boosting is another ensemble learning method that uses decision trees as base learners.[47,48] In contrast to bagging algorithm of the random forests, the boosting algorithm learns each decision tree sequentially. The first tree is used to coarsely segment the observations in the data. Then, the subsequent trees were fit in such a way that the error of the current model was further minimized. The loss function used in the model is a binomial log-likelihood function. The final prediction of the class is estimated as the sum of predictions of the individual trees. There are three tuning parameters: (i) the number of trees, (ii) the maximum number of splits in each tree, and (iii) the learning rate of the gradient of loss function. We fixed the maximum number of splits in each tree and the learning rate hyperparameters to the default values. We tuned the number of trees using the *k*-fold cross-validation method. We utilized the gradient tree boosting method as implemented in the GBM **R**-package.[49]

## Density functional theory

DFT calculations were performed within the generalized gradient approximation (GGA) as implemented in QUANTUM ESPRESSO.[50] The PBEsol exchange-correlation functional[51] was used, and the core and valence electrons were treated with ultrasoft pseudopotentials. The Brillouin zone integration was performed using a Monkhorst-Pack *k*-point mesh centered at Γ and a 60-Ry plane-wave cutoff for wavefunctions (600 Ry kinetic energy cutoff for charge density and potential). The scalar relativistic pseudopotentials were taken from the PSLIBRARY.[52] The atomic positions and the cell volume were allowed to relax until an energy convergence threshold of $10^{-8}$ eV and Hellmann–Feynman forces <2 meV/Å, respectively, were achieved. We considered the following space groups to determine the lowest energy apatite structure: (i) in the case of fluorapatites (X=F), hexagonal ($P6_3/m$), and triclinic ($P\bar{1}$) space groups, (ii) in the case of chlorapatites (X=Cl), bromapatites (X=Br), and iodoapatites (X=I), hexagonal ($P6_3/m$), monoclinic ($P2_1/b$), and triclinic ($P\bar{1}$) space groups, and (iii) in the case of hydroxyapatites (X=OH), monoclinic ($P2_1/b$ and $P2_1$) and hexagonal ($P6_3$) crystal symmetries. The choice of these symmetries was motivated by the earlier work in the literature.[7,53] We did not consider substitutional defects, vacancies, or interstitials in our calculations.

To evaluate the thermodynamic stability and decomposition reaction of each apatite compound, we considered the large pool of compounds assembled in the OQMD.[24,25] To construct the CH for a given A–B–O–X element set, the OQMD uses the total energy data computed by DFT of all (both existing and hypothetical) elements and compounds in the OQMD. Due to computational challenges involved in treating solid solutions, DFT calculations were only run on a subset of the



https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

total training data that represented simple compounds (Supplementary Table 1). For simplicity, all magnetic materials and those containing partially filled *d*-orbitals in the divalent ($Fe^{2+}$, $Mn^{2+}$, and $Ni^{2+}$) or pentavalent charge states ($Fe^{5+}$, $Mn^{5+}$, $Cr^{5+}$, $Re^{5+}$, etc.) were not considered in this analysis. In total, the $\Delta E^{D}$ values were computed for 38 apatite compounds in the training set. We extracted the reactants from the OQMD and performed spot DFT calculations. Our product is the apatite phase in its lowest energy structure. We then calculated the distance from the CH as $\Delta E^{D} = E^{\text{product}} - \sum_{i=1}^{n} E_i^{\text{reactants}}$ (where $i = 1, \ldots, n$ represent the distinct reacting species and E is the total energy from DFT calculations), which can take either a positive [the hypothetical $A_{10}(BO_4)_6X_2$ compound is metastable or unstable] or a negative value [the hypothetical $A_{10}(BO_4)_6X_2$ compound is more stable than other compounds that lie on the CH, and thus, the apatite is more likely to form].

## Results and discussion

The hyperparameters for the random forest models were optimized by iterating over a grid space of *ntree* and *mtry*. We explored a range of 1–2000 trees for *ntree* and 1–13 for *mtry*. The optimal OOB error rate was calculated to be 13.1% for the combined *ntree* and *mtry* values of 955 and 3, respectively. As a result, the final predictions about apatite formability were made by combining predictions from 955 models. Similarly, the hyperparameters for the gradient tree boosting (number of trees) were optimized using a *k*-fold cross-validation, where we explored several values of *k*-fold. The best result was obtained for a *k*-value of 15 (number of trees = 500). In this case, the final predictions were made by combining predictions from 500 models. In order to reliably compare random forest and gradient tree boosting methods, we randomly split our dataset into training and test sets. A 80%–20% training set–test set split was used to build and test the trained ML models, respectively. We also repeated the process five times with five different random seeds, and the average test set error rate from the five random trials was calculated. The random forest and gradient tree boosting models performed with an average test set error rate of 15% and 18%, respectively. These results indicate that the random forest models perform relatively better compared with the gradient tree boosting models on this dataset.

In Supplementary Table 1, the results from DFT-CH analysis are given for 38 compounds for which apatite-phase formation data are known from experimentally measured XRD data. Ideally, we expect the DFT-CH data to consist of two independent probability distributions, one corresponding to experimentally formable and the other to nonformable compounds. The violin plots of this data appear to show that the distributions are independent (Fig. 3). To verify independence statistically, we conducted a nonparametric Mann–Whitney *U*-test (also referred to as the Wilcoxon–Mann–Whitney test or the rank-sum test). This nonparametric test is commonly used to determine the independence of any two distributions,[54] where a P-value of <0.05 generally means that two samples are from

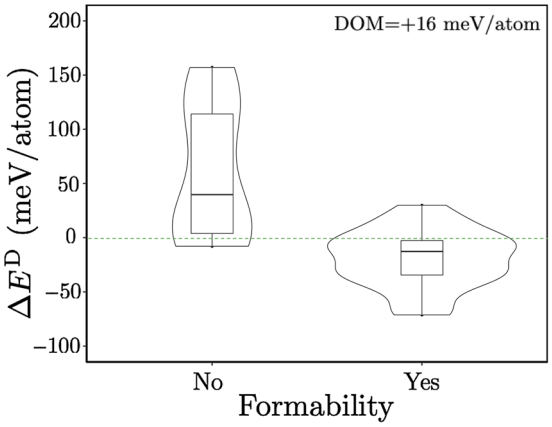



**Figure 3.** Distributions of formable and nonformable compounds based on their DFT-CH predicted $\Delta E^{D}$. One of the nonformable compounds, $Ga_{10}(PO_4)_6F_2$, is identified as an outlier in the violin plot because of an uncharacteristically large $\Delta E^{D}$ in the DFT-CH analysis. The degree of metastability (DOM) is set at +16 meV/atom, and any compound with $\Delta E^{D} \leq$ DOM are predicted to be potentially formable.

distributions with separate median values. The P-value for the test was $2.04 \times 10^{-5}$, confirming that these were indeed independent distributions.

Comparing several different cutoff energies, there is some ambiguity in the area of formability between the values of −10 to +16 meV/atom. General thermodynamical stability intuition would lead us to believe that compounds with a positive $\Delta E^{D}$ would be energetically unfavorable. For this reason, we defined a "degree of metastability (DOM)" and referred to compounds whose $\Delta E^{D}$ falls in this range (≥0 and ≤+16 meV/atom) to be metastable and potentially formable using nonequilibrium synthesis methods. Additional details can be found in Supplementary Section III.

On the basis of the DOM analysis, we recommend re-evaluation of $Pb_{10}(PO_4)_6I_2$ compound. Cao *et al.*[55] used wet chemical synthesis methods and led to the conclusion that the incorporation of $(PO_4)$ in $Pb_{10}(VO_4)_6I_2$ had a compositional limit (4.77 wt%), beyond which the single-phase structure was not experimentally observed. Therefore, we labeled $Pb_{10}(PO_4)_6I_2$ as nonformable in our training set. However, the DFT-CH analysis predicts a $\Delta E^{D}$ value of +5.14 meV/atom (only weakly metastable) for $Pb_{10}(PO_4)_6I_2$. Therefore, nonequilibrium processing methods can potentially stabilize an apatite phase in $Pb_{10}(PO_4)_6I_2$. Further, random forest and gradient tree boosting methods predicted that $Pb_{10}(PO_4)_6I_2$ is nonformable with a probability of 51% and 87%, respectively. Therefore, the random forest models are uncertain about assigning a label for $Pb_{10}(PO_4)_6I_2$, which further indicates that the compound should be re-evaluated. On the other hand, gradient tree boosting is fairly confident (87%) to assign a nonformable label to $Pb_{10}(PO_4)_6I_2$.

Having trained the ML models using formable and nonformable experimental data and having established a


https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

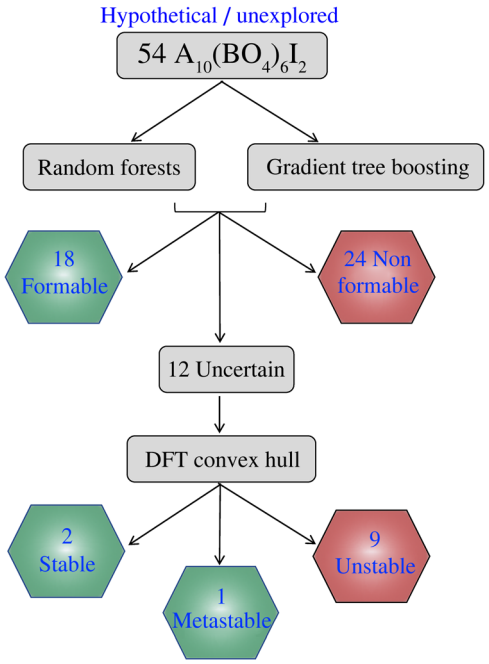



**Figure 4.** The strategy to predict the formability and thermodynamic stability of previously unexplored or hypothetical 54 $A_{10}(BO_4)_6I_2$ using two ML methods (random forest and gradient tree boosting) and DFT-CH thermodynamical stability analysis. The ML models predict 18 as formable and 24 as nonformable. There were 12 compounds for which the outcomes from ML models were uncertain (gradient tree boosting predicted them to be formable, but random forest predicted them to be nonformable). We performed DFT-CH analysis on those 12 uncertain compounds to gain insights into their thermodynamic stability, which predicted two to be stable, one to be metastable, and nine to be unstable.

thermodynamic stability and DOM criteria ($\Delta E^D$) using known data from DFT-CH analysis, we now predict the formability of previously unexplored iodoapatites. Our strategy is shown in Fig. 4. A dataset was constructed by including a list of potential iodoapatites considering possible permutations of several different atoms in the A- and B-sites. We included Ba, Ca, Cd, Ga, Pb, and Sr as possible A-site atoms and As, Mo, Ga, Nb, P, Re, Os, Ta, V, and W as B-site atoms. The combinations of these A- and B-site atoms gave 60 potential iodoapatite stoichiometries, and six have been already experimentally synthesized. Out of the remaining 54 compounds, 42 agreed between random forest and gradient tree boosting, which included predictions of 18 compositions as formable and 24 as nonformable. The list of 18 formable novel apatites is given in Table II. For the 12 compounds where random forest and gradient tree boosting predictions disagreed (i.e., gradient tree boosting predicted them to be formable, whereas random forest predicted them to be nonformable), we performed DFT-CH analysis to provide additional insights into the thermodynamic stability. The results are given in Table III. From the data, it can be seen that the thermodynamic stability analysis inclines toward the random forest predictions.

**Table II.** List of 18 novel and formable iodoapatites predicted by random forest and gradient tree boosting.

| |
|---|
| $Ba_{10}(MoO_4)_6I_2$, $Ba_{10}(AsO_4)_6I_2$ (+21.27), $Ba_{10}(OsO_4)_6I_2$, $Ba_{10}(PO_4)_6I_2$ (+11.70), $Ba_{10}(VO_4)_6I_2$ (+1.77), $Ca_{10}(OsO_4)_6I_2$ |
| $Ca_{10}(ReO_4)_6I_2$, $Ca_{10}(AsO_4)_6I_2$ (−72.5), $Ca_{10}(VO_4)_6I_2$, $Cd_{10}(ReO_4)_6I_2$, $Cd_{10}(OsO_4)_6I_2$, $Pb_{10}(AsO_4)_6I_2$ |
| $Pb_{10}(MoO_4)_6I_2$, $Pb_{10}(VO_4)_6I_2$, $Sr_{10}(MoO_4)_6I_2$, $Sr_{10}(OsO_4)_6I_2$, $Sr_{10}(AsO_4)_6I_2$ (−15.93), $Sr_{10}(VO_4)_6I_2$ (+2.13) |

The $\Delta E^D$ (meV/atom) from DFT-CH analysis are also given for a few compounds in parentheses. Any compound with $\Delta E^D \leq +16$ meV/atom will be predicted as formable in the apatite structure, whereas compounds with $\Delta E^D > +16$ meV/atom will predicted as nonformable from the DFT-CH analysis.

We predicted two compounds to be thermodynamically stable with $\Delta E^D < 0$, namely [$Ba_{10}(NbO_4)_6I_2$ and $Ba_{10}(TaO_4)_6I_2$]. However, the calculated $\Delta E^D$ values were uncharacteristically small. Plausible reasons include mismatch in the choice of DFT codes. For instance, our DFT calculations were performed using an ultrasoft pseudopotential and PBEsol-GGA functional in Quantum ESPRESSO, whereas the decomposition energies were taken from the OQMD that uses projector-augmented wave pseudopotential and PBE-GGA functional in VASP (The Vienna Ab initio simulation package). Other source for discrepancy can come from the consideration of only relatively fewer or nonrepresentative compounds in the construction of CH for the Ba–Nb–O–I and Ba–Ta–O–I element sets. More discussions on this topic can be found in Ref. [15]. For consistency, we also checked whether Materials Project[56] can predict the decomposition reactions for $Ba_{10}(NbO_4)_6I_2$ and $Ba_{10}(TaO_4)_6I_2$. However, our query returned with "No data available." This is likely because there is no experimental data on $Ba_{10}(NbO_4)_6I_2$ and $Ba_{10}(TaO_4)_6I_2$ anywhere in the published literature. Therefore, we recommend the readers to interpret the $\Delta E^D$ for $Ba_{10}(NbO_4)_6I_2$ and $Ba_{10}(TaO_4)_6I_2$ in Table III with cautious optimism. Furthermore, the random forest models predicted that there is only 32% and 28% probability of forming an apatite phase in $Ba_{10}(NbO_4)_6I_2$ and $Ba_{10}(TaO_4)_6I_2$, respectively.

We now turn our attention to $Sr_{10}(PO_4)_6I_2$ in Table III that has a $\Delta E^D$ value of +8.03 meV/atom. The $\Delta E^D$ is below the DOM threshold of +16 meV/atom. Thus, we predict $Sr_{10}(PO_4)_6I_2$ to be metastable and potentially formable. The random forest models, on the other hand, predicted that $Sr_{10}(PO_4)_6I_2$ has a 44.3% probability of forming an apatite phase. These results can be used as guidelines for the synthesis of novel iodoapatite compounds.

We also performed DFT-CH calculations on six hypothetical compounds that were predicted to be formable by both random forest and gradient tree boosting methods (see Table II): $Ca_{10}(AsO_4)_6I_2$, $Sr_{10}(VO_4)_6I_2$, $Ba_{10}(PO_4)_6I_2$, $Sr_{10}(AsO_4)_6I_2$, $Ba_{10}(AsO_4)_6I_2$, and $Ba_{10}(VO_4)_6I_2$. Five of the compounds have $\Delta E_D \leq +16$ meV/atom, indicating that they are only



https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

**Table III.** DFT-CH predictions for 12 novel iodoapatites, where random forest and gradient tree boosting methods disagree in their predictions.

| $A_{10}(BO_4)_6X_2$ | Space group | GTB | RF | $\Delta E^D$ (meV/atom) | Final prediction |
|---|---|---|---|---|---|
| $Ba_{10}(GaO_4)_6I_2$ | $P2_1/b$ | Yes | No | 29.92 | Unstable |
| $Ba_{10}(WO_4)_6I_2$ | $P2_1/b$ | Yes | No | 156.97 | Unstable |
| $Ba_{10}(NbO_4)_6I_2$ | $P6_3/m$ | Yes | No | −943.83 | Stable |
| $Ba_{10}(TaO_4)_6I_2$ | $P6_3/m$ | Yes | No | −1533.19 | Stable |
| $Ca_{10}(MoO_4)_6I_2$ | $P6_3/m$ | Yes | No | 137.96 | Unstable |
| $Cd_{10}(MoO_4)_6I_2$ | $P2_1/b$ | Yes | No | 177.06 | Unstable |
| $Cd_{10}(NbO_4)_6I_2$ | $P\bar{1}$ | Yes | No | 64.49 | Unstable |
| $Sr_{10}(GaO_4)_6I_2$ | $P2_1/b$ | Yes | No | 50.69 | Unstable |
| $Sr_{10}(NbO_4)_6I_2$ | $P2_1/b$ | Yes | No | 46.52 | Unstable |
| $Sr_{10}(TaO_4)_6I_2$ | $P2_1/b$ | Yes | No | 47.33 | Unstable |
| $Sr_{10}(PO_4)_6I_2$ | $P2_1/b$ | Yes | No | 8.03 | |

$\Delta E^D$ (in meV/atom) is the decomposition energy or the distance from the CH. When $\Delta E^D \leq 0$, then the compound is predicted to be thermodynamically stable. When $\Delta E^D > +16$, then the compound is predicted to be unstable. The inequality condition, $0 < \Delta E^D \leq +16$, represents the degree of metastability (DOM) region.

weakly metastable and potentially formable. The exception is $Ba_{10}(AsO_4)_6I_2$ that has a calculated $\Delta E_D$ of +21.27 meV/atom. Since the value of $\Delta E_D$ for $Ba_{10}(AsO_4)_6I_2$ is >+16 meV/atom (DOM), the prediction from DFT-CH analysis is that $Ba_{10}(AsO_4)_6I_2$ is not formable in the apatite crystal structure. This result is in disagreement with the ML predictions. All compounds have the monoclinic space group ($P2_1/b$) as the lowest-energy crystal structure.

Finally, we compare our predictions with the work of Wang[22] who predicted that when the A-site is $Sr^{2+}$, $Pb^{2+}$, $Cd^{2+}$, and $Ba^{2+}$ and when the B-site is $V^{5+}$, $As^{5+}$, $Cr^{5+}$, and $Mo^{5+}$, then the structure will have the suitable channel size for incorporating iodine in the lattice. Our ML predictions generally agree with Wang's work, when A={$Sr^{2+}$ and $Pb^{2+}$} and B={$V^{5+}$, $As^{5+}$, and $Mo^{5+}$} (see Table II). Our work also predicts new and formable $Ca^{2+}$-based iodine-containing apatites that were missed by Wang. However, there were disagreements when considering the Cd-containing compounds, where DFT-CH analyses predict some of the Cd-containing iodoapatites to be thermodynamically unstable (see Table III). Experimental synthesis and characterization will be needed to provide validation and feedback.

## Conclusion

Predicting the formability and thermodynamic stability is one of the important first steps for assessing the potential of realizing any compound as a waste form for immobilizing toxic elements, such as radioactive iodine. We have made progress by exploring two complementary computational methods, such as data-driven ML and physics-based DFT, to address the problem. ML methods predict 18 new iodoapatites as formable. In addition, there were 12 compounds that posed a challenge to ML methods. We performed DFT-CH analysis on those 12 compounds and predicted two to be thermodynamically stable and one to be metastable (DOM ≤+16 meV/atom). The remaining nine were predicted to be unstable. We recommend the promising formable new compounds for experimental synthesis and validation. In addition to line compounds, the ML models also have the capability to rapidly predict the formability of complex solid solutions, which are beyond the scope of the current state-of-the-art DFT methods. Although the analysis and results presented focused on the formability and stability of iodine-containing apatites, our ML models are capable of exploring other chemical substitutions, including toxic Hg.[57] Furthermore, our computational approach is generic and can seamlessly be adapted to other crystal systems (e.g., lanthanide phosphates[58]) to address problems in environmental remediation and related applications, where phase formability and stability are important for design considerations.

## Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1557/mrc.2019.103.

## Acknowledgments

All authors acknowledge support from the Advanced Research Computing Services at the University of Virginia (UVa), which maintains the Rivanna supercomputing cluster where we performed the DFT calculations. M.V.A and P.V.B acknowledge support from the UVa start-up funds. T.Q.H acknowledges support from the DARPA-Topological Excitations in Electronics program (grant# D18AP00009). The content of the information


https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. Approved for public release; distribution is unlimited. All authors thank Dr Shruba Gangopadhyay for providing insightful comments on the manuscript.

## References


1. W.J. Weber, R.C. Ewing, C.R.A. Catlow, T.D. de la Rubia, L.W. Hobbs, C. Kinoshita, H. Matzke, A.T. Motta, M. Nastasi, E.K.H. Salje, E. R. Vance and S. J. Zinkle: Radiation effects in crystalline ceramics for the immobilization of high-level nuclear waste and plutonium. *J. Mater. Res.* **13**, 1434–1484 (1998).
2. I.W. Donald, B.L. Metcalfe, and R.N.J. Taylor: The immobilization of high level radioactive wastes using ceramics and glasses.. *J. Mater. Sci.* **32**, 5851–5887 (1997).
3. Y. Watanabe, T. Ikoma, H. Yamada, Y. Suetsugu, Y. Komatsu, G.W. Stevens, Y. Moriyoshi, and J. Tanaka: Novel long-term immobilization method for radioactive iodine-129 using a Zeolite/Apatite composite sintered body. *ACS Appl. Mater. Interfaces* **1**, 1579–1584 (2009).
4. K. Masuda, O. Kato, Y. Tanaka, S. Nakajima, S. Okamoto, T. Sakuragi, and S. Yoshida: Iodine immobilization: Development of solidification process for spent silver-sorbent using hot isostatic press technique. *Prog. Nucl. Energy* **92**, 267–272 (2016).
5. B.J. Riley, J.D. Vienna, D.M. Strachan, J.S. McCloy, and J.L. Jerden Jr: Materials and processes for the effective capture and immobilization of radioiodine: a review. *J. Nucl. Mater.* **470**, 307–326 (2016).
6. T.J. White and D. ZhiLi: Structural derivation and crystal chemistry of apatites. *Acta Crystallogr. B* **59**, 1–16 (2003).
7. T. White, C. Ferraris, J. Kim, and S. Madhavi: Apatite-An adaptive framework structure. *Rev. Mineral. Geochem.* **57**, 307–401 (2005).
8. J.Y. Kim, Z. Dong, and T.J. White: Model apatite systems for the stabilization of toxic metals: II, cation and metalloid substitutions in chlorapatites. *J. Am. Ceram. Soc.* **88**, 1253–1260 (2005).
9. B. Wopenka and J.D. Pasteris: A mineralogical perspective on the apatite in bone. *Mater. Sci. Eng. C* **25**, 131–143 (2005).
10. L.I. Ardanova, E.I. Get'man, S.L. Loboda, V.V. Prisedsky, T.V. Tkachenko, V.I. Marchenko, V.P. Antonovich, N.A. Chivireva, K.A. Chebishev, and A.S. Lyashenko: Isomorphous substitutions of rare earth elements for calcium in synthetic hydroxyapatite. *Inorg. Chem.* **49**, 10687–10693 (2010).
11. M. Stennett, I. Pinnock, and N. Hyatt: Rapid synthesis of $Pb_5(VO_4)_3I$, for the immobilisation of iodine radioisotopes, by microwave dielectric heating. *J. Nucl. Mater.* **414**, 352–359 (2011).
12. T. Yao, F. Lu, H. Sun, J. Wang, R.C. Ewing, and J. Lian: Bulk iodoapatite ceramic densified by spark plasma sintering with exceptional thermal stability. *J. Am. Ceram. Soc.* **97**, 2409–2412 (2014).
13. F. Legrain, J. Carrete, A. van Roekeghem, G.K. Madsen, and N. Mingo: Materials screening for the discovery of new half-heuslers: Machine learning versus ab Initio methods. *J. Phys. Chem. B* **122**, 625–632 (2018).
14. P.V. Balachandran, S.R. Broderick, and K. Rajan: Identifying the "inorganic gene" for high temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. Lond. A: Math. Phys. Eng. Sci.* **467**, 2271–2290 (2011).
15. P.V. Balachandran, A.A. Emery, J.E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger: Predictions of new $ABO_3$ perovskite compounds by combining machine learning and density functional theory. *Phys. Rev. Mater.* **2**, 043802 (2018).
16. G. Pilania, P.V. Balachandran, C. Kim, and T. Lookman: Finding new perovskite halides via machine learning. *Front. Mater.* **3**, 19 (2016).
17. Z. Li, Q. Xu, Q. Sun, Z. Hou, and W.-J. Yin: Thermodynamic stability landscape of halide double perovskites via high-throughput computing and machine learning. *Adv. Funct. Mater.* **29**, 1807280 (2019).
18. W. Li, R. Jacobs, and D. Morgan: Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **150**, 454–463 (2018).
19. C.J. Bartel, C. Sutton, B.R. Goldsmith, R. Ouyang, C.B. Musgrave, L.M. Ghiringhelli, and M. Scheffler: New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019). doi:10.1126/sciadv.aav0693.
20. N. Islam, W. Huang, and H.L. Zhuang: Machine learning for phase selection in multi-principal element alloys. *Comput. Mater. Sci.* **150**, 230–235 (2018).
21. F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, and A. Mehta: Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018), doi:10.1126/sciadv.aaq1566.
22. J. Wang: Incorporation of iodine into apatite structure: A crystal chemistry approach using artificial neural network. *Front. Earth Sci.* **3**, 20 (2015).
23. D. Optiz and R. Maclin: Popular ensemble methods: An empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).
24. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton: Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
25. S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, and C. Wolverton: The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
26. J. Ito: Silicate apatites and oxyapatites. *Am. Miner.* **53**, 890–907 (1968).
27. A.G. Cockbain: The crystal chemistry of apatites. *Mineral. Mag.* **37**, 654–660 (1968).
28. D.A. Grisafe and F.A. Hummel: Crystal chemistry and color in apatites containing cobalt, nickel and rare-earth ions. *Am. Mineral.* **55**, 1131–1145 (1970).
29. D.A. Grisafe and F.A. Hummel: Pentavalent ion substitutions in the apatite structure: Part a. crystal chemistry. *J. Solid State Chem.* **2**, 160–166 (1970).
30. E.R. Kriedler and F.A. Hummel: The crystal chemistry of apatite: Structure fields of fluor- and chlorapatite. *Am. Mineral.* **55**, 170–184 (1970).
31. T. Dordević, S. Šutovic, J. Stojanovic, and L. Karanović: Sr, Ba and Cd arsenates with the apatite-type structure. *Acta Crystallogr. C* **64**, i82–i86 (2008).
32. Z. Dong, T.J. White, B. Wei, and K. Laursen: Model apatite systems for the stabilization of toxic metals: I, calcium lead vanadate. *J. Am. Ceram. Soc.* **85**, 2515–2522 (2002).
33. C.C. Silva, I.F. Vasconcelos, A.S.B. Sombra, and M.A. Valente: Magnetic properties study on fe-doped calcium phosphate. *Phys. Scr.* **80**, 055706 (2009).
34. T. Baikie, P.H.J. Mercier, M.E. Elcombe, J.Y. Kim, Y.L. Page, L.D. Mitchell, T.J. White, and P.S. Whitfield: Tricilinic apatites. *Acta Crystallogr. B* **63**, 251–256 (2007).
35. T. Suzuki, T. Hatsushika, and M. Miyake: Synthetic hydroxyapatites as inorganic cation exchangers: Part 2. *J. Chem. Soc.: Faraday Trans. I* **78**, 3605–3611(1982).
36. N.J. Flora, C.H. Yoder, and H.D.B. Jenkins: Lattice energies of apatites and the estimation of $\Delta H_f^\circ$ ($PO_4^{3-}$, g). *Inorg. Chem.* **43**, 2340–2345 (2004).
37. S. Sugiyama: Approach using apatite to studies of energy and environment. *Phosphorus Res. Bull.* **21**, 1–8 (2007).
38. K. Matsunaga, H. Inamori, and H. Murata: Theoretical trend of ion exchange ability with divalent cations in hydroxyapatite. *Phys. Rev. B* **78**, 094101 (2008).
39. P.V. Balachandran and K. Rajan: Structure maps for $A_4^{I}A_6^{II}(BO_4)_6X_2$ apatite compounds via data mining. *Acta Crystallogr. B* **68**, 24–33 (2012).
40. O. Muller and R. Roy: *The Major Ternary Structural Families* (Springer-Verlag, New York, 1974).
41. L.H. Ahrens: Anion affinity and polarizing power of cations. *Nature* **169**, 463 (1952).
42. R.D. Shannon: Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. A* **32**, 751–767 (1976).
43. L. Pauling: *The Nature of the Chemical Bond* (Cornell University Press, Ithaca, 1960).
44. E.V. Anslyn and D.A. Dougherty: Introduction to Structure and Models of Bonding. In *Modern Physical Organic Chemistry*, edited by J. Murdzek (University Science Books, Sausalito, California, 2006).




https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press

45. L. Breiman: Random forests. *Mach. Learn.* **45**, 5–32 (2001).
46. A. Liaw and M. Wiener: Classification and regression by randomForest. *R News* **2**, 18 (2002).
47. J.H. Friedman: Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
48. J.H. Friedman: Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
49. B. Greenwell, B. Boehmke, J. Cunningham, and G. Developers: *gbm: Generalized Boosted Regression Models*, r package version 2.1.5 (2019).
50. P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G.L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal 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: QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
51. J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, and K. Burke: Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).
52. A.D. Corso: Pseudopotentials periodic table: From H to Pu. *Comput. Mater. Sci.* **95**, 337–350 (2014).
53. P.V. Balachandran, K. Rajan, and J.M. Rondinelli: Electronically driven structural transitions in $A_{10}(BO_4)_6F_2$ apatites (A = Ca, Sr, Pb, Cd and Hg). *Acta Crystallogr. B* **70**, 612–615 (2014).
54. M.P. Fay and M.A. Proschan: Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat. Surv.* **4**, 1 (2010).
55. C. Cao, S. Chong, L. Thirion, J.C. Mauro, J.S. McCloy, and A. Goel: Wet chemical synthesis of apatite-based waste forms–A novel room temperature method for the immobilization of radioactive iodine. *J. Mater. Chem. A* **5**, 14331–14342 (2017).
56. 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: Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013), doi:10.1063/1.4812323.
57. Q. Wang, D. Kim, D.D. Dionysiou, G.A. Sorial, and D. Timberlake: Sources and remediation for mercury contamination in aquatic systems - a literature review. *Enviorn. Pollut.* **131**, 323–336 (2004).
58. M. Toro-González, D.M. Clifford, R. Copping, S. Mirzadeh, and J.V. Rojas: Synthesis and characterization of intrinsically radiolabeled lanthanide phosphate nanoparticles toward biomedical and environmental applications. *J. Nanopart. Res.* **20**, 238 (2018).


https://doi.org/10.1557/mrc.2019.103 Published online by Cambridge University Press