Computational Materials Science xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Crystal structure classification in ABO_3 perovskites via machine learning

Santosh Behara [a], Taher Poonawala [a], Tiju Thomas [a,b,*]

[a] Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[b] DST Solar Energy Harnessing Center – An Energy Consortium, Indian Institute of Technology Madras, Chennai 600036, India

ARTICLE INFO

Keywords:
ABO_3 perovskites
Crystal structure prediction
Light GBM
Feature importance graph
SHAP analysis

ABSTRACT

Crystal structure classification of perovskites (ABO_3) is done using the Light Gradient Boosting Machine (Light GBM) algorithm. In this work, we have identified features such as electronegativity, ionic radius, valence, and bond lengths of A-O and B-O pairs that enable a priori crystal structure prediction. We have taken 5329 ABO_3 perovskites and applied the proposed model to 675 compounds. It successfully categorized the compounds into cubic, tetragonal, orthorhombic, and rhombohedral structures with 80.3% best accuracy using 5-*fold* cross-validation. Therefore, the model can be used as a preliminary, fast, and inexpensive method to classify perovskites into their respective crystal systems. Feature importance graph and SHapley Additive exPlanations (SHAP) are used in feature ranking and crystal structure prediction. These composition-structure predictions will find applications in ceramic engineering and solid-state chemistry.

1. Introduction

Perovskite structures have been extensively studied in ceramic science and engineering, materials physics, and solid-state inorganic chemistry because of their compositional flexibility, distortion of the cation configuration, and mixed valence state electronic structure. This becomes a means for tuning the material's properties [1]. An ideal perovskite has an ABX_3 structure where A and B are two differently sized cations, and X is an anion (X is oxygen in this work). The A cation is surrounded by 12 oxygen ions in a dodecahedral environment, and B cation is octahedrally coordinated by 6 oxygen ions. The ideal ABO_3 belongs to the cubic space group $Pm\bar{3}m$ [2]. The deviation from the ideal cubic structure is generally attributed to one of the three mechanisms: (i) displacement of the cations, (ii) distortions of the octahedra, and (iii) tilting of the octahedra [3].

The first two mechanisms (*i.e.,* (i) and (ii), given above) are driven by the electronic instabilities of the cation. The displacement of titanium in ferroelectric $BaTiO_3$ [4] and the John-Teller distortion in $LaMnO_3$ [5] are examples of electronic instabilities. The third mechanism is due to the tilting of BO_6 octahedra while maintaining their corner-sharing connectivity [6]. These structural deviations result in tetragonal, orthorhombic, rhombohedral, monoclinic, and triclinic crystal structures in perovskites [7]. The crystal structures and its role in simple and complex perovskite applications are mentioned in Table 1. Most of the metallic ions in the periodic table can form a perovskite structure. From the work of Emery et al. [8], we have identified 5329 compounds of ABO_3 perovskite-type oxides (Ref: Fig. 1).

In this paper, we have identified a list of features used to classify 5329 ABO_3 compounds into cubic, tetragonal, orthorhombic and rhombohedral systems based on the proposed model. Monoclinic and triclinic systems are omitted due to their negligible fraction (~0.3%) in the entire dataset.

2. Methodology

We have created a dataset of 5329 ABO_3 perovskite-type oxides (Ref: Supplementary information file - SIF-1) along with its features (Ref: Table 2). Out of 5329 compounds, 53 compounds are removed as the report does not specify any lowest distortion crystal type [8]. One of the features used in our work is "Valence". The valence of the ions is calculated by using the bond-valence (BV) model [36]. The ABO_3 compounds which have the entries "element not in BV" and "not balanced" in their 'Valence A' column are removed. With a loss of 3063 compounds, 2213 compounds remain and are used for the data analysis.

'Valence A' has 5 input values (1, 2, 3, 4, and 5). So, it is treated here as a categorical variable. Categorical variables are those that have two or more categories, and which can take an input belonging only to one of these categories. Categorical variables have to be encoded [37], and hence one-hot encoding (in one-hot encoding, a binary variable is created for every possible input value) is performed on 'Valence A'. In

* Corresponding author at: Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India.
E-mail address: tt332@cornell.edu (T. Thomas).

https://doi.org/10.1016/j.commatsci.2020.110191
Received 25 July 2020; Received in revised form 14 October 2020; Accepted 16 November 2020

Table 1
ABO_3 perovskite compounds and its applications.

S. No	Compound	Crystal structure	Property	Applications
1	$BaZrO_3$	Cubic	Proton and ionic conductivity	Protonic Fuel cell [9] and hydrogen separation membrane [10]
2	$(BaK)BiO_3$	Cubic	Superconductivity	Superconductor [11]
3	$AgSbO_3$	Cubic	Photocatalytic	Visible-light sensitive photocatalyst [12]
4	$BaTiO_3$	Tetragonal	Ferroelectricity	Multilayer capacitor [13]
5	$Pb(Zr,Ti)O_3$	Tetragonal	Piezoelectricity	Piezoelectric transducer [14]
6	$PbTiO_3$	Tetragonal	Pyroelectricity	Pyroelectric infrared detector [15]
7	$LaCrO_3$, $LaFeO_3$	Orthorhombic	Mixed conductivity	Solid oxide fuel cell cathode [16]
8	$GdFeO_3$, $LaMnO_3$	Orthorhombic	Magnetic	Magnetic memory and ferromagnetism [17,18]
9	$YAlO_3$, $KNbO_3$	Orthorhombic	Optical	Laser [19,20]
10	$BiFeO_3$	Rhombohedral	Multiferroic	Spintronics and memory devices [21]
11	$Na_{0.5}Bi_{0.5}TiO_3$	Rhombohedral	Piezoelectricity	Lead-free piezoelectric [22,23]
12	$LaAlO_3$	Rhombohedral	Catalytic	Industrial catalyst for oxidative coupling of methane (OCM) [24]

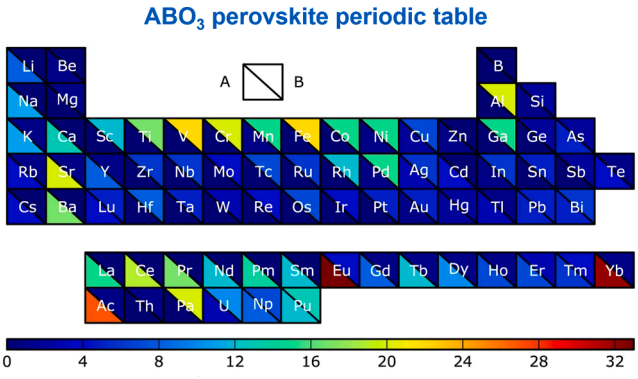

Fig. 1. The solid-state periodic table shows the frequency at which each element appears on the A and B-sites of the ABO_3 perovskite structure. Elements with frequency 0 are shown in *blue* color. From this work, 73 elements are identified in A and B-sites of ABO_3 structure. It results in a total of 5329 ($\sim 73^2$) perovskite-type oxides, and all are used in this study. Adapted from ref. [8] with permission from the Nature ©2017 under the Creative Commons license CC 4.0 (http://creativecommons.org/licenses/by/4.0/).

this case, since there are five valence categories, 5 binary variables are created. If the compound has 'Valence A' as 1, then the binary variable corresponding to Valence 1 takes the value of 1, whereas the other binary variables corresponding to Valence 2, 3, 4, and 5 take the value of 0. We have dropped the binary variable columns of 'Valence 4′ and 'Valence 5′ since all the known perovskites have $A^{1+}B^{5+}O_3$, $A^{2+}B^{4+}O_3$, and $A^{3+}B^{3+}O_3$ structures only [1,38]. One-hot encoding is done to prevent the model from assuming any false relationship between the 'Valence A' values. Ionic radii (r) values are taken from Shannon [29] and radii values at other coordination numbers (12 or 6) are calculated using the linear extrapolation method [8]. Experimentally reported perovskites are taken from the literature [25–28]. The preferred crystal structure with the lowest formation energies is reported from the work of Emery et al. [8]. Average electronegativity (*EN*) values of ions are taken from [30] and electronegativity difference with radius is modified from the reference [33]. Ionic radius and electronegativity of oxygen are taken as 1.40 Å and 3.44, respectively. Bond lengths (l) of A-O and B-O pairs are calculated by using the BV model [31,32].

Goldschmidt tolerance factor $\left(t_G = \frac{(r(A)+r(B))}{\sqrt{2}(r(B)+r(O))}\right)$ and octahedral factor $\left(\mu = \frac{r(B)}{r(O)}\right)$ are important in governing the stability and formability of perovskite structures [39]. A t_G value of 1 shows a perfectly cubic structure. If $t_G > 1$, it indicates large-sized A cation and results the formation of hexagonal crystal structure, as observed in $BaNiO_3$ (t_G = 1.13) [40]. One the other hand, if $t_G < 0.82$, it excludes the perovskite formation and leads to other alternative structures like ilmenite ($FeTiO_3$, t_G = 0.81) [41], corundum ($AlAlO_3$, t_G = 0.71) [42] etc. So, we have chosen the range, $0.82 < t_G < 1.10$, to classify perovskites. Another important parameter is the octahedral factor (μ). To form regular BO_6 octahedra, the accepted values of μ are in the range 0.414 to 0.732 [43]. With these bounds of t_G and μ, we have further eliminated 1538 compounds and examined our model on 675 ABO_3 perovskites (Ref: SIF-2 and Fig. S1 in SIF-3). A new tolerance factor (τ) proposed by Bartel et al. [35] is taken here as another feature.

3. Prediction model

Here, a multiclass classification model is proposed for classifying 675 ABO_3 type perovskites into cubic, tetragonal, orthorhombic, and rhombohedral crystal structures (monoclinic and triclinic systems are omitted due to their negligible fraction (~0.3%)). We have used the *Scikit-learn* library [44] to carry out preprocessing and feature encoding. *Scikit-learn* library has in-built functions that enable us to carry out the preprocessing steps such as oversampling, and splitting the data into the k equal folds required for using cross-validation. For the multiclass classification problem, we have used the Light Gradient Boosting Machine (Light GBM) algorithm.

Light GBM is a fast, distributed, tree-based algorithm. The nodes of the tree represent splits based on the input feature, and the data points at the leaves (a leaf is a collection of data points that have similar feature values) correspond to the prediction (Ref: Fig. S2 in SIF-3). The advantage this offers is that it takes rather nominal memory to run and can handle large datasets (with thousands of features) quite effectively [45]. It uses an ensemble of classifiers to improve accuracy as it combines the predictions of many classifiers. It also gives the influence of one or more features on the final result [46]. This allows for the ranking of the features by their relevance, which has value to materials design, in this case.

Out of the 17 features proposed above, only 13 are selected for the model. The features such as 'Compounds', 'A', 'B', 'In literature', and 'Lowest distortion' describe ABO_3 perovskite-type oxides and its reported crystal structure information in the work of Emery et al. [8]. The features 'Compounds', 'A', 'B' and 'In literature' are not used in the model. The feature 'Lowest distortion' is the output variable that the model has to predict. The feature 'Valence A' has three separate values as it shows three different valences in ABO_3 ($A^{1+}B^{5+}O_3$, $A^{2+}B^{4+}O_3$, and $A^{3+}B^{3+}O_3$) structures [1,38]. 'Valence B' is omitted because it does not contain any new information other than what is already provided by 'Valence A' ($ValenceB = 3 \times ValenceO - ValenceA$) in a neutral compound. The feature 'Radius A' is divided into two features, which are Radius A at XII coordination and Radius A at VI coordination. This is so since ionic radius is a function of the coordination number. In an ideal

perovskite structure, the coordination number of A-site cation is reported to be XII. And, in low-symmetries like orthorhombic and rhombohedral structures, smaller coordination numbers (VIII or VI) are usually reported [47,48]. Here we have taken the ionic radii of A-site cations at VI coordination. Another feature, 'Radius B' is removed as it is completely correlated with the 'Octahedral factor (μ)' feature and hence provides no new information to the model (Fig. 2). So, the final list of features which are used as input parameters to build our classification model is:

1. Valence A – *1* ($v(A^{1+})$)
2. Valence A – *2* ($v(A^{2+})$)
3. Valence A – *3* ($v(A^{3+})$)
4. Radius A at 12 coordination ($r(A_{XII})$)
5. Radius of A at 6 coordination ($r(A_{VI})$)
6. Electronegativity of A (*EN(A)*)
7. Electronegativity of B (*EN(B)*)
8. Bond length of A-O pair (*l(A-O)*)
9. Bond length of B-O pair (*l(B-O)*)
10. Electronegativity difference with radius (ΔENR)
11. Goldschmidt tolerance factor (t_G)
12. New tolerance factor (τ)
13. Octahedral factor (μ)

The conventional gradient boosting algorithm does not work very well if the features are highly correlated [49]. So, we have checked the Spearmen's rank correlation [50] among variables to ensure that the features are not highly correlated. The formula for the Spearmen's rank correlation (ρ) is given as follows:

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\left(R(x_i)-\overline{R(x)}\right)\left(R(y_i)-\overline{R(y)}\right)\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i)-\overline{R(x)}\right)^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i)-\overline{R(y)}\right)^2\right)}} \quad (1)$$

Where $R(x_i)$ and $R(y_i)$ represent the rank of each data point of the two features and $\overline{R(x)}$ and $\overline{R(y)}$ represent the rank of the mean values of the two features whose correlation is to be measured, n is the number of data points. In order to determine the rank, order the data points from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest, and so on. The value of 'ρ' is ranging from −1 to 1. A high value of 'ρ' indicates a high positive correlation between those features. From the values of 'ρ' in this work, one can say that the correlation among features is sufficiently weak except 'Radius B'. It is showing a strong correlation with 'Octahedral factor μ'.

4. Working of the Light GBM

Light GBM utilizes gradient boosting decision trees, which use boosting to combine individual decision trees [51]. A model built on an individual decision tree is prone to overfitting. This is due to the amount of specificity we look at while classifying based on the feature values. It leads to a smaller sample of events that meet the previous assumptions. This small sample may lead to unsound conclusions. This problem is avoided when we build multiple decision trees. Boosting combines individual decision trees (weak learners) in series to form a strong learner (which comprises of sequentially connected weak learners). Each new tree that is added attempts to minimize the error of the previous trees. This improves the accuracy of the model. The final model then combines the result from each step so that a strong learner is achieved. This prevents overfitting and improves the model quality (Ref: Fig. S2 in SIF-3) [51]. Other model parameters used in this study are given in SIF-3.

Unlike other tree-based algorithms like eXtreme Gradient Boosting (XGBoost) that grow horizontally, Light GBM tree grows vertically. This means that trees in the Light GBM model grow leaf-wise and not level-wise [52]. A leaf-wise algorithm selects the leaf with the largest delta loss (change in the loss function) to grow on. This is beneficial as the leaf-wise growth method is shown to cause a greater decrease in the loss function than a level-wise growth algorithm. Hence, it provides better accuracy than other gradient boosting tree models (Ref: Fig. S3 in SIF-3).

5. Results and discussion

We have used k-fold cross-validation [44] for carrying out the train-

Table 2
List of features selected in this study.

S. No	Feature	Type	Unit	Description
1	Compound	String	None	ABO_3 compound
2	A	String	None	Chemical element on the A-site
3	B	String	None	Chemical element on the B-site
4	In literature	Boolean	None	TRUE or FALSE. TRUE means perovskites experimentally reported in literature [25–28].
5	Lowest distortion	String	None	Preferred crystal structure with the lowest formation energy from the work of Emery et al. [6]
6	Valence A ($v(A)$)	Number or String	None	Valence of A is calculated by bond-valence (BV) model [7]. If ABO_3 is not balanced, it is denoted by "not balanced". If the compound contains at least one element without a BV parameter, it is considered as "element not in BV".
7	Valence B ($v(B)$)	Number or String	None	Valence of B is calculated by bond-valence (BV) model. If ABO_3 is not balanced, it is denoted by "not balanced". If the compound contains at least one element without a BV parameter, it is considered as "element not in BV".
8	Radius A ($r(A)$)	Number	Å	Shannon [29] ionic radius of A cation. Ionic radii of 12 and 6 coordination are estimated from the given coordination value in the database.
9	Radius B ($r(B)$)	Number	Å	Shannon [29] ionic radius of B cation. Ionic radius of 6 coordination is estimated from the given coordination value in the database.
10	Electronegativity of A ($EN(A)$)	Number	None	Average electronegativity value of A cation from the reference [30]
11	Electronegativity of B ($EN(B)$)	Number	None	Average electronegativity value of B cation from the reference [30]
12	Bond length of A-O ($l(A\text{-}O)$)	Number	Å	Bond length of A-O pair is estimated by BV model [31,32]
13	Bond length of B-O ($l(B\text{-}O)$)	Number	Å	Bond length of B-O pair is estimated by BV model [31,32]
14	Electronegativity difference with radius (ΔENR)	Number	None	Electronegativity difference with radius is modified as follows [33]: $\left(\left(\frac{r(A)}{r(O)}\right)(EN(A)-2\times EN(O))+\left(\frac{r(B)}{r(O)}\right)(EN(B)-EN(O))\right)$
15	Goldschmidt tolerance factor (t_G)	Number	None	Goldschmidt [34] tolerance factor is calculated as follows: $\left(t_G=\frac{(r(A)+r(B))}{\sqrt{2}(r(B)+r(O))}\right)$
16	New tolerance factor (τ)	Number	None	New tolerance factor [35] is calculated as follows: $\left(\tau=\frac{r(O)}{r(B)}-v(A)\left(v(A)-\frac{r(A)/r(B)}{ln(r(A)/r(B))}\right)\right)$
17	Octahedral factor (μ)	Number	None	Octahedral factor is calculated as follows: $\left(\mu=\frac{r(B)}{r(O)}\right)$

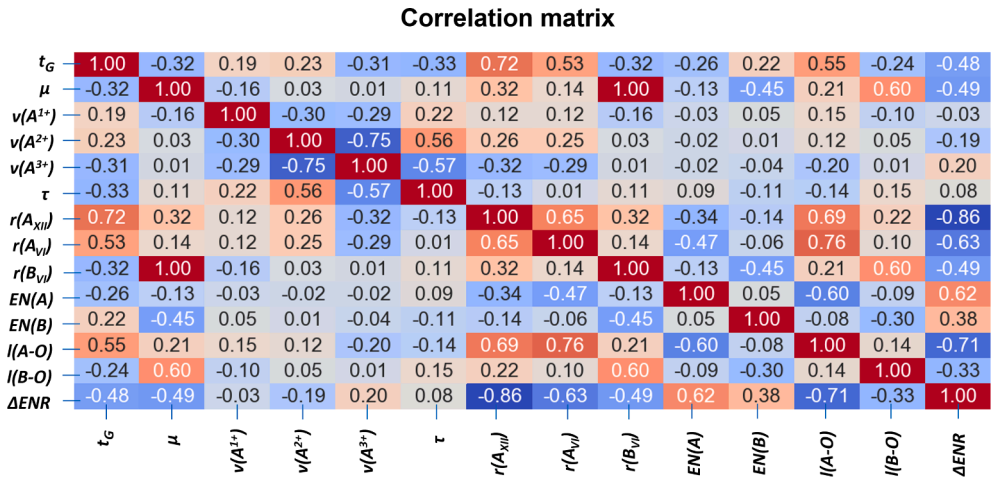

Fig. 2. Shows the correlation matrix. The Spearmen's rank correlation (ρ) does not show high positive values among features. It indicates that the correlation among features is sufficiently weak allowing their independent treatment in feature tables. The feature 'Radius B ($r(B_{VI})$)' is removed as it shows a high correlation with 'Octahedral factor μ'.

test split. In k-fold cross-validation, the data points split into k sets of equal sizes. The model is then trained on k-1 sets with 1 set being used as the test set. Each time the model is trained on a slightly different training set and gets a different accuracy. The reported accuracy is the mean of the k accuracies obtained. Hence k-*fold* cross-validation provides a more robust performance estimation than if it runs on just one particular train-test split. The model quality is judged using the 'accuracy_score' feature of the *Scikit-learn* package. The accuracy_score feature gives the ratio of the number of data points correctly classified to the total number of data points present. On running the model with 5-*fold* cross-validation using the Light GBM model, we obtained 80.3% accuracy and a standard deviation of 0.036.

Stratified k-fold cross-validation is carried out to ensure the proportion of each class in the folds remains constant. Our dataset contains 675 data points out of which 19 are tetragonal, 261 are cubic, 362 are orthorhombic, and 33 are rhombohedral. The dataset is highly imbalanced as 53.63% of the data in the dataset is orthorhombic, whereas only 2.81% is tetragonal, and 4.88% is rhombohedral. Due to the lower number of tetragonal and rhombohedral compounds, it is important to carry out oversampling on the folds so that the model is sufficiently trained on the minority classes.

For each run of the model on the k-1 folds, we oversample the data so that the number of data points in each minority class (tetragonal, rhombohedral, and cubic) becomes equal to the number of data points in the majority class (orthorhombic). Oversampling involves randomly duplicating entries from each of the minority classes so that the total number of entries in each class is the same [51]. Generally, oversampling overfits the data on the minority classes so that the model is trained on a balanced dataset. Hence, it becomes more robust when it has to be applied to the test set. This prevents the model from behaving as a dummy model and labelling all the data as belonging to the majority class.

If our model behaved as a dummy model and labelled all the points as orthorhombic (majority class), we would have $\sim$ 54% accuracy. However, our model gives an accuracy of 80.3%, which shows that our model behaves much better than a dummy model and that the oversampling has helped the model predictions. We have compared our Light GBM with other classification models, such as support vector machines (SVM, kernel-Radial Basis Function), Random forest (RF), and XGBoost. The accuracy of Radial Basis Function kernel SVM, RF, XGBoost, and Light GBM are 76.6%, 62.8%, 74.8%, and 80.3% respectively. Hence we choose the Light GBM as the algorithm for predicting crystal structures of ABO_3 perovskites for our dataset.

Kernel Density Estimation (KDE) [53] plots are generated to visualize how the relative energy for the compounds incorrectly classified by the model is different from the correctly classified compounds. Typically, KDE plots represent the relative frequency of a specific continuous random variable. The KDE plots for the formation energy of all compounds belonging to or predicted as a particular crystal structure are shown in Fig. 3. The 'green-colored' ones are the original formation energy values of a particular crystal structure taken from the work of Emery et al. [8]. The mean formation energy values for cubic, tetragonal, orthorhombic, and rhombohedral crystal structures are −2.12 eV, −2.05 eV, −2.57 eV, and −2.17 eV respectively. Compounds that are correctly classified by the Light GBM are represented using 'orange-color'. Incorrectly classified compounds are divided into two types. They are represented using two different colors (*blue* and *red*). Type 1, the '*blue-colored*' ones, are the compounds that do not belong to a particular crystal structure but are incorrectly classified as so. For example, the compounds that belong to tetragonal/orthorhombic/rhombohedral are assigned as cubic in Type 1 inaccurate assignment. On the other hand, Type 2 ('red color') inaccurate assignment involves a compound that belongs to a particular crystal structure but is assigned to any of the incorrect crystal structure possibilities. For instance, compounds assigned to tetragonal/orthorhombic/rhombohedral by our model actually belong to the cubic crystal system. All incorrectly classified crystal structures and their corresponding reported crystal structures by Emery et al. [8] are given in SIF-3. It is observed that the energy difference between formation energy for the structure reported by Emery et al. [8] and the one assigned in this work is significant in cases where misclassification occurs. Overall, however, the Light GBM results are in good agreement with the reports [8].

5.1. Feature importance graph:

Feature importance graph [46] arranges all features according to which one gives the most gain. Gain tells us the relative contribution of the particular feature to the model calculated by taking each feature's contribution for each tree in the model. The measures are based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees. The features having higher gain values are more important for generating a prediction. For our dataset, t_G, *EN(B)*, and *ΔENR* are the most important features in the classification of perovskite structures (Fig. 4). From the work of Bartel et al. [35], a value of $\tau < 4.18$ is considered as perovskite compound. For our dataset, our work shows the upper limit for τ is 5.86.

Below is the confusion matrix for the model (Fig. 5). We can see that

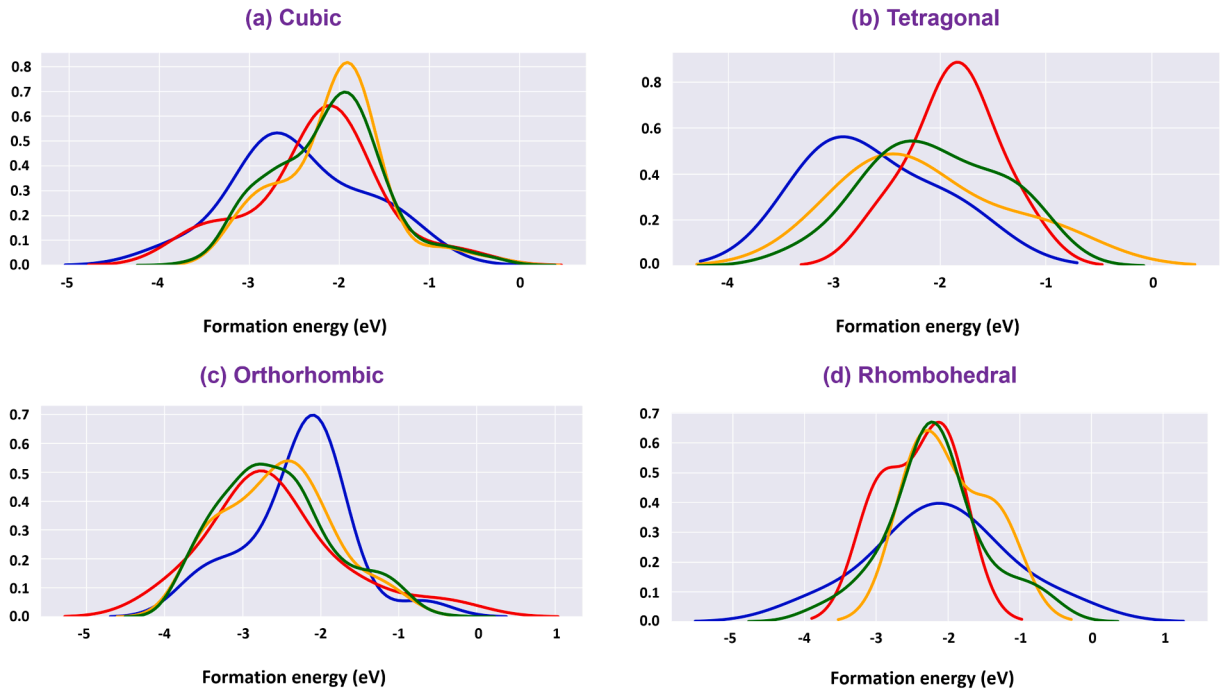

Fig. 3. Shows the Kernel Density Estimation (KDE) plots for the formation energy of all compounds belonging to or predicted as a particular crystal structure. The 'green-colored' ones are the original formation energy values of a particular crystal structure taken from the work of Emery et al. [8]. Compounds that are correctly classified by the Light GBM are represented using 'orange color'. Incorrectly classified compounds are divided into two types (*blue* and *red*). Type 1, the '*blue-colored*' ones, are the compounds that do not belong to a particular crystal structure but are incorrectly classified as so. On the other hand, Type 2 ('red color') inaccurate assignment involves a compound that belongs to a particular crystal structure but is assigned to any of the incorrect crystal structure possibilities. Overall, however, the Light GBM results are in good agreement with the reports [8].

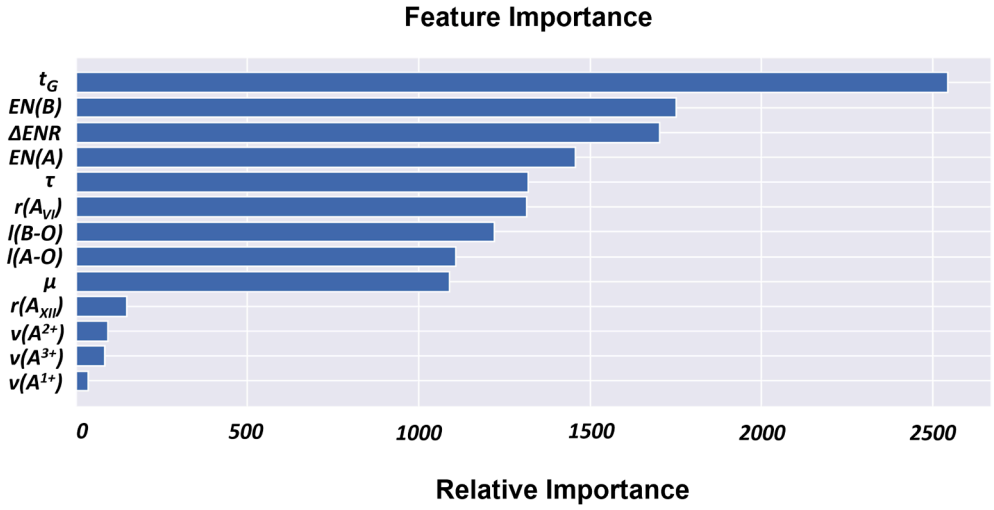

Fig. 4. Shows the feature importance graph. For our dataset, t_G, *EN(B)* and *ΔENR* consistently rank as the most important features in the classification of the 675 taken perovskite-type oxide structures. Crystal structures differ the most on the basis of the features at the top of the table.

the model has been trained to take the minority classes into account and not ignore them. The confusion matrix enables us to see the true predictions as well as false predictions. The diagonal elements represent true predictions, whereas the non-diagonal elements represent false predictions. Our model is evidently good at classifying the given crystal structures of ABO_3 perovskite-type oxides.

5.2. Shapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) are used to determine each feature's effect on the prediction of a particular class [54]. It provides information on how the value of that particular feature affects the probability of a predicted class. It utilizes the Shapley value for its analysis. Shapley value is the average of the marginal contribution of the feature to the class prediction across all permutations.

Shapley values can be explained as follows: assume there are N features, and S is a subset of the N features. Let $v(S)$ be the total contribution of the S features in the prediction model. When feature i is added to the S features, feature i's marginal contribution is $v(S\cup\{i\}) - v(S)$. If we take the average of the contribution over the possible different permutations in which the coalition can be formed, we can get the right contribution of the feature i [54].

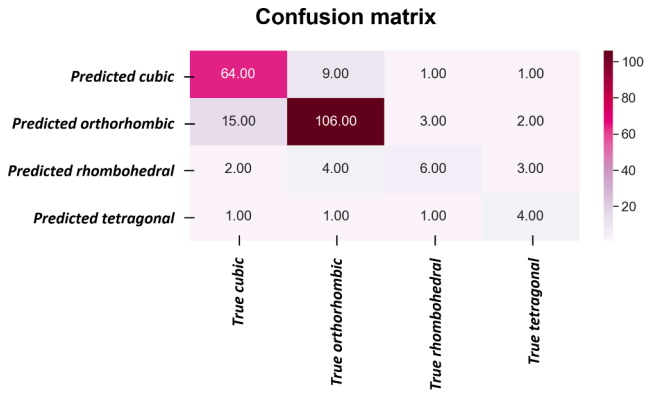

Fig. 5. Confusion matrix for the model proposed. The diagonal elements represent true predictions, whereas the non-diagonal elements represent false predictions. This explains how good the model is at correctly classifying the crystal structures and shows the contrast between the true crystal structure and what was predicted by the model.

Here, SHAP summary plots show how input feature's magnitude affects the probability of prediction of the particular crystal system (Fig. 6). The '*x*' axis has the Shapley values, and they represent the change in log odds (log odds is the logarithm of a ratio of the number of times the class occurs to the number of times the class does not occur). This indicates the value of how likely it is for a particular class to be predicted (a positive value means that it is more likely to predict that class). The features are listed on the '*y*' axis in descending order of feature importance. The red (or blue) color indicates that the feature's value at that data point is high (or low). The important point is to know that SHAP values help identify the features' contribution to predict the classes; however, they do not guarantee causality between feature value and prediction probability. Based on the information we get from SHAP summary plots, we can determine how the feature values affect the prediction of the crystal structure and subsequently alter particular feature values in an attempt to alter the crystal structure.

From the SHAP summary plot we may state the following,

(i) For cubic structures, low *EN(B)* and high τ have a large positive impact on the probability of the crystal structure's prediction. From the geometrical relations and the hard-sphere model

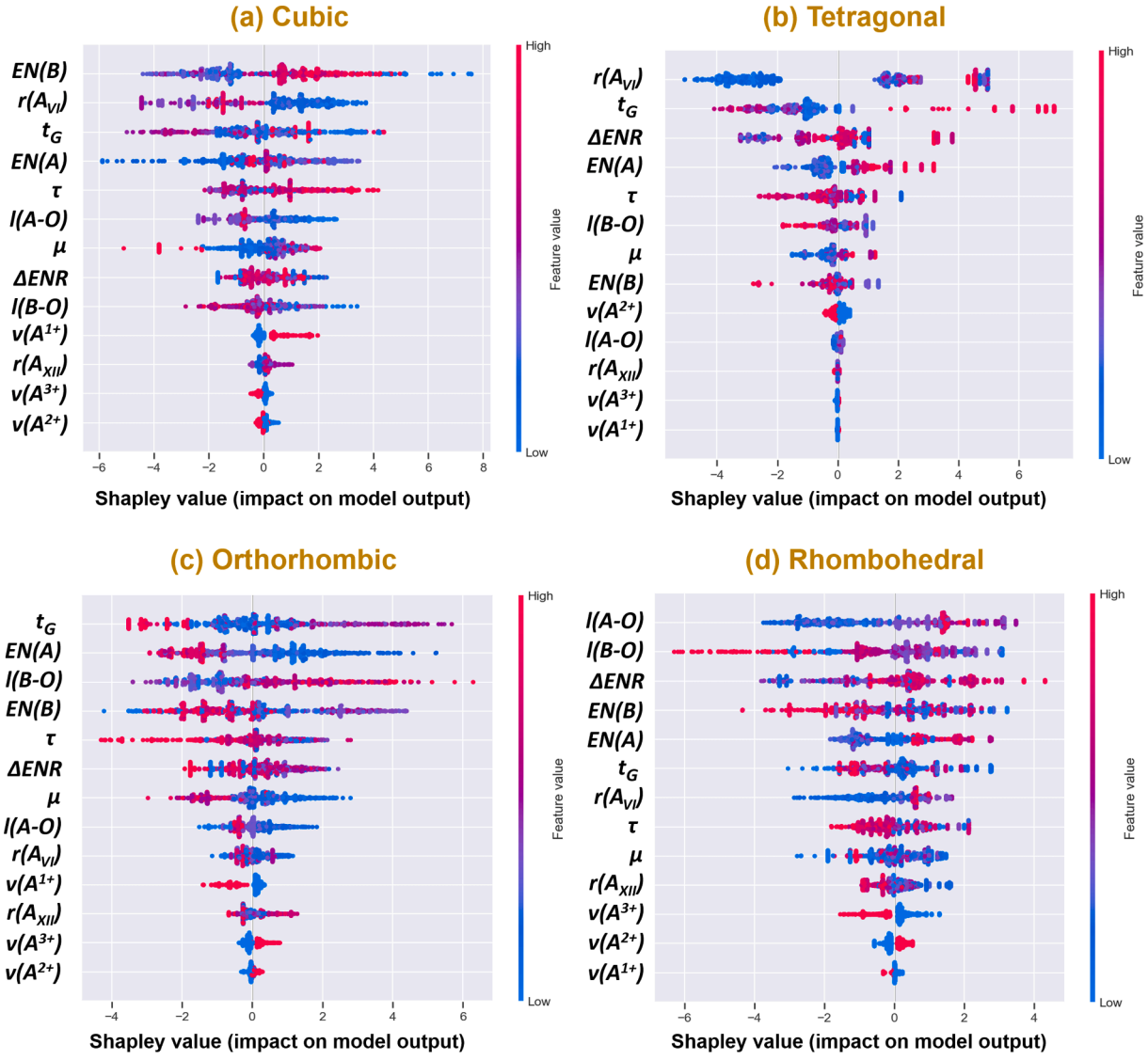

Fig. 6. SHAP summary plots for (a) cubic (b) tetragonal (c) orthorhombic and (d) rhombohedral systems. The 'x' axis has the Shapley values, and a positive value indicates it is more likely to predict that class. The features are listed on the 'y' axis in descending order of feature importance. The red (or blue) color indicates the high (or low) value of feature at that particular data point.

suggested by Goldschmidt, a t_G value of 1 indicates a perfectly cubic structure. This assumption is more valid in the structures, which shows a high degree of ionicity in the bonding [55]. As the B-site's electronegativity decreases, the ionic bonding between B-O pair increases due to the decreased difference in electronegativity between the B cation and oxygen anion. So, low *EN(B)* value means probability of prediction of cubic structure is high. This is consistent with the work of Bartel et al. [35], wherein it has been shown that high τ values are good at classifying perovskites.

(ii) For tetragonal structures, the low value of $r(A_{VI})$ has a negative impact. And, a high value of *EN(A)* has a positive impact on the probability of the prediction of this crystal structure. If we compare $BaTiO_3$ and $SrTiO_3$, a high value of $r(A_{VI})$ in $BaTiO_3$ shows tetragonal structure. In tetragonal structures, the covalent bonding of both A-O and B-O pairs are the deciding parameter. Even the phase transition from cubic to tetragonal takes place (*e.g.,* $PbTiO_3$) if the covalent interaction of A-O is more than that of B-O pair [56]. So, the high value of *EN(A)* is preferred in tetragonal structures as it shows a more covalent nature. And also, the high values of $t_G \geq 1$ result in the rattling of B cation inside the tetragonal unit cell causing its ferroelectric behavior.

(iii) For orthorhombic structures, the high value of l(B-O) and low value of EN(A) have a large positive impact on the probability of the structure's prediction. It is already shown from the work of Thomas et al. [57] that closely coordinated B cations (high value of *l(B-O)*) and over-sized A cations in ABO_3 structure lead to orthorhombic structures. Usually, low symmetry structures (orthorhombic) show higher band gaps [58]. This is due to the low *EN(A)* or the higher electronegativity difference between A cation and oxygen anion, which makes more ionic bonds and results in higher band gaps [59].

(iv) For rhombohedral systems, the high value of *ΔENR* has a positive impact, and a low value of *l(A-O)* has a negative impact on the probability of the prediction of its crystal structure. Here, we are proposing that the high value of *ΔENR* shows positive results in the classification of rhombohedral systems. From the work of Thomas et al. [60], the ratio of V_A/V_B (V_A is the volume of AO_{12} polyhedra, and V_B is the volume of BO_6 octahedra) found to be a reasonable indicator to classify rhombohedral structures. The high value of V_A/V_B minimizes the repulsions of O-O pairs' repulsion in AO_{12} polyhedra, resulting in high values of *l(A-O)*. In $A^{1+}B^{5+}O_3$, $A^{2+}B^{4+}O_3$ and $A^{3+}B^{3+}O_3$ perovskite structures, there is a high value of V_A/V_B (high value of *l(A-O)*) shows rhombohedral systems. Therefore, the low value of *l(A-O)* shows the negative effect of the rhombohedral structure classification.

From this SHAP analysis, we can analyze how each feature is related to the crystal structure's prediction probability (Ref: Fig. 7). Using the Shapley values from a materials design standpoint, one may try to alter the most significant features to get a particular crystal structure. Some more points to consider here while using our results for materials design would be that crystal structure may undergo phase transitions with temperature and pressure. Our model will work well for both experimental and high-throughput DFT calculated ABO_3 perovskites ($T = 0$ K and $P = 0$ bar) [8]. The experimental dataset may also be imperfect as compounds can have different crystal structures as a function of different synthetic conditions [35]. Here, our machine learning model (Light GBM) used its existing data (in literature and high-throughput DFT calculations) and classified perovskites into its respective crystal systems. From this study, new perovskite compounds (apart from known experimental reports) and its predicted crystal systems are reported, and they have implications to current activity in ceramic science and engineering, and solid-state inorganic chemistry. Our work also has implications for structure–property correlations in the disciplines mentioned.

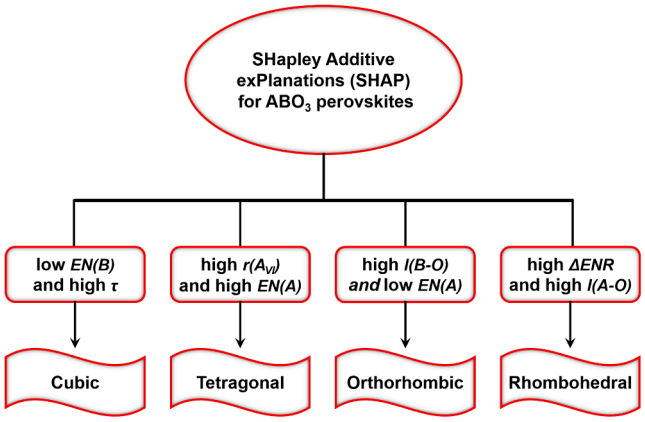

Fig. 7. SHAP summary plots show how input feature's magnitude affects the probability of prediction of the particular crystal system.

6. Conclusion

Out of 5329 ABO_3 perovskites, 675 compounds have been classified by using a Light GBM model. The model can, therefore, be used as a preliminary, fast and inexpensive method to classify perovskites into their respective crystal systems (*i.e.,* cubic, tetragonal, orthorhombic, and rhombohedral). The prediction accuracy is as high as 80.3%. In addition to this, the feature importance graph and Shapley value analysis are done to find the relationship between features used and the crystal structures predicted. Using the Shapley values reported, further work may be carried out wherein features are duly engineered in an attempt to pursue a target crystal structure. This renders this work directly relevant to the rational design of perovskites.

CRediT authorship contribution statement

Santosh Behara: Conceptualization, Methodology, Writing - original draft, Software, Formal analysis, Visualization. **Taher Poonawala:** Methodology, Writing - original draft, Software, Validation, Visualization. **Tiju Thomas:** Supervision, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the Department of Science and Technology, Government of India for supporting this work through EDD1819042DSTXNILE and SOL1819001DTSXHOCX. Santosh Behara also acknowledges the support from HTRA fellowship.

Data availability

The raw and processed data required to produce these finding are downloaded from Supplementary Information Files 1 and 2.

Supplementary Information File 1 (SIF-1): List of all 5329 ABO_3 perovskite-type oxides (Microsoft Excel Worksheet (.xlsx)).

Supplementary Information File 2 (SIF-2): List of all 675 ABO_3 perovskite-type oxides used in this study (Microsoft Excel Worksheet (.xlsx)).

Supplementary Information File 3 (SIF-3): List of all incorrectly classified crystal structures and their corresponding reported crystal structures by Emery et al. [8] (Microsoft Excel Worksheet (.xlsx)).

Supplementary Information File 4 (SIF-4): (.doc) file contains.

1. Model parameters used in this study.
2. Supplementary figures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.commatsci.2020.110191.

References

[1] A.S. Bhalla, R. Guo, R. Roy, The perovskite structure – a review of its role in ceramic science and technology, Mater. Res. Innov. 4 (1) (2000) 3–26.
[2] M. Johnsson, P. Lemmens, Crystallography and Chemistry of Perovskites, in: H. Kronmüller, S. Parkin (Eds.), Handbook of Magnetism and Advanced Magnetic Materials, John Wiley & Sons Ltd, Chichester, UK, 2007, p. p. hmm411., https://doi.org/10.1002/9780470022184.hmm411.
[3] P.M. Woodward, Octahedral tilting in perovskites. I. Geometrical considerations, Acta Crystallogr. B Struct. Sci. 53 (1) (1997) 32–43.
[4] G. Shirane, H. Danner, R. Pepinsky, Neutron diffraction study of orthorhombic BaTi O 3, Phys. Rev. 105 (1957) 856–860, https://doi.org/10.1103/PhysRev.105.856.
[5] M.W. Lufaso, P.M. Woodward, Jahn–Teller distortions, cation ordering and octahedral tilting in perovskites, Acta Crystallogr. B Struct. Sci. 60 (1) (2004) 10–20.
[6] P.M. Woodward, Octahedral tilting in perovskites. II. Structure stabilizing forces, Acta Crystallogr. B Struct. Sci. 53 (1) (1997) 44–66.
[7] P.M. Woodward, Structural distortions, phase transitions, and cation ordering in the perovskite and tungsten trioxide structures, (1996).
[8] A.A. Emery, C. Wolverton, High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites, Sci. Data 4 (1) (2017), https://doi.org/10.1038/sdata.2017.153.
[9] K. Bae, D.Y. Jang, H.J. Choi, D. Kim, J. Hong, B.-K. Kim, J.-H. Lee, J.-W. Son, J. H. Shim, Demonstrating the potential of yttrium-doped barium zirconate electrolyte for high-performance fuel cells, Nat. Commun. 8 (2017) 14553.
[10] S.H. Morejudo, R. Zanón, S. Escolástico, I. Yuste-Tirados, H. Malerød-Fjeld, P. K. Vestre, W.G. Coors, A. Martínez, T. Norby, J.M. Serra, C. Kjølseth, Direct conversion of methane to aromatics in a catalytic co-ionic membrane reactor, Science 353 (6299) (2016) 563–566.
[11] A. Soldatov, Crystal structure and possible superconductivity in BaBiO3–KBiO3 system outside the cubic phase, Physica B 284-288 (2000) 1059–1060.
[12] T. Kako, N. Kikugawa, J. Ye, Photocatalytic activities of AgSbO3 under visible light irradiation, Catal. Today 131 (1-4) (2008) 197–202.
[13] S. Thirumalai, B.P. Shanmugavel, Microwave assisted synthesis and characterization of barium titanate nanoparticles for multi layered ceramic capacitor applications, J. Microw. Power Electromagn. Energy. 45 (3) (2011) 121–127.
[14] D. Dimos, C.H. Mueller, Perovskite thin films for high-frequency capacitor applications, Annu. Rev. Mater. Sci. 28 (1) (1998) 397–419.
[15] C. Ye, T. Tamagawa, P. Schiller, D.L. Polla, Pyroelectric PbTiO3 thin films for microsensor applications, Sens. Actuators, A 35 (1992) 77–83.
[16] S.H. Chan, A review of anode materials development in solid oxide fuel cells, J. Mater. Sci. 39 (2004) 4405–4439.
[17] B.J. Prakash, B.H. Rudramadevi, S. Buddhudu, Analysis of ferroelectric, dielectric and magnetic properties of GdFeO 3 nanoparticles, Ferroelectr. Lett. Sect. 41 (4-6) (2014) 110–122.
[18] M. Baldini, T. Muramatsu, M. Sherafati, H.-K. Mao, L. Malavasi, P. Postorino, S. Satpathy, V.V. Struzhkin, Origin of colossal magnetoresistance in LaMnO 3 manganite, Proc. Natl. Acad. Sci. USA 112 (35) (2015) 10869–10872.
[19] R. Diehl, G. Brandt, Crystal structure refinement of YAlO3, a promising laser material, Mater. Res. Bull. 10 (2) (1975) 85–90.
[20] T. Fukuda, Y. Uematsu, Preparation of KNbO3 single crystal for optical applications, Jpn. J. Appl. Phys. 11 (1972) 163.
[21] J. Wu, Z. Fan, D. Xiao, J. Zhu, J. Wang, Multiferroic bismuth ferrite-based materials for multifunctional applications: Ceramic bulks, thin films and nanostructures, Prog. Mater Sci. 84 (2016) 335–402.
[22] P.K. Panda, B. Sahoo, PZT to lead free piezo ceramics: A review, Ferroelectrics 474 (1) (2015) 128–143.
[23] S. Behara, L. Ghatti, S. Kanthamani, M. Dumpala, T. Thomas, Structural, optical, and Raman studies of Gd doped sodium bismuth titanate, Ceram. Int. 44 (11) (2018) 12118–12124.
[24] H. Imai, T. Tagawa, Oxidative coupling of methane over LaAlO3, J. Chem. Soc., Chem. Commun. (1) (1986) 52, https://doi.org/10.1039/c39860000052.
[25] D.M. Giaquinta, H.-C. zur Loye, Structural predictions in the ABO3 Phase diagram, Chem. Mater. 6 (4) (1994) 365–372.
[26] C. Li, K.C.K. Soh, P. Wu, Formability of ABO3 perovskites, J. Alloy. Compd. 372 (1-2) (2004) 40–48.
[27] H. Zhang, N.a. Li, K. Li, D. Xue, Structural stability and formability of AB O 3 -type perovskite compounds, Acta Crystallogr. B Struct. Sci. 63 (6) (2007) 812–818.
[28] A.A. Demkov, A.B. Posadas (Eds.), Integration of Functional Oxides with Semiconductors, Springer New York, New York, NY, 2014.
[29] R.D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Cryst. A 32 (5) (1976) 751–767.
[30] A.L. Allred, Electronegativity values from thermochemical data, J. Inorg. Nucl. Chem. 17 (3-4) (1961) 215–221.
[31] H. Chen, S. Adams, Bond softness sensitive bond-valence parameters for crystal structure plausibility tests, IUCrJ. 4 (2017) 614–625. https://doi.org/10.1107/S2052252517010211.
[32] I.D. Brown, Predicting bond lengths in inorganic crystals, Acta Crystallogr B Struct Crystallogr Cryst Chem. 33 (1977) 1305–1310. https://doi.org/10.1107/S0567740877005998.
[33] G. Pilania, P.V. Balachandran, C. Kim, T. Lookman, Finding new perovskite halides via machine learning, Front. Mater. 3 (2016), https://doi.org/10.3389/fmats.2016.00019.
[34] V.M. Goldschmidt, Die Gesetze der Krystallochemie, Naturwissenschaften 14 (21) (1926) 477–485.
[35] C.J. Bartel, C. Sutton, B.R. Goldsmith, R. Ouyang, C.B. Musgrave, L.M. Ghiringhelli, M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Sci. Adv. 5 (2) (2019) eaav0693, https://doi.org/10.1126/sciadv.aav0693.
[36] I.D. Brown, Chemical and steric constraints in inorganic solids, Acta Crystallogr. B Struct. Sci. 48 (5) (1992) 553–572.
[37] K. Potdar, T. S., C. D., A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers, IJCA. 175 (2017) 7–9. https://doi.org/10.5120/ijca2017915495.
[38] F.S. Galasso, R. Smoluchowski, N. Kurti, Structure, Properties and Preparation of Perovskite-Type Compounds International Series of Monographs in Solid State Physics, Elsevier Science, Burlington, 2013.
[39] A. Kumar, A.S. Verma, S.R. Bhardwaj, Prediction of formability in perovskite-type oxides~!2008-08-05~!2008-10-08~!2008-12-05~!, TOAPJ 1 (1) (2008) 11–19.
[40] Y. Takeda, F. Kanamura, M. Shimada, M. Koizumi, The crystal structure of BaNiO3, Acta Crystallogr B Struct Crystallogr Cryst Chem. 32 (1976) 2464–2466. https://doi.org/10.1107/S056774087600798X.
[41] Y. Ishikawa, S. Sawada, The study on substances having the ilmenite structure I. Physical properties of synthesized $FeTiO_3$ and $NiTiO_3$ Ceramics, J. Phys. Soc. Jpn. 11 (5) (1956) 496–501.
[42] K. Davis, Material Review: Alumina (Al2O3), School Doctoral Stud. Eur. Union J. (2010).
[43] Z. Wang, Z.C. Kang, Functional and smart materials: structural evolution and structure analysis, Springer Science & Business Media, 2012.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[45] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, 2017: pp. 3146–3154.
[46] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.
[47] S. Aleksovska, S. Dimitrovska, I. Kuzmanovski, Crystal Structure Prediction in Orthorhombic ABO 3 Perovskites by Multiple Linear Regression and Artificial Neural Networks, Acta Chim. Slov. 54 (2007).
[48] S. Behara, T. Thomas, Stability and amphotericity analysis in rhombohedral ABO3 perovskites, Materialia. 13 (2020), 100819.
[49] J. Elith, J.R. Leathwick, T. Hastie, A working guide to boosted regression trees, J Anim Ecol. 77 (4) (2008) 802–813.
[50] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, 2009, pp. 1–4.
[51] Jerome H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[52] H. Shi, Best-first decision tree learning, The University of Waikato, 2007.
[53] J. Kim, C.D. Scott, Robust kernel density estimation, J. Mach. Learn. Res. 13 (2012) 2529–2565.
[54] S.B. Cohen, E. Ruppin, G. Dror, Feature selection based on the shapley value, IJCAI (2005) 665–670.
[55] W. Travis, E.N.K. Glover, H. Bronstein, D.O. Scanlon, R.G. Palgrave, On the application of the tolerance factor to inorganic and hybrid halide perovskites: A revised system, Chem. Sci. 7 (7) (2016) 4548–4556.
[56] K. Singh, S. Acharya, D.V. Atkare, Qualitative analysis of tolerance factor, electronegativity and chemical bonding of some ferroelectric perovskites through MOT, Ferroelectrics 315 (1) (2005) 91–110.
[57] N.W. Thomas, The compositional dependence of octahedral tilting in orthorhombic and tetragonal perovskites, Acta Crystallogr. B Struct. Sci. 52 (1) (1996) 16–31.
[58] W.B. Park, S.U. Hong, S.P. Singh, M. Pyo, K.-S. Sohn, Systematic approach to calculate the band gap energy of a disordered compound with a low symmetry and large cell size via density functional theory, ACS Omega 1 (3) (2016) 483–490.
[59] M. Pazoki, T. Edvinsson, Metal replacement in perovskite solar cell materials: Chemical bonding effects and optoelectronic properties, Sustain. Energy Fuels 2 (7) (2018) 1430–1445.
[60] N.W. Thomas, A re-examination of the relationship between lattice strain, octahedral tilt angle and octahedral strain in rhombohedral perovskites, Acta Crystallogr. B 52 (1996) 954–960.