Solar Energy 267 (2024) 112209

Contents lists available at ScienceDirect

Solar Energy

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

Machine learning-aided discovery of bismuth-based transition metal oxide double perovskites for solar cell applications

Siddharth Sradhasagar, Omkar Subhasish Khuntia [1], Srikanta Biswal [1], Sougat Purohit, Amritendu Roy [*]

School of Minerals, Metallurgical and Materials Engineering, Indian Institute of Technology, Bhubaneswar, Odisha 752050, India

ARTICLE INFO

Keywords:
Photovoltaics
Bandgap
Machine learning
Double perovskites
Transition metal oxide

ABSTRACT

Due to their importance in semiconductor device designing, especially in photovoltaic solar cells and light emitting diodes, methods that can promptly and reliably forecast material's bandgap (E_g) and its character, *viz.*, direct or indirect, are in demand. In this context, data-driven machine learning (ML) methodologies are considered promising. In the present work, several ML models were developed using easy-to-find instrumental variables such as unit-cell volume, structural parameters (a, b, c, α, β, γ), space group, number of constituent atoms, and standard atomic properties (*viz.*, atomic number, atomic mass, ionisation energy, electronegativity) to forecast the bandgap and its character for double perovskites. The LGBMRegressor and XGBClassifier models were identified to best predict the magnitude and nature of the bandgap with an accuracy of ~ 0.89 and 0.95, respectively. Subsequently, the above models were employed to predict the bandgap for novel bismuth-based transition metal oxide double perovskites. The accuracy of the present models, especially over the range of 1.2–1.8 eV, makes them particularly suitable for designing bismuth-based double perovskites for photovoltaic applications.

1. Introduction

Despite impressive conversion efficiency and easiness of synthesis, lead toxicity, material instability, and consequent performance degradation have been major deterrents towards the commercialization of lead halide-based perovskite solar cells [1]. This leaves a massive opportunity for designing novel, affordable, and efficient solar photovoltaic (PV) materials. In this regard, several materials have been explored in recent years such as Cs_2ScAgX_6 (X: I, Cl, Br) [2], Rb_2YInX_6 (X: Cl, Br, I) [3], X_2ScAuI_6 (X: Rb, Cs) [4] and K_2AgBiX_6 (X: Cl, Br) double perovskites [5].

Other viable alternative materials include transition metal oxide (TMO)-based perovskites and double perovskites, which have superior stability and innocuousness. Unfortunately, not much literature is available in the above domain, presumably due to unfavorable optoelectronic properties translating to a poor PV response in the above class of materials. There are, however, certain reports of oxide perovskites [6] and double perovskites [7] with decent PV performance. Thus, there may be many more similar oxide systems with even superior PV performance. In this regard, oxide double perovskites are more promising owing to their excellent structural stability and greater susceptibility to functional control [8]. Thus, in the present work, we attempt to design novel TMO-based PV materials with double perovskites as the preferred structural backbone.

Since the electronic structure of lead is critical to impart the desired PV conversion efficiency [9], it would have been desirable to include lead while designing TMO-based double perovskites for prospective PV applications. However, Pb^{2+} ions are toxic and should be avoided. Instead, isoelectronic and nontoxic Bi^{3+} ions were considered a better choice to design and develop environment-friendly TMO-based double perovskites. Therefore, we considered the following chemical formulae for designing prospective PV materials: A_2BBiO_6 and $Bi_2B'B''O_6$, where "A" could be alkali/ alkaline-earth/ transition/ rare-earth metal ions. At the same time, B, B' and B'' represent transition metal ions, which would be commensurate with the emergence of double perovskites structure. However, for A_2BBiO_6, B, in some cases, may include non-metals such as phosphorus, arsenic, etc. Nevertheless, a suitable material selection strategy should be adapted to fast-track exploration.

* Corresponding author.
E-mail address: amritendu@iitbbs.ac.in (A. Roy).
[1] OSK and SB contributed equally.

https://doi.org/10.1016/j.solener.2023.112209
Received 31 August 2023; Received in revised form 24 October 2023; Accepted 19 November 2023
Available online 28 November 2023

While all efficient lead-based halide perovskites demonstrate a host of favourable optoelectronic properties [9], the conversion efficiency is critically dependent on the optical bandgap of the material. Bandgap determines the onset of optical absorption, which decides the number of photo-excited charge carriers. It has been found that a direct bandgap of $\sim$ 1.5 eV results in optimum PV response, given that other parameters are kept constant. Thus, the most critical screening parameters for selecting solar-absorbing materials are the nature and magnitude of the material's bandgap. Therefore, the proposed bismuth-based double perovskites systems must possess a favorable bandgap close to $\sim$ 1.5 eV.

Conventional ways of measuring this intrinsic attribute accurately require cumbersome synthesis steps and expensive equipment [10]. On the contrary, theoretical prediction using techniques such as density functional theory (DFT) is time-consuming. It is also prone to underestimating the actual value [11] using usual approximations such as GGA and LDA. Therefore, a fast prediction of reasonably accurate bandgap data would be critical to screening and, thus, designing novel bismuth-based double perovskites.

The bandgap of materials has been routinely predicted using machine learning (ML) [12–18] algorithms that recognise significant patterns in bandgap values of thousands of compounds and can predict the bandgaps of new systems. For example, Gladkikh et al. employed ML to predict the bandgap of ABX_3 single perovskites using elemental properties [19]. Wu et al. utilised the random forest technique to evaluate the bandgap of inorganic compounds [20]. Zhuo et al. have used ML to forecast bandgaps in inorganic solids, achieving an $R^2 \sim 0.90$ with the support vector regression (SVR) model [18]. Pilania et al. forecasted bandgaps of double perovskites using linear least square regression with kernel ridge regression (KRR) [13]. Ward et al. achieved relatively high model prediction accuracy using a random forest approach with a categorised dataset based on the bandgap range and element group [21]. However, the partitioning of the materials decreased the model's robustness.

Additionally, categorising the bandgap type can provide valuable insights into these materials' electronic properties and potential applications. It would also help experimentalists to identify likely systems before using sophisticated experimental setups to crosscheck with the value already predicted using the ML algorithms. However, despite considerable progress in research on predicting the bandgap of materials using different ML techniques, very few works have implemented the above models in proposing new materials satisfying the criteria for prospective solar cell applications. Thus, keeping these points in mind, the present study develops regression and classification ML models for predicting the bandgap and its type (direct or indirect) for double perovskites using elemental features and crystal structure information collected from the Materials Project database (https://next-gen.materialsproject.org/) [22]. Since the goal is to employ relatively inexpensive characteristics to compute and apply in large prediction datasets, formation energy has not been used as a descriptor. Thus, by leveraging the available materials data, the proposed models facilitate the accurate prediction of bandgap values and their type in the optimum bandgap range of 1.2–1.8 eV [23], contributing to advancements in solar cell research. Finally, a list of novel Bi-based double perovskite oxides with their predicted bandgap values and types is prepared from a vast chemical space using the above-developed models.

2. Dataset and screening constraints used

Since bandgap data are critical in the present work, we chose the open-access structure–property database of Materials Project (MP) to extract the material's bandgap and crystal structure-related features using its application program interface (API). The DFT data used in the present work were generated using Vienna *Ab-Initio* Simulation Package (VASP) software [24–26]. In Generalized Gradient Approximation [27], GGA/GGA + U, the core electrons were treated using the projector augmented wave (PAW) [28,29] method with a kinetic energy cutoff of 520 eV. A k-point mesh of 1000/(number of atoms in the cell) was employed for all computations. Similarly, PAW was employed for r2SCAN exchange-correlational functional [30], and the energy cutoff was set to 680 eV. The Monkhorst-Pack method [31] was used for k-point meshing (with Γ-centred for hexagonal cells) in both GGA/GGA + U and r2SCAN, while the tetrahedron approach was utilised for k-point integration. The energy difference for ionic convergence was fixed at 0.0005 × no of atoms in the cell. All computations were performed at 0 K, with spin polarisation on and magnetic ions in a high-spin ferromagnetic initialisation. Alternate spin states were used for some materials. The input structures were taken from sources like the Inorganic Crystal Structure Database (ICSD) [32]. The structure's cell and atomic positions were then relaxed in two consecutive runs. For materials exhibiting multiple crystal structures, all unique structures identified by an affine mapping technique [33] were examined. The information obtained is then analysed using the Pymatgen (Python Materials Genomics) package, which uses the MPrester module to pool the HTTP connection. At this juncture, it should be noted that the bandgap data calculated using conventional DFT techniques, such as those using GGA and LDA, brings in considerable error, and it would be desirable to have either experimental data or calculated data using techniques such as GW or hybrid-functionals. Without any such database for double perovskites, the present work uses DFT-calculated data and intends to improve upon it by adding more accurate datasets in the future.

Using MP API, we have extracted 4735 DFT calculated data of double perovskite materials having a general formula of ABC_2D_6. The bandgap data comprised conducting metals (bandgap $<$ 0.1 eV) and insulator non-metal double perovskites. Thus, 1647 metals and 3088 non-metal double perovskites are present in the current data set. Fig. 1(a) and (b) show the bin count of corresponding bandgaps of all double perovskites and all double perovskites except metals. It illustrates the wide range of bandgap values of materials used for training the models, with most of the compounds ranging from 0.1 to 4 eV. However, to understand the effect of metals on bandgap and bandgap type predictions, two different materials datasets were made, one with metals and the other without metals. The former has 1125 direct and 3610 indirect bandgap materials, with the latter having 1093 direct and 1995 indirect bandgap materials (Fig. 2). The dataset of the latter case has a uniform distribution of compounds with bandgap values in the range of 0.5–2.5 eV, suitable in utilising it to find potential double perovskite materials.

3. Bandgap-based computations

3.1. Descriptor selection

It has been observed that the bandgap of materials is a strong function of its composition and crystal structure [16]. Thus, the input to the models used the structural (a, b, c, α, β, γ in conventional settings) and elemental descriptors/features of each compound. We have opted not to use formation energy as a descriptor since we aim to use comparatively inexpensive features to calculate to enable them to be implemented in large prediction datasets. Thus, using the pymatgen.core.periodic_table module of pymatgen.core.package, several unique elemental descriptors of the constituent ions of ABC_2D_6 double perovskites, such as radius (atomic, ionic, covalent), atomic mass, melting point, hardness, molar volume, thermal conductivity, period number, electronegativities (Pauling, Mulliken, Allen), valence electrons, shell electrons (s, p, d, f), etc. have been used to help generalise the ML model and enable it to rapidly and accurately predict the bandgap of any composition for different types of compounds. Few other elemental descriptors, such as the ionic highest occupied molecular orbital (HOMO) level, ionic lowest unoccupied molecular orbital (LUMO) level, and ionic HOMO-LUMO gap taken from ref [34], have also been included. He [35], Huang [15], and Zhuo et al. [18] also used similar elemental properties in their work. The entire list of features used is given in the supporting information (Appendix A). However, to improve the variety of training data

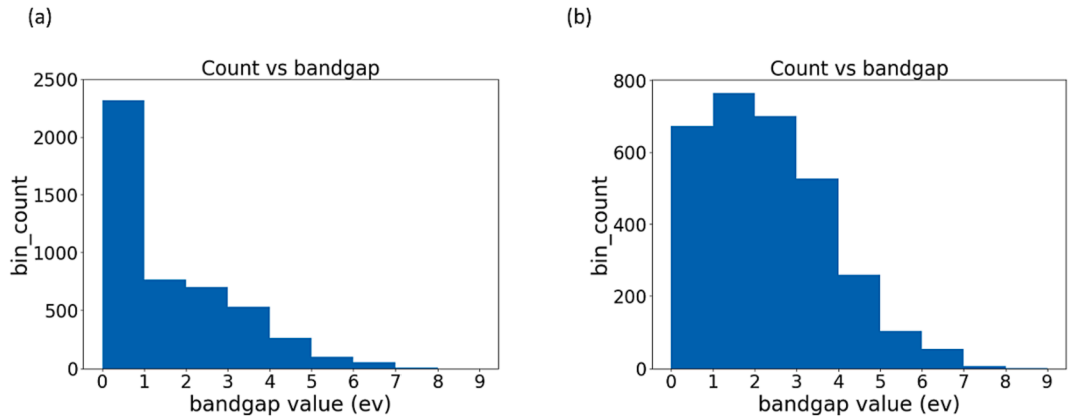

Fig. 1. The bin count of corresponding bandgaps of (a) all double perovskites. (b) all double perovskites except metals.

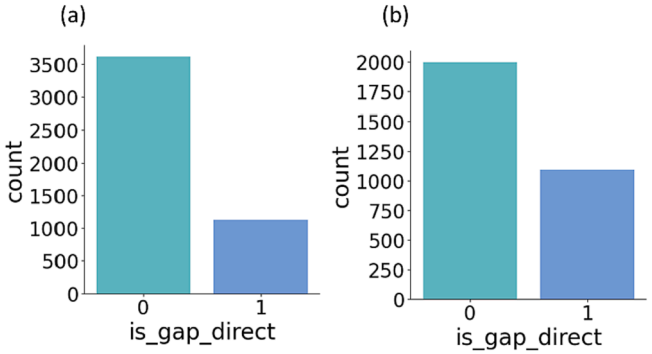

Fig. 2. The count of the number of direct (1) and indirect (0) bandgap type materials for (a) all double perovskites. (b) all double perovskites except metals.

and reduce overfitting, a simple data augmentation method has been adopted [18,36], increasing the number of descriptors to 303 by describing the compositions of the compounds using several mathematical formulae involving a simple combination of the constituent elemental properties (e.g., mean, standard deviation) or the extremes (max, min). Each combination introduces new possible characteristics that explicitly define the target variable. All the compounds and their feature vectors are then stored in the Pandas data frame.

3.2. Preprocessing and feature engineering of the dataset

To develop accurate models by gaining a deep understanding of the dataset, exploratory data analysis (EDA) was performed. Also, feature engineering techniques improve the dataset's quality and optimise the model's performance. It includes dealing with both datasets, in the present case, one with metals and the other without metals, which have an imbalanced distribution of direct and indirect bandgap materials. This is appropriately handled using the Synthetic Minority Oversampling Technique (SMOTE) [37]. It balances the dataset by creating synthetic samples for the minority class (in this case, direct bandgap), increasing its representation and lessening the impact of class imbalance on model training.

Furthermore, the data from the dataset cannot be used directly in the original format as they may contain erroneous, missing, or inconsistent data points that must be homogenised and cleaned before use. This step is crucial in building more accurate predictors through ML. For this, the duplicate entries (identical chemical formulas and space groups) were initially removed to avoid biasing the fitted model. Some feature columns with the same numeric values were also dropped as they give no unique information for each row or compound. The distribution plots of all independent numerical variables were done to provide a concise summary of the distribution of the dataset. A boxplot or a box-and-whisker plot was used, which displayed key descriptive statistics such as the median, quartiles, and potential outliers lying outside the whisker. This visualisation analysis helped us understand the data's shape, central tendency, spread, and skewness. These are valuable for EDA as they help identify patterns, detect anomalies, make feature engineering and form modelling decisions.

Since the total percentage of missing values upon inspection in both datasets was found to be very small (~1.47 % with metals and ~ 1.59 % without metals), they could be dealt with pipelines that use different imputing methods like replacing missing values with KNN imputation, mean imputation, median imputation, filling with constant values (e.g., zero) and with the 'Multiple Imputation by Chained Equation' (MICE) approach to impute missing data by predicting them using other features from the dataset. Apart from the earlier imputing methods, the principle component analysis or PCA with imputing constant zero was also implemented for bandgap type classification only. Kauwe et al. [16] also employed a similar mean imputation strategy to replace missing values in their work. The different imputation techniques also resulted in different datasets containing values of various magnitudes of features that must be scaled within a single dimension, making them parsable without any information loss. This step is crucial for handling data with outliers and different distribution patterns. This was done using the Robustscaler package of Sklearn [38].

3.3. Training set

The entire materials dataset is randomly split (e.g., by allocating a random seed and rearranging the data set) into having training to a testing ratio of 80:20. This means 80 % of the materials of the dataset will be used for training, with the rest (20 %) assessing the model's performance. Dataset rearrangement is necessary to make a similar presentation of all sorts of data, helping build a robust model for achieving higher accuracy. It will also ensure that no data that has been trained gets tested, which may inaccurately lead to higher accuracy.

3.4. Model selection

3.4.1. Based on regression for bandgap prediction

From an ML point of view, bandgap prediction represents a regression task that aims to find a linear relation $y = f(x)$ between an input vector x as independent variables (widely referred to as features or descriptors) and an output y as dependent variable (widely referred to as target, bandgap in the present case). The quality of the bandgap estimation depends on the type and size of the training set, the selection of features, the regression methods and their corresponding model parameters used [10]. As mentioned earlier, the different datasets were then dealt with typical classical and statistical ML regression models,

such as LinearRegression, Ridge, and BayesianRidge. These algorithms were further enhanced by boosting ensemble techniques, which decrease variance or bias by combining multiple model predictions [39]. The ensemble models used in the current work include GBR, XGBoost, LightGBM (LGBM), and CatBoost. All the regression models were from the Sklearn package [16].

3.4.2. Based on classification for bandgap-type prediction

The bandgap type prediction represents a classification task that aims to distinguish the bandgap of the double perovskite material as direct (1) or indirect (0). For this purpose, typical classification models, such as logistic regression, K-neighbours classifiers, and decision tree classifiers, were implemented. Also, these algorithms were further enhanced by bagging (RandomForestClassifiers) and boosting ensemble techniques like GradientBoostingClassifier, XGBClassifier, LGBMClassifier, AdaBoostClassifier, and CatBoostClassifier. All the classifiers were from the Sklearn package [16].

3.5. Model performance evaluation

Upon employing different algorithms, their performance was evaluated and compared to identify the best model for regression and classification tasks.

3.5.1. Regression

To check whether the developed model captures the underlying patterns of the data, it is vital to evaluate the performance of the regression task using computed test metrics such as R^2 (Coefficient of determination), (root) mean squared error for (RMSE/MSE) and mean absolute error (MAE) [40]. In the present work, all four have been calculated as follows:

$$R^2\left(y, y_i^{'}\right) = 1 - \left(\sum_{i=1}^{N}\left(y_i - y_i^{'}\right)^2 / \sum_{i=1}^{N}(y_i - \bar{y})^2\right) \quad (1)$$

$$\mathrm{MSE} = \left(\sum_{i=0}^{N}(y_i - y_i^{'})^2\right)/N \quad (2)$$

$$\mathrm{RMSE} = \sqrt{MSE} \quad (3)$$

$$\mathrm{MAE} = \left(\sum_{i=0}^{N}|y_i - y_i^{'}|\right)/N \quad (4)$$

where N is the number of data points, $y_i/y_i^{'}$ are the actual/model-predicted values and $\bar{y}$ is the mean of the data points, respectively. Furthermore, to detect overfitting, the training data is exposed to a better statistical technique of k-fold cross-validation (CV) [25]. Here, the dataset is divided into k subsets (known as folds, 5 in this case), and the model is trained on the combined data of k-1 subsets and then evaluated on the k^{th} subset for each k of the data subsets k = 1,2,3… k. The k prediction error results are then averaged to measure the model's genuine predictive performance.

3.5.2. Classification

The performance measuring parameters for classification calculated in the present work include accuracy, precision, recall, F1 score, and AUC (Area under the ROC or receiver operating characteristic curve) [41]. Their formulations are given below:

$$\mathrm{Accuracy} = (TP+TN)/(TP+FN+FP+TN) \quad (5)$$

$$\mathrm{Precision} = (TP)/(TP+FP) \quad (6)$$

$$\mathrm{Recall} = (TP)/(TP+FN) \quad (7)$$

$$\mathrm{F1\ score} = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (8)$$

$$\mathrm{AUC} = (S_P - N_P(N_N+1)/2)/(N_P N_N) \quad (9)$$

where True positive (TP) refers to the correct reporting of a class (e.g., predicted bandgap type that turns out to be correct), and a false positive (FP) is the incorrect reporting of a class (e.g., predicted bandgap type that turns out to be wrong). Similar definitions can be deduced for false negative (FN) and true negative (TN) [42]. S_P represents the sum of all positive examples ranked, while N_P and N_N denote the number of positive and negative examples, respectively. The predicted and actual result labels are then assessed by a 2 × 2 confusion matrix, frequently employed in ML for classification tasks. It offers a detailed description of the model's predictions and the actual values from the dataset. Also, the ROC-AUC curve measures a classifier's ability to discriminate between positive and negative classes across threshold settings. The ROC curve shows the tradeoff between the true positive rate (TPR) (sensitivity) and the false positive rate (FPR) (1-specificity) for different classification thresholds. The AUC, the area under the ROC curve, is a single numerical measure of the model's performance. Moreover, it is not affected by class imbalance in the dataset [17].

4. Results and discussion

4.1. Bandgap prediction model performance

The complete results of the bandgap prediction by different ML models on both datasets (without and with metals) with different imputing methods are given in Table A1 (without metals) and A.2 (with metals) of the supporting information, respectively. Upon comparison of the different performance metrics of ML regression models for bandgap prediction, it is found that in both cases, the classical and statistical ML models (LinearRegression, Ridge, and BayesianRidge) did not show satisfactory performance, which significantly improved upon using ensemble techniques (GBM, XGBoost, LightGBM (LGBM) and CatBoost). The best result was obtained for the dataset, which contained no metals and where the missing values were imputed with the KNN imputation method. LGBMRegressor was the corresponding ML algorithm that performed best with this dataset. Further optimisation of the hyperparameter values of the LGBMRegressor using the grid search method did not yield any improvement in the evaluation metrics, prompting us to consider the default hyperparameter values the best. Fig. 3. shows the various performance metrics on the training and the test set by the

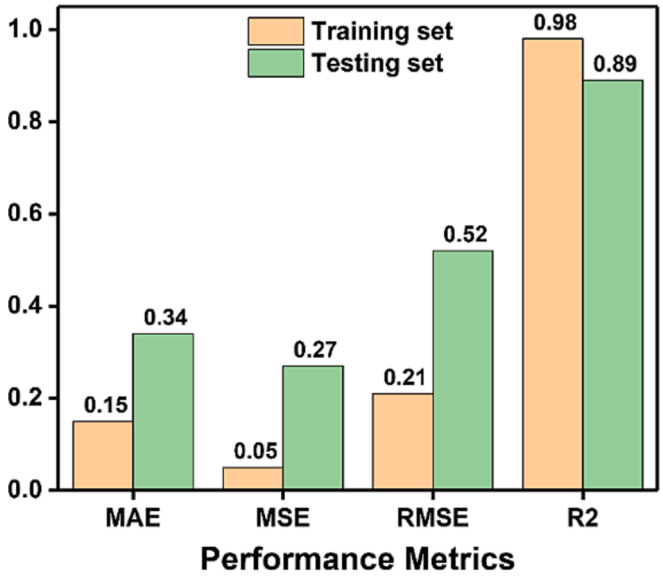

Fig. 3. Various performance metrics of the LGBMRegressor model on the training and the test set.

LGBMRegressor model. These metrics indicated that the LGBMRegressor model achieved a comparatively minimum mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and maximum R^2 on both the training and test datasets. It also showed a comparable CV performance of ~ 85 %.

As shown in Table A1 and A2 of the supporting information, irrespective of the imputation method and ML model used, the R^2 of the dataset, which contained both metal and non-metal, has a value of ~ 0.87, slightly lower than the ~ 0.89 obtained in the dataset containing only non-metals. This could be due to the models getting affected by weights and biases induced by ~ 35 % of metals present in the dataset during training, leading to underestimation of the bandgaps. Nevertheless, our model's performance is comparable to that of the gradient-boosting decision tree model based on a property-labelled materials fragment descriptor, which had an R^2 of 0.90 and RMSE of 0.51 eV [43] and a crystal graph convolutional neural network (CGCNN) using a descriptor set that relies on atomic coordinates derived from the generation of crystal graphs which has an MAE of 0.388 [44]. Fig. 4(a). shows a scatter plot with predicted bandgap values on the y-axis and the DFT calculated actual values on the x-axis. It shows that most data points are closely aligned to the ideal diagonal line, where each prediction perfectly matches the measured value. The error distribution analysis was further done to study the bandgap prediction deviation. It revealed that most errors have a difference of ± 2 eV, as depicted in the inset of Fig. 4(b). The present work assumes that the prediction is deemed highly inaccurate when the percentile error exceeds 25 % (Note: The percentile error is calculated using the formula [(actual bandgap – predicted bandgap) × 100/actual bandgap)]. Here, the total size of the test case used for prediction is 618, of which, 178 observations exhibited a percentile error above 25 %. This represented approximately 29 % of the total observations with errors above this threshold. Fig. 4(c) illustrates that most errors fall in the bandgap range of 0.1–0.7 eV with 85 observations of the 178 total observations. This constituted approximately 14 % of test materials have errors within this range. Similarly, the fact that just 5 % of the total test materials (32 observations) exhibit erroneous bandgap predictions within the optimum bandgap range of 1.2–1.8 eV [23] indicates that bandgap predictions in this range are reliable. This is particularly significant from the perspective of the scope of the present work for predicting the bandgap of prospective photovoltaic materials.

The electronic bandstructure of bulk-crystalline materials plots a relationship between the energy and momentum of the carriers. This relationship is complex in real materials, with several factors affecting the bands' dispersion vis-à-vis magnitude and bandgap characteristics. The nature of the correlation between the bonding atoms (ions) and the chemical identity of the bonding atoms are two critical factors. The relative positions of electronic energy levels of the bonding ions critically determine the band gap. Furthermore, the nature of the bond is also vital. Generally, a large difference in the electronegativity of the bonding atoms (leading to predominantly ionic bonding) increases the bandgap. The periodic manner in which the molecules are arranged in three dimensions is also critical in determining the electronic structure. This is manifested by the dependence of electronic structure on the crystal symmetry and lattice parameters. It has been noticed that bandgap in semiconductors is inversely proportional to the inter-atomic bond distance. In addition, the finite size of the crystal also affects the bandstructure, especially when the crystal size is really small where quantum confinement sets in. In developing the ML models, the features used are associated with the above physical parameters in one way or the other.

Thus, these relationships between the different features and bandgap have been explored by computing the Pearson correlation coefficients. Table 1 lists the top twenty features with positive and negative correlations to bandgap in the present dataset. The difference of Waber radii, minimum Ionic HOMO-LUMO gap, the difference in the first ionisation energy, and the difference of Mulliken electronegativity are some elemental properties of compounds that possess relatively higher positive correlation values. The electronegativity also takes into account both the ionisation and electron affinity energy [36]. If a compound contains elements with high electronegativity, its propensity to attract electrons is also high [45]. Consequently, it is tough for the electrons in the valance band to go to the conduction band. And as the electronegativity difference rises, the corresponding bandgaps also rises. Electronegativity also explains bond polarity or the dipole moment between atoms [46]. As a result, the electronegativity of elements in the system has a significant impact on its band structure. Huang et al. also reported the importance of electronegativity and valence in the bandgap prediction of materials [15]. Talapatra et al., in their work, also mentioned HOMO and LUMO energies, ionisation energy IE, and electronegativity as some of the essential atomic descriptors affecting the bandgap of double perovskite oxides [16]. Other features, such as the maximum number of d electrons, have also been reported as critical to bandgap prediction in Zhuo's [18] and Wang et al. [36] study. Gladkikh et al. determined electronegativities, electron affinities, ionisation energies,

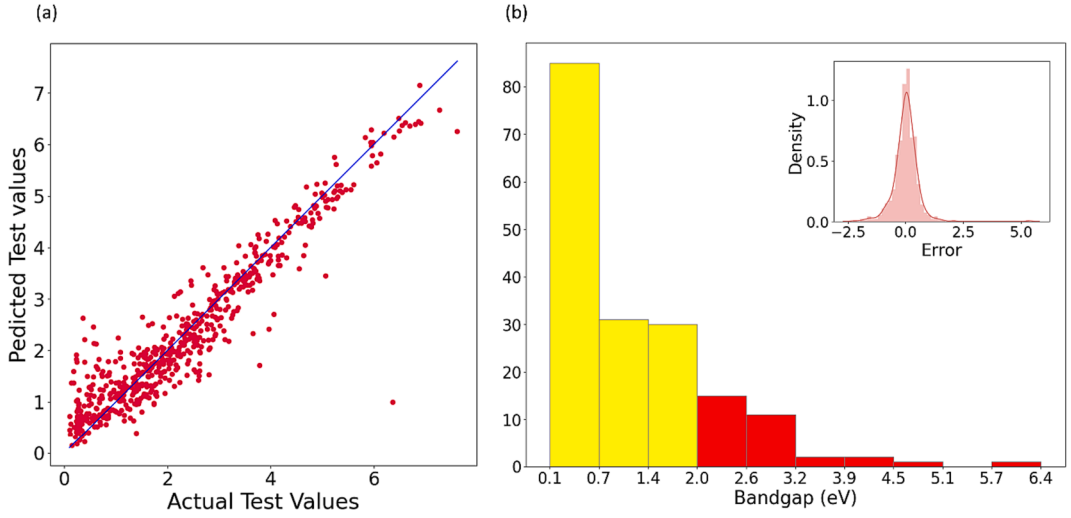

Fig. 4. (a) LGBMRegressor predicted bandgap E_g vs DFT predicted bandgap E_g for only non-metals (i.e., $E_g \geq 0.1$ eV as defined in the present work). (b) Histogram showing the distribution of the 178 observations according to bandgap values, which exhibited percentile error above 25 %. Inset showing most bandgap errors having a difference of ± 2 eV.

Table 1
Top twenty features according to positive and negative importance score.

Serial No	Feature Name	Positive Importance score
1	diff_Values of Radii metal (Waber)	0.37
2	min_Ionic HOMO-LUMO gap / eV	0.35
3	diff_Mulliken EN	0.34
4	diff_Distance from valence electron (Schubert)	0.33
5	std_average_ionic_radius	0.32
6	diff_First ionisation energy (kJ/mol)	0.32
7	min_Volume of atom (Villars. Daams)(10^{-2} nm^3)	0.31
8	diff_Ionic radius (Å)	0.31
9	diff_Atomic radius (Å)	0.31
10	max_Distance from valence electron (Schubert)	0.31
11	max_Metallic valence	−0.49
12	avg_Density (g/mL)	−0.47
13	mean_density_of_solid	−0.46
14	diff_Metallic valence	−0.45
15	avg_Metallic valence	−0.44
16	max_Number of d electrons	−0.44
17	diff_Number of d electrons	−0.44
18	mean_youngs_modulus	−0.37
19	mean_brinell_hardness	−0.36
20	mean_bulk_modulus	−0.35

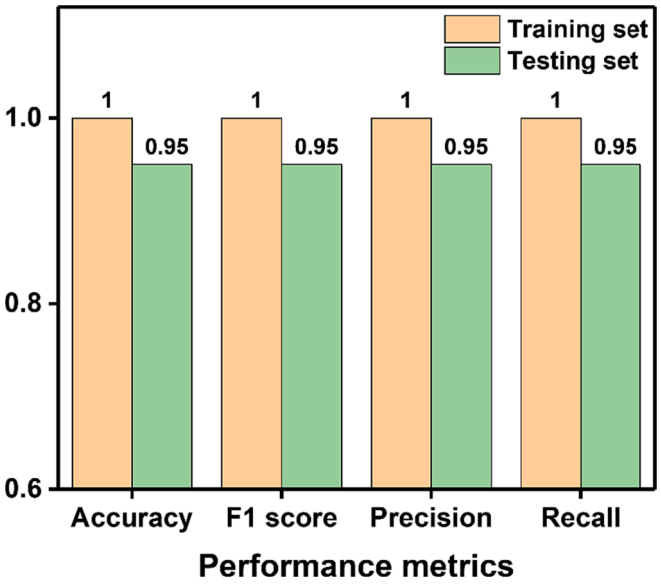

Fig. 5. Various performance metrics on the training and the test set by the XGBClassifier model.

and atomic radii of the constituents as important descriptors affecting the material's bandgap [19]. Also, the bulk modulus is important as it is determined by atom compressibility, influencing the extent of valence atomic orbital overlap and, hence, the material's bandgap [47]. However, it is found that the model performance deteriorates when only certain features are used for model training and testing. Hence, all features have been included in the LGBMRegressor model for bandgap prediction. Zhuo et al. also used a similar approach to predict the bandgap of inorganic solid materials [18].

4.2. Bandgap type prediction model performance

Using the same feature set as the regression model, the complete results of the classification of the bandgap type into direct (1) and indirect bandgap (0) classes by the different ML models on both datasets (without and with metals) with different imputing methods are given in Table A3 and A4 of the supporting information, respectively. Like the regression work, in both cases, the various performance metrics of ML classification models found that the ensemble models perform better than the conventional classification models. However, XGBClassifier performed the best on the dataset, which contained both metals and non-metals and where the missing values were imputed with the mean. Like regression, hyperparameter optimisation of the XGBClassifier using the grid search method was ineffective, suggesting that the default parameter values are more suited to the present problem. Fig. 5. shows the various performance metrics on the training and the test set by the XGBClassifier model. The accuracy in the present case is measured in terms of true positive and false negative predictions, and the XGBClassifier model achieved maximum accuracy and balanced performance in terms of F1 score, precision, and recall on both the training and test datasets. It also showed an excellent CV performance of ~ 94 %. The result obtained here is also better than the direct–indirect classifier model developed by Weston et al., which had an accuracy of ~ 89 % [23].

The principal diagonal of the confusion matrix, shown in Fig. 6(a), supports the model's ability to classify bandgap types. The instances in each row and column of the matrix correspond to the actual and predicted classes, respectively. This matrix helps calculate several performance metrics, including accuracy, precision, recall, and F1 score, which provide light on the model's effectiveness and capacity to classify classes accurately. As can be seen, 1343 materials (True-positive: 658 and True-negative: 685 materials) out of 1415 materials have been correctly predicted, showing an accuracy of ~ 0.95. The rest, 5 % or 72 materials (False-negative: 39 and False-positive: 32 materials), were incorrectly predicted, meaning only one out of twenty predictions went wrong. Overall, the confusion matrix demonstrates minimal bias, and the model's success rate for correctly classifying the bandgap type as direct or indirect is nearly identical.

In the ROC-AUC curve (Fig. 6(b)), an AUC score of ~ 0.95 is achieved, which is closer to 1, indicating an excellent performing model with a higher ability to classify instances accurately. Thus, the classification model works significantly well for the distinction of classes.

Thus, our results indicated that the classification models achieved better accuracy when the metal dataset was included, while the regression models yielded better results when the metal data were excluded. Based on these observations, we will focus on presenting the experiments for classification models using the metal dataset and regression models, excluding the metal data for predicting the bandgap and bandgap type of unknown dataset of Bi and O containing double perovskite materials, enabling an assessment of their generalisation capabilities.

4.3. Prediction of bandgap and bandgap type for novel Bi-based TMO double perovskites

This section evaluates the performance of the present models in predicting the bandgap and bandgap type of Bi-based TMO double perovskites. The reason for selecting Bi-based TMO double perovskites systems has already been elucidated in the introduction. Here, we first evaluate possible combinations of the elements that could form Bi-based TMO-double perovskites. For predicting the formability of perovskites ABX_3, a parameter 'τ' called the new tolerance factor was proposed by Bartel et al. [48]. If $\tau < 4.18$, ABX_3 is classified as a perovskite else, the compound is a non-perovskite. τ is defined as:

$$\tau = (r_x/r_B) - (n_a(n_a - ((r_A/r_B)/ln(r_A/r_B)))) \quad (10)$$

where r_i is the ionic radius of ion 'i'' and n_a is the ionic charge of A. τ was developed using data analytics [48] and was shown to label 94 % of perovskites and 89 % of non-perovskites correctly. When extended to double perovskites, τ correctly classified almost 91 % of the compounds [42].

For finding prospective oxide double perovskites for photovoltaic

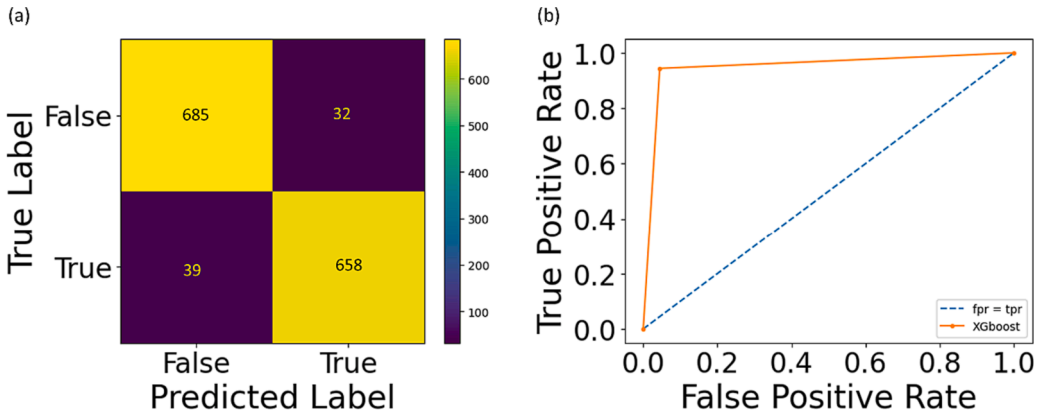

Fig. 6. (a) Confusion matrix for classification of bandgaps into direct and indirect. (b) ROC-AUC plot of XGBOOST.

applications, two compositions are chosen: $Bi_2B'B''O_6$ and A_2BBiO_6. Combinations of cations B', B'', A and B spanning the periodic table are chosen to preserve electrical neutrality in the stoichiometric composition. This is followed by the corresponding τ value calculation using eq. (12). Since the work aims to find prospective double perovskite photovoltaic materials that could be commercially manufactured, expensive and toxic elements are excluded. Also, halogens and elements that make the perovskite structure unstable are excluded. These excluded elements are marked in the periodic table, as shown in Fig. 7.

The Shanon ionic radii are used to calculate the τ values, giving ionic radius as a function of oxidation state and coordination number [49]. The Shanon radius of ions with coordination number 12 is unavailable for many elements. Since the 'A' atoms of $A_2B'B''X_6$ have coordination number 12, it is necessary to get the corresponding radius. This radius was calculated by getting the ionic radius of the element for various coordination numbers and fitting a straight line through these data points. The straight-line equation gives the ionic radius as a function of the coordination number of an element. Using this equation, the desired ionic radius was obtained. Once τ values were calculated for all the required combinations of elements, the probability P(τ) that the compound forms a perovskite structure was estimated by running a logistic regression algorithm on the obtained τ values. The probability P(τ) plots for $Bi_2B'B''O_6$ and A_2BBiO_6 for various combinations of B', B'', A and B elements are shown in Fig. 8. Each box in Fig. 8(a) represents a combination of B' and B'' in the compound $Bi_2B'B''O_6$. The colour of each box represents the probability P(τ) to form a perovskite with the colour

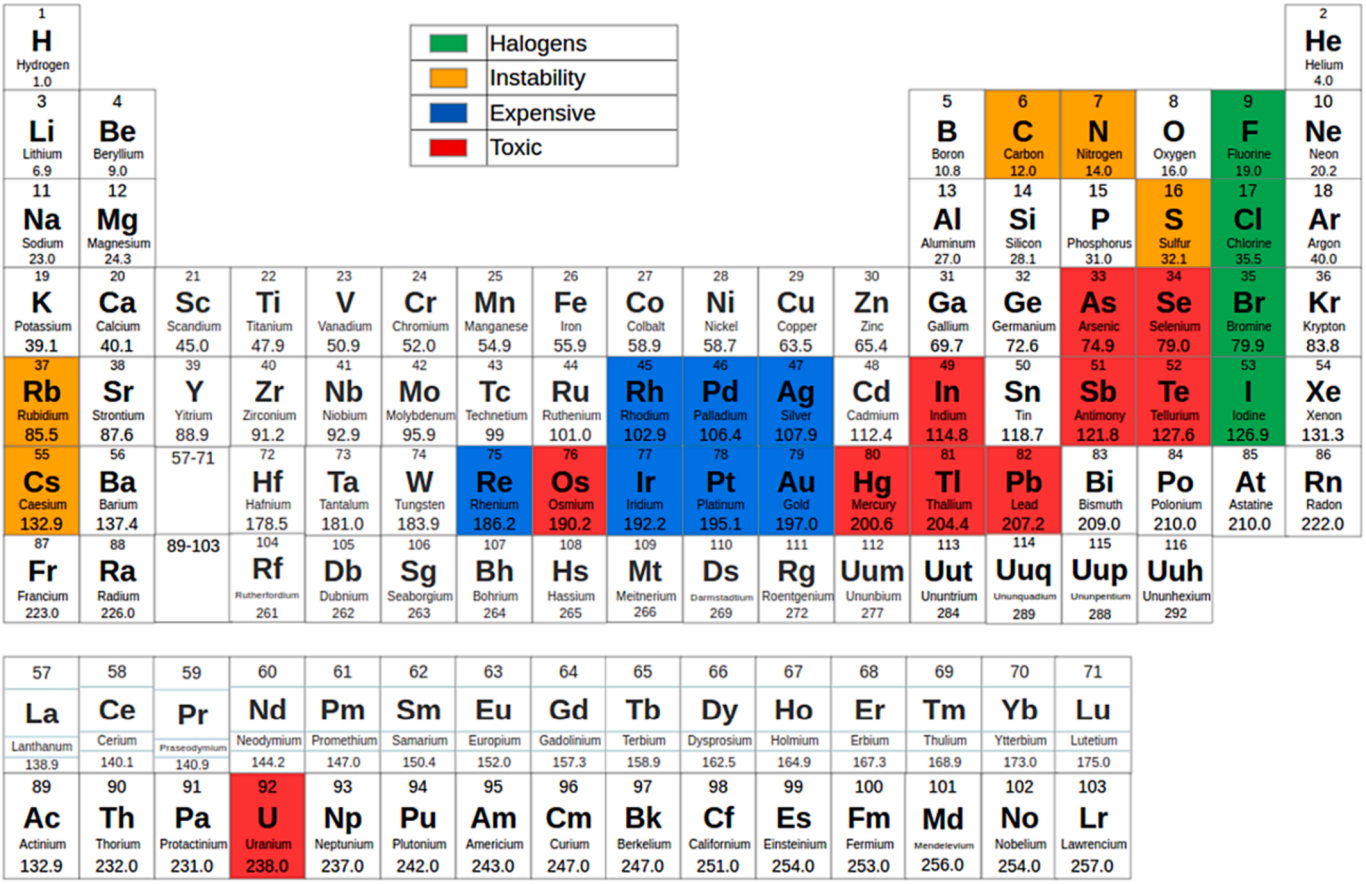

Fig. 7. Combinations of B', B'', A, and B elements are chosen to form $Bi_2B'B''O_6$ and A_2BBiO_6 double perovskite. The current study did not consider expensive, unstable, toxic elements and halogens.

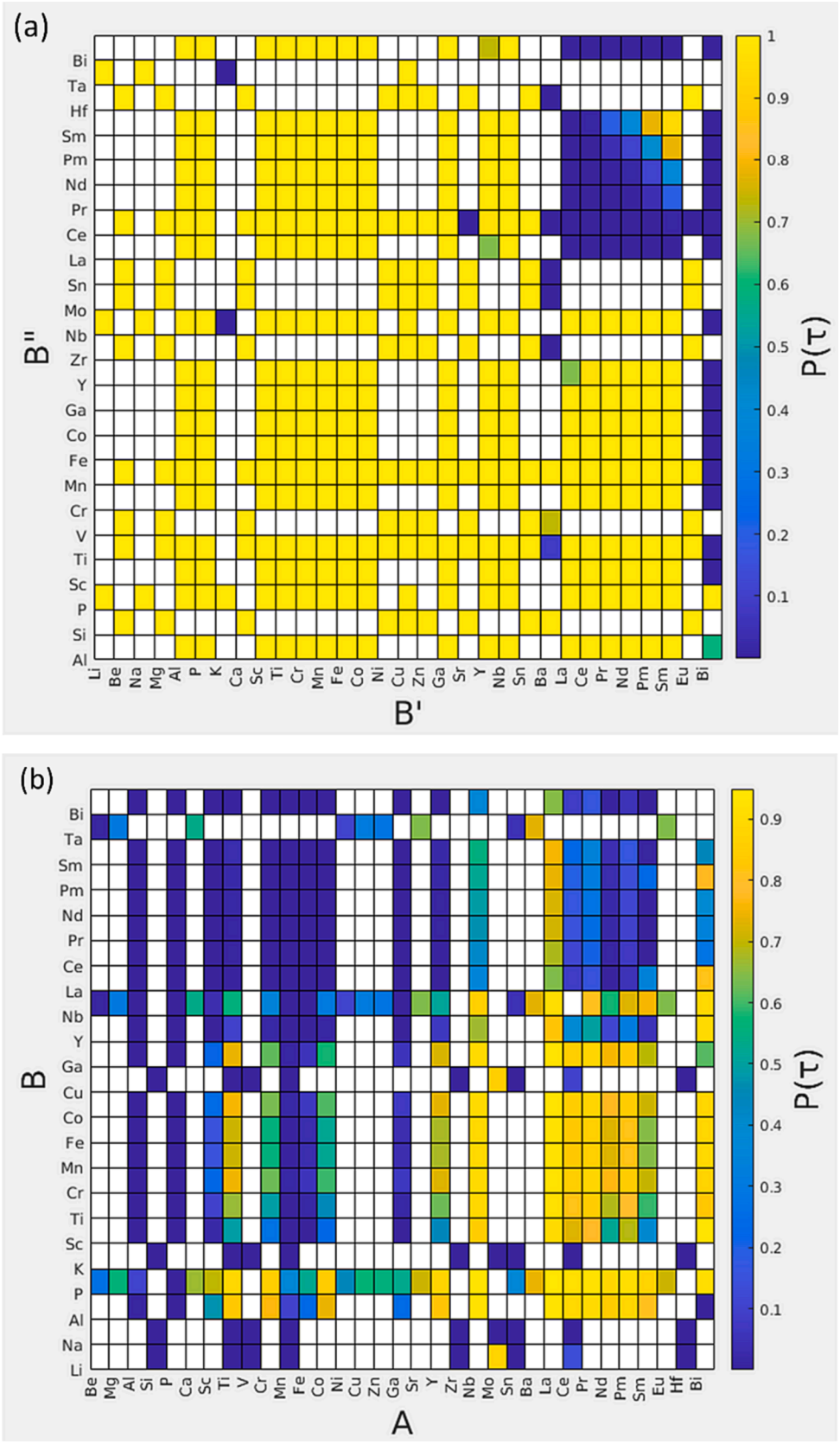

Fig. 8. Probability P(τ) plots for (a) $Bi_2B'B''O_6$ and (b) A_2BBiO_6 for various combinations of B', B'', A and B elements.

bar to the right giving the corresponding probability of the colour. The combination of elements corresponding to white coloured boxes (in both figures (a) and (b)) result in compounds with non-zero charge and hence are invalid combinations. Similarly, Fig. 8(b) gives the probability P(τ) that the combination of A and B elements forms A_2BBiO_6 double perovskites. All the yellow boxes present feasible combinations.

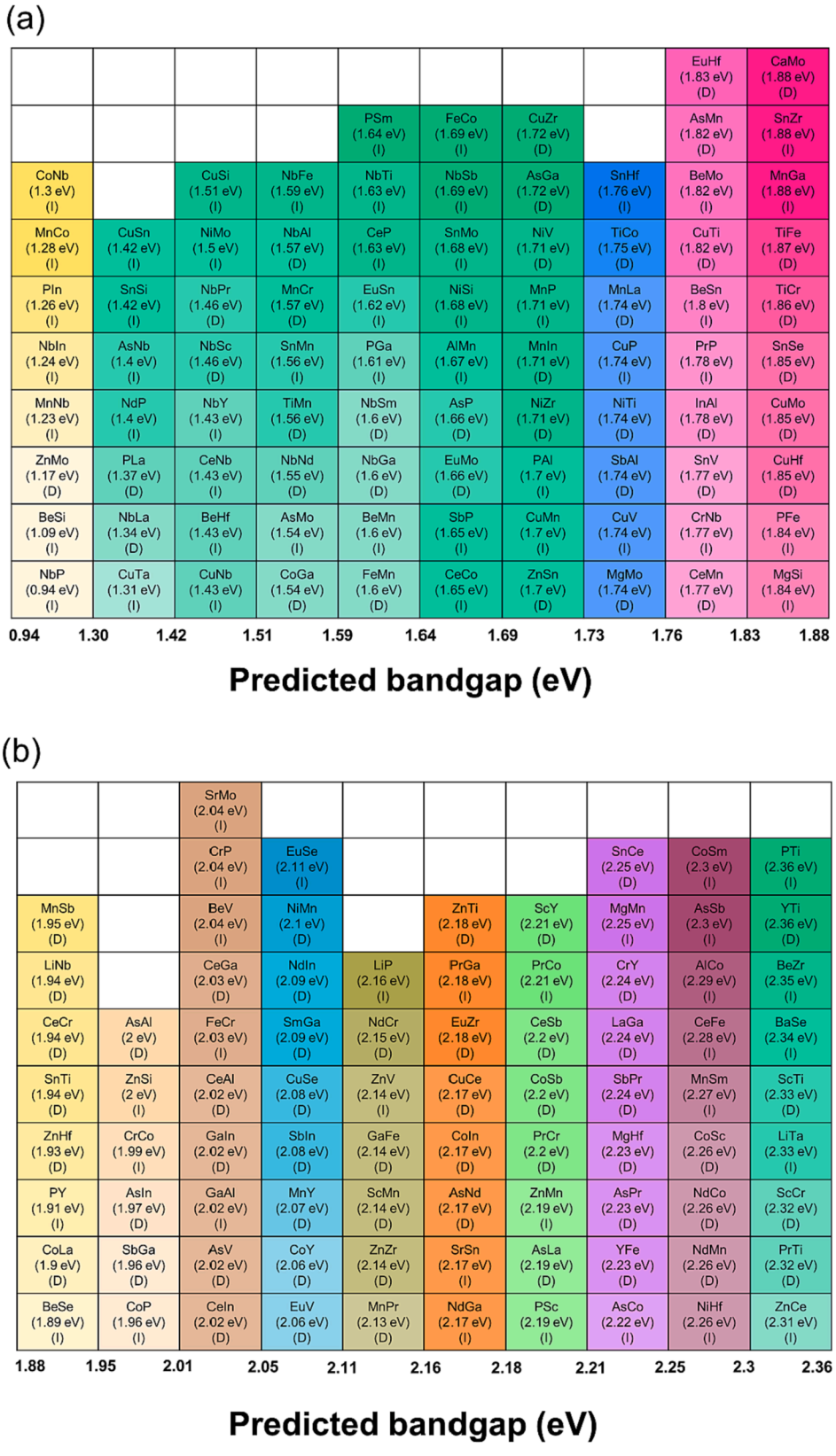

Fig. 9. Predicted bandgap values and type (I- Indirect and D-Direct) for $Bi_2B'B''O_6$ materials in bandgap range (a) 0.94–1.88 eV, (b) 1.88–2.36 eV and (c) 2.36–3.27 eV. (d) For A_2BBiO_6 materials in the bandgap range of 1.4–3.12 eV. Only the elements occupying B', B'', A, and B sites are shown here.

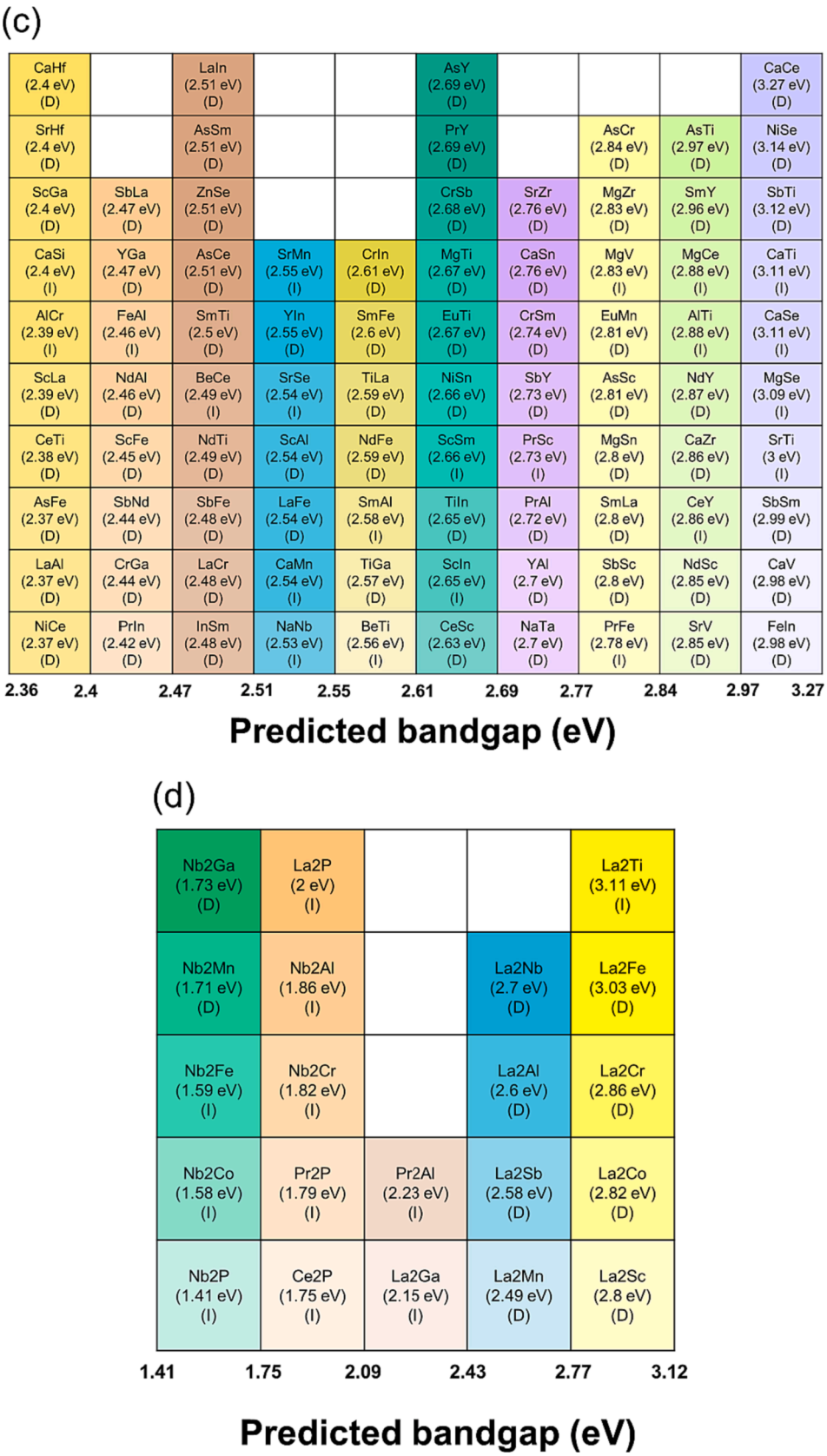

Fig. 9. (*continued*).

To further obtain the relative structural stability and to get structural information on the prospective double perovskites, we calculated the global instability index (GII) [50] of the competing phases using SPuDs software [51]. Typically, phases with GII < 0.2 are expected to become stable. We conjectured that the structure with the lowest GII had the most stable structure. It was found that structures with rhombohedral $R\bar{3}$

symmetry corresponded to the lowest GII and was therefore considered for further study. SPuDs was also used to generate crystallographic information files (CIF) with $R\bar{3}$ symmetry for the Bi-based double perovskites. Then, using the above-developed ML models, the bandgap and its type were predicted for the list of the $Bi_2B'B''O_6$ and A_2BBiO_6 compounds and are shown by color maps in Fig. 9. The bandgap values ranged from 0.9 to 3.3 eV and 1.4–3.2 eV for the $Bi_2B'B''O_6$ and A_2BBiO_6 compounds, finding potential wide-scale applications. Since Bi and O are common for all, only the elements occupying B', B'', A and B sites have been shown here. In a particular bandgap range, the shade of the colour becomes darker from the bottom to the top, signifying the bandgap value increase. The white boxes denote no accompanying materials. Most of these double perovskites oxides with $R\bar{3}$ symmetry are unexplored and, to the best of our knowledge, have never been experimentally synthesised. Since the spacegroup number has a comparatively low positive correlation (0.17) with bandgap values, the effect of the crystal symmetry change on the bandgap will not be dramatic. Some model-predicted bandgap values of Bi-based double perovskite oxides have also shown a good match with known bandgap values of similarly structured materials as reported in the literature. For example, the predicted bandgap of Bi_2FeCrO_6 (BFCO) is 2.03 eV and is an excellent match to its rhombohedral *R*3 symmetry structure having an experimental bandgap of 1.9/2.6 eV for the ordered/disordered film and 2.07 eV as reported in the literature [52] and [53]. It also matches well with the DFT-predicted bandgap values (2.3/2.50–2.55 eV for ordered/disordered configuration) performed using the HSE06 functional by Quattropani et al. [52]. Similarly, Bi_2FeMnO_6 is predicted to have a bandgap of ~ 1.6 eV with a direct bandgap type that closely matches its *R*3*c* symmetry structure having a direct bandgap of ~ 1.23 eV and measured using UV–Vis absorption spectroscopy [54].

Certain reports, however, demonstrate significant discrepancies in the DFT-predicted bandgap values. Wen et al., for example, predicted the bandgap of BFCO using LDA, GGA, and mBJ functionals and discovered it to be 0.26, 0.42, and 1.48 eV, respectively [55]. Tablero also investigated BFCO theoretically using various U values for the GGA + U functional. The BFCO energy gaps for up and down spin with U = 6, 8, and 10 eV were estimated to be 1.10/2.23 eV, 1.22/2.39 eV, and 1.41/2.58 eV, respectively [56].

Nonetheless, 67 materials of the $Bi_2B'B''O_6$ family and 7 materials in the A_2BBiO_6 family possess bandgap in the optimum range (1.2–1.8 eV) favourable for PV applications. We believe the predicted list of materials with favourable bandgaps will trigger further theoretical and experimental studies of Bi-based transition metal oxide double perovskites for solar cell applications.

5. Limitations and scope of improvement in the current work

The current work uses DFT-calculated bandgaps computed using common approximation schemes like the LDA and GGA. The LDA/GGA predictions are severely underestimated with respect to the experiments [57]. For example, an internal test done at Materials Project [22] on the experimental versus computed bandgaps for 237 compounds found that the computed gaps are underestimated by about 40 %, with the MAE being 0.6 eV. Several well-known insulators were also predicted as metals. The errors can be attributed to exchange–correlation functional approximations [58], self-interaction errors (SIE) [59,60], and a derivative discontinuity term resulting from the true density functionals being discontinuous with the total number of electrons in the system [61]. The latter's contribution is often greater and adds more to the inaccuracy. GW approximations could partially solve the problem but at a hefty computational expense. Other methods for improving bandgap predictions at low to moderate computing costs include delta-sol [62], hybrid functionals [63], and empirical fits [64].

Furthermore, the current work does not consider the experimental conditions, like the synthesis technique used, synthesis temperature, pressure, and time, that may affect the grain size of the polycrystalline sample, influencing its bandgap. Other factors, such as defect concentration/doping level, dielectric constant, and lattice disorder, are also not considered here, which could potentially impact the experimentally measured bandgap of the material. Another issue might be that we build the models with DFT-relaxed structures from the MP database. DFT-predicted structures might demonstrate deviation of lattice constants by up to several percent with respect to the experimentally reported values. This can lead to different values of the computed bandgap compared to its experimentally observed one. Yet another issue of the present model could be its inability to correctly predict the bandgap and its character for disordered systems and systems whose crystal structure is unknown since the models were trained using certain features having crystal structure information of the ordered structures. Also, due to insufficient training data, substantial errors may occur. With more data being used for training, the learning of the model improves, enhancing prediction performance.

Another scope of improvement in the current work is to predict the transport properties of double perovskite oxides. A few studies used ML techniques to probe relationships between local atomic structure and global electronic transport coefficients. For example, Pimachev et al. [65] established a DFT-based electronic transport-informatics (ETI) framework for estimating electronic transport coefficients of semiconductor heterostructures based on DFT-calculated structures and electronic band characteristics. The framework hypothesises that the material's crystal structure determines its physical properties (mechanical or electronic). Thus, the proposed ML model uses local structures determined using Voronoi tessellations as input to predict the global electronic energy bands. Since the carrier's effective mass is strongly tied to its mobility, a reduced effective mass increases band edge dispersion and mobility. High carrier (electron and hole) mobility is also predicted for band configurations with higher diffusion around the Fermi surface [66]. A similar approach could also be used to determine the carrier mobilities of double perovskites with sufficient bandstructure data.

6. Conclusion

DFT-based first-principles modelling is common but expensive and inaccurate in predicting materials bandgap. Machine learning can model fundamental electronic structural features like the bandgap. The current machine learning approach leverages structural and elemental descriptors to efficiently forecast the nature and magnitude of bismuth-based transition metal oxide double perovskites using over 4000 DFT-computed bandgaps. The model found that electronegativity, HOMO-LUMO energies, ionisation energy, the maximum number of d electrons, atomic radii, and bulk modulus are some of the essential atomic descriptors that affect the bandgap, providing physically interpretable descriptions of the electronic structure. LGBMRegressor had R^2 of ~ 0.89 in bandgap prediction, and XGBClassifier showed 0.95 accuracy in determining the bandgap type of double perovskite structures. Finally, we used the developed models to list environmentally benign and affordable novel Bi-based double perovskite oxides with their projected bandgap values and types. This dataset is expected to be useful in follow-up studies in materials research as it can guide the synthesis of compounds with specific bandgap requirements, especially for novel photovoltaic materials.

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

SS thanks the Ministry of Education, Government of India (GOI), for the research fellowship. This work was supported by SERB, Govt. of India through project no. CRG/2019/003828.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.solener.2023.112209.

References

[1] A.K. Jena, A. Kulkarni, T. Miyasaka, Halide Perovskite Photovoltaics: Background, Status, and Future Prospects, Chem. Rev. 119 (2019) 3036–3103, https://doi.org/10.1021/acs.chemrev.8b00539.
[2] M. Aslam Khan, H.A. Alburaih, N.A. Noor, A. Dahshan, Comprehensive investigation of Opto-electronic and transport properties of Cs2ScAgX6 (X = Cl, Br, I) for solar cells and thermoelectric applications, Sol. Energy 225 (2021) 122–128, https://doi.org/10.1016/J.SOLENER.2021.07.026.
[3] P. Aymen Nawaz, G.M. Mustafa, S. Sagar Iqbal, N.A. Noor, T. Shahzad Ahmad, A. Mahmood, R. Neffati, Theoretical investigations of optoelectronic and transport properties of Rb2YInX6 (X = Cl, Br, I) double perovskite materials for solar cell applications, Sol. Energy 231 (2022) 586–592, https://doi.org/10.1016/J.SOLENER.2021.11.076.
[4] O. Xiu-Juan Li, Z.-B. Zhang, C.-T. Zhang, al -, M.A. Muñoz-Gutiérrez, B. Pichardo, A. Peimbert -, A.J. Verbiscer, P. Helfenstein, H.A. Alburaih, M. Bououdina, R. Sharma, A. Laref, R. Neffati, N.A. Noor, Opto-electronic and thermoelectric properties of free-lead inorganic double perovskites Rb/Cs2ScAuI6 for energy devices, Phys. Scr. 98 (2023) 085925. https://doi.org/10.1088/1402-4896/ACE4FF.
[5] M. Waqas Iqbal, M. Manzoor, N.A. Noor, I. Rehman, N. Mushahid, S. Aftab, Y. M. Alanazi, H. Ullah, A. Muhammad Afzal, Tuning of the electronic bandgap of lead-free double perovskites K2AgBiX6 (X = Cl, Br) for solar cells applications and their thermoelectric characteristics, Sol. Energy 239 (2022) 234–241, https://doi.org/10.1016/J.SOLENER.2022.05.018.
[6] D.S. Knoche, M. Steimecke, Y. Yun, L. Mühlenbein, A. Bhatnagar, Anomalous circular bulk photovoltaic effect in BiFeO3 thin films with stripe-domain pattern, Nat. Commun. 121(12) (2021) 1–8. https://doi.org/10.1038/s41467-020-20446-z.
[7] R. Nechache, C. Harnagea, S. Li, L. Cardenas, W. Huang, J. Chakrabartty, F. Rosei, Bandgap tuning of multiferroic oxide solar cells, Nat. Photonics 91(9) (2014) 61–67. https://doi.org/10.1038/nphoton.2014.255.
[8] A.M. Glazer, Perovskites modern and ancient ., Roger H. Mitchell. Thunder Bay, Ontario: Almaz Press, 2002., Acta Crystallogr. Sect. B Struct. Sci. 58 (2002) 1075–1075. https://doi.org/10.1107/S0108768102020220.
[9] J.Y. Kim, J.W. Lee, H.S. Jung, H. Shin, N.G. Park, High-efficiency perovskite solar cells, Chem. Rev. 120 (2020) 7867–7918, https://doi.org/10.1021/acs.chemrev.0c00107.
[10] J. Lee, A. Seko, K. Shitara, I. Tanaka, Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B 93 (2015), https://doi.org/10.1103/PhysRevB.93.115104.
[11] A.C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.R. Lee, A.K. Singh, Machine-learning-assisted accurate band gap predictions of functionalized mxene, Chem. Mater. 30 (2018) 4031–4038, https://doi.org/10.1021/acs.chemmater.8b00686.
[12] M. Guo, X. Xu, H. Xie, Predicting the band gap of binary compounds from machine-learning regression methods, ChemRxiv (2021). https://doi.org/10.26434/CHEMRXIV-2021-JHG7B.
[13] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites, Sci. Rep. 6 (2016), https://doi.org/10.1038/srep19375.
[14] J. Zhang, Y. Li, X. Zhou, Machine-Learning Prediction of the Computed Band Gaps of Double Perovskite Materials, (2022). https://doi.org/10.26434/CHEMRXIV-2022-BLKMP.
[15] Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, Y. Jia, Band gap and band alignment prediction of nitride-based semiconductors using machine learning, J. Mater. Chem. C 7 (2019) 3238–3245, https://doi.org/10.1039/C8TC05554H.
[16] A. Talapatra, B.P. Uberuaga, C.R. Stanek, G. Pilania, Band gap predictions of double perovskite oxides using machine learning, Commun. Mater. 4 (2023) 1–14, https://doi.org/10.1038/s43246-023-00373-4.
[17] V. Venkatraman, The utility of composition-based machine learning models for band gap prediction, Comput. Mater. Sci. 197 (2021), 110637, https://doi.org/10.1016/j.commatsci.2021.110637.
[18] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, Predicting the Band Gaps of Inorganic Solids by Machine Learning, J. Phys. Chem. Lett. 9 (2018) 1668–1673, https://doi.org/10.1021/acs.jpclett.8b00124.
[19] V. Gladkikh, D.Y. Kim, A. Hajibabaei, A. Jana, C.W. Myung, K.S. Kim, Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties, J. Phys. Chem. C 124 (2020) 8905–8918, https://doi.org/10.1021/acs.jpcc.9b11768.
[20] L. Wu, Y. Xiao, M. Ghosh, Q. Zhou, Q. Hao, Machine Learning Prediction for Bandgaps of Inorganic Materials, ES Mater. Manuf. 9 (2020) 34–39, https://doi.org/10.30919/ESMM5F756.
[21] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, Npj Comput. Mater. 2016 21. 2 (2016) 1–7. https://doi.org/10.1038/npjcompumats.2016.28.
[22] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (2013), 011002, https://doi.org/10.1063/1.4812323.
[23] L. Weston, C. Stampfl, Machine learning the band gap properties of kesterite I2-II-IV- V4 quaternary compounds for photovoltaics applications, Phys. Rev. Mater. 2 (2018), 085407, https://doi.org/10.1103/PhysRevMaterials.2.085407.
[24] G. Kresse, J. Furthmüller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169, https://doi.org/10.1103/PhysRevB.54.11169.
[25] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50, https://doi.org/10.1016/0927-0256(96)00008-0.
[26] G. Kresse, J. Hafner, *Ab initio* molecular dynamics for liquid metals, Phys. Rev. B 47 (1993) 558, https://doi.org/10.1103/PhysRevB.47.558.
[27] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77 (1996) 3865, https://doi.org/10.1103/PhysRevLett.77.3865.
[28] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758, https://doi.org/10.1103/PhysRevB.59.1758.
[29] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953, https://doi.org/10.1103/PhysRevB.50.17953.
[30] J.W. Furness, A.D. Kaplan, J. Ning, J.P. Perdew, J. Sun, Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation, J. Phys. Chem. Lett. 11 (2020) 8208–8215, https://doi.org/10.1021/acs.jpclett.0c02405.
[31] H.J. Monkhorst, J.D. Pack, Special points for Brillouin-zone integrations, Phys. Rev. B 13 (1976) 5188, https://doi.org/10.1103/PhysRevB.13.5188.
[32] G. Bergerhoff, R. Hundt, R. Sievers, I.D. Brown, The Inorganic Crystal Structure Data Base, J. Chem. Inf. Comput. Sci. 23 (1983) 66–69, https://doi.org/10.1021/ci00038a003.
[33] R. Hundt, J.C. Schon, M. Jansen, CMPZ– an algorithm for the efficient comparison of periodic structures, J. Appl. Crystallogr. 39 (2006) 6–16, https://doi.org/10.1107/S0021889805032450.
[34] S. Kanno, Y. Imamura, M. Hada, Alternative materials for perovskite solar cells from materials informatics, Phys. Rev. Mater. 3 (2019), 075403, https://doi.org/10.1103/PhysRevMaterials.3.075403.
[35] Y. He, E.D. Cubuk, M.D. Allendorf, E.J. Reed, Metallic metal-organic frameworks predicted by the combination of machine learning methods and ab initio calculations, J. Phys. Chem. Lett. 9 (2018) 4562–4569, https://doi.org/10.1021/acs.jpclett.8b01707.
[36] T. Wang, K. Zhang, J. Thé, H. Yu, Accurate prediction of band gap of materials using stacking machine learning model, Comput. Mater. Sci. 201 (2022), 110899, https://doi.org/10.1016/J.COMMATSCI.2021.110899.
[37] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/JAIR.953.
[38] F. Pedregosa FABIANPEDREGOSA, V. Michel, O. Grisel OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, andÉdouard Duchesnay, Fré. Duchesnay EDOUARDDUCHESNAY, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html (accessed May 27, 2023).
[39] C.H. Chen, K. Tanaka, M. Kotera, K. Funatsu, Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications, J. Cheminform. 12 (2020) 1–16, https://doi.org/10.1186/s13321-020-0417-9.
[40] A.-Y.-T. Wang, R.J. Murdock, S.K. Kauwe, A.O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T.D. Sparks, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, Chem. Mater. 32 (2020) 4954–4965, https://doi.org/10.1021/acs.chemmater.0c01907.
[41] H. M, S. M.N, A Review on Evaluation Metrics for Data Classification Evaluations, Int. J. Data Min. Knowl. Manag. Process. 5 (2015) 01–11. https://doi.org/10.5121/ijdkp.2015.5201.
[42] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S.P. Ong, A Critical Review of Machine Learning of Energy Materials, Adv. Energy Mater. 10 (2020) 1903242, https://doi.org/10.1002/AENM.201903242.
[43] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nat. Commun. 2017 81. 8 (2017) 1–12. https://doi.org/10.1038/ncomms15679.
[44] T. Xie, J.C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, Phys. Rev. Lett. 120 (2018), 145301, https://doi.org/10.1103/PhysRevLett.120.145301.
[45] L. Pauling, The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms, J. Am. Chem. Soc. 54 (1932) 3570–3582, https://doi.org/10.1021/ja01348a011.
[46] D.C. Ghosh, T. Chakraborty, Gordy's electrostatic scale of electronegativity revisited, J. Mol. Struct. Theochem. 906 (2009) 87–93, https://doi.org/10.1016/j.theochem.2009.04.007.

[47] K. Li, C. Kang, D. Xue, Electronegativity calculation of bulk modulus and band gap of ternary ZnO-based alloys, Mater. Res. Bull. 47 (2012) 2902–2905, https://doi.org/10.1016/J.MATERRESBULL.2012.04.115.
[48] 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 (2019), https://doi.org/10.1126/sciadv.aav0693.
[49] R.D. Shannon, IUCr, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Urn:Issn:0567–7394 32 (1976) 751–767, https://doi.org/10.1107/S0567739476001551.
[50] A. Salinas-Sanchez, J.L. Garcia-Muñoz, J. Rodriguez-Carvajal, R. Saez-Puche, J. L. Martinez, Structural characterization of R2BaCuO5 (R = Y, Lu, Yb, Tm, Er, Ho, Dy, Gd, Eu and Sm) oxides by X-ray and neutron diffraction, J. Solid State Chem. 100 (1992) 201–211, https://doi.org/10.1016/0022-4596(92)90094-C.
[51] M.W. Lufaso, P.M. Woodward, Prediction of the crystal structures of perovskites using the software program SPuDS, Acta Crystallogr. B 57 (2001) 725–738, https://doi.org/10.1107/S0108768101015282.
[52] A. Quattropani, D. Stoeffler, T. Fix, G. Schmerber, M. Lenertz, G. Versini, J. L. Rehspringer, A. Slaoui, A. Dinia, S. Colis, Band-Gap Tuning in Ferroelectric Bi2FeCrO6 Double Perovskite Thin Films, J. Phys. Chem. C 122 (2018) 1070–1077, https://doi.org/10.1021/acs.jpcc.7b10622.
[53] D.S. Walch, Y. Yun, N. Ramakrishnegowda, L. Mühlenbein, A. Lotnyk, C. Himcinschi, A. Bhatnagar, Resistive Switching in Ferroelectric Bi2FeCrO6 Thin Films and Impact on the Photovoltaic Effect, Adv. Electron. Mater. 8 (2022) 2200276, https://doi.org/10.1002/AELM.202200276.
[54] Z. Pei, K. Leng, W. Xia, Y. Lu, H. Wu, X. Zhu, Structural characterization, dielectric, magnetic and optical properties of double perovskite Bi2FeMnO6 ceramics, J. Magn. Magn. Mater. 508 (2020), 166891, https://doi.org/10.1016/J.JMMM.2020.166891.
[55] Z.-W. Song, 宋哲文, B.-G. Liu, 刘邦贵, Electronic structure and magnetic and optical properties of double perovskite Bi2FeCrO6 from first-principles investigation, Chinese Phys. B. 22 (2013) 047506. https://doi.org/10.1088/1674-1056/22/4/047506.
[56] C. Tablero, Photovoltaic application of the multiferroic Bi2FeCrO6 double perovskite, Sol. Energy 137 (2016) 173–178, https://doi.org/10.1016/J.SOLENER.2016.08.004.
[57] P. Scharoch, M. Winiarski, An efficient method of DFT/LDA band-gap correction, Comput. Phys. Commun. 184 (2013) 2680–2683, https://doi.org/10.1016/J.CPC.2013.07.008.
[58] R.M. Martin, Electronic Structure: Basic Theory and Practical Methods (2004), https://doi.org/10.1017/CBO9780511805769.
[59] V.I. Anisimov, F. Aryasetiawan, A.I. Lichtenstein, First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+ U method, J. Phys. Condens. Matter. 9 (1997) 767, https://doi.org/10.1088/0953-8984/9/4/002.
[60] L. Wang, T. Maxisch, G. Ceder, Oxidation energies of transition metal oxides within the GGA+U framework, Phys. Rev. B - Condens. Matter Mater. Phys. 73 (2006), 195107, https://doi.org/10.1103/PhysRevB.73.195107.
[61] J.P. Perdew, M. Levy, Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities, Phys. Rev. Lett. 51 (1983) 1884, https://doi.org/10.1103/PhysRevLett.51.1884.
[62] M.K.Y. Chan, G. Ceder, Efficient band gap prediction for solids, Phys. Rev. Lett. 105 (2010), 196403, https://doi.org/10.1103/PhysRevLett.105.196403.
[63] J. Heyd, J.E. Peralta, G.E. Scuseria, R.L. Martin, Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional, J. Chem. Phys. 123 (2005), https://doi.org/10.1063/1.2085170.
[64] W. Setyawan, R.M. Gaume, S. Lam, R.S. Feigelson, S. Curtarolo, High-throughput combinatorial database of electronic band structures for inorganic scintillator materials, ACS Comb. Sci. 13 (2011) 382–390, https://doi.org/10.1021/co200012w.
[65] A.K. Pimachev, S. Neogi, First-principles prediction of electronic transport in fabricated semiconductor heterostructures via physics-aware machine learning, Npj Comput. Mater. (2021), https://doi.org/10.1038/s41524-021-00562-0.
[66] A. Chen, S. Ye, Z. Wang, Y. Han, J. Cai, J. Li, Machine-learning-assisted rational design of 2D doped tellurene for fin field-effect transistor devices, Patterns 4 (2023), 100722, https://doi.org/10.1016/j.patter.2023.100722.