Discover Materials

RESEARCH Open Access

Exploring nontoxic perovskite materials for perovskite solar cells using machine learning

W. G. A. Pabasara[1,2], H. A. H. M. Wijerathne[1], M. G. M. M. Karunarathne[1], D. M. C. Sandaru[1], Pradeep K. W. Abeygunawardhana[3] and Galhenage A. Sewvandi[1*]

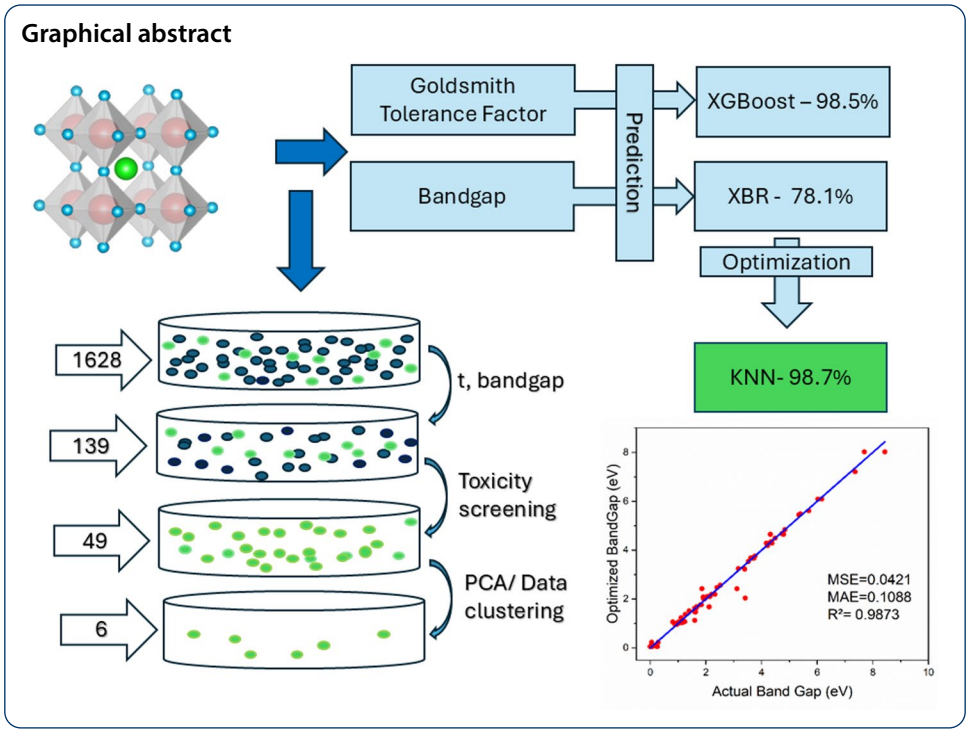
*Correspondence:
Galhenage A. Sewvandi
galhenagea@uom.lk
[1]Department of Materials Science and Engineering, Faculty of Engineering, University of Moratuwa, Bandaranayake Mawatha, Katubedda, Moratuwa, Sri Lanka
[2]Department of Engineering Technology, Faculty of Technology, University of Ruhuna, Karagoda, Uyangoda, Kamburupitiya, Sri Lanka
[3]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Abstract
Perovskite solar cells are promising renewable energy technology that faces significant challenges due to the Pb induced toxicity. The current study addresses this issue by leveraging machine learning techniques to explore Pb-free perovskite materials that ensure environmental sustainability and human safety. A highly accurate machine learning model was developed to predict Goldschmidt factor and the band gap, aiming to discover lead-free perovskites. Extreme Gradient Boost (XGBoost), Random Forest (RF), Gradient Boost Regression (GBR), and Ada Boost Regression (ABR) models were employed for this purpose. The findings exhibit that XGBoost delivers the most precise and reliable results for Goldsmith tolerance factor prediction with an accuracy of 98.5%. Furthermore, GBR model, combined with K-nearest neighbors (KNN) model delivers an impressive accuracy of 98.7% for the band gap predictions. 49 Pb-free perovskite materials were screened out considering the toxicity and the abundance. Utilizing Principal Component Analysis (PCA) and K-means clustering, six optimal materials ($KBiBr_3$, $KZnBr_3$, $RbBiBr_3$, $RbZnBr_3$, $MAGeI_3$, and $FAGeI_3$null) were identified as the potential environment-friendly materials for photovoltaic applications. These results show the crucial role of machine learning and statistical analysis in discovering nontoxic and environmental-friendly perovskite materials, advancing the development of sustainable energy solutions.

Keywords Perovskite solar cells, Machine learning, Lead alternatives, Bandgap prediction

1 Introduction

Perovskite solar cells (PSCs) have become a promising third-generation photovoltaic technology exhibiting superior power conversion efficiencies with affordable manufacturing practices [1–3]. Within a decade, the power conversion efficiency (PCE) of PSC has improved by 20% with the rapid technological advancement of the photovoltaic field, showing signs of rapid early growth of PCS technology. By 2025, it has achieved 27% PCE at the laboratory scale, demonstrating substantial progression [4]. Perovskite solar cell consists of several layers of materials, including a transparent conductive electrode,



an electron transport layer, a perovskite structured light absorbing layer, a hole transport layer, and a back electrode. ABX_3 is the standard formula of perovskite structure. Here, A is denoted as a monovalent cation such as $CH_3NH_3^+$, Cs^+, Rb^+, or their mixtures. B is a divalent cation such as Pb^{2+}, Sn^{2+}, Ge^{2+}, or a blend of these metals. X indicates monovalent anions, typically halides, such as I^-, Br^-, Cl^-, or a mixture of these halides.

Studies have proved that Pb-based perovskites more favorably achieve higher efficiency levels of PSCs [5, 6]. Pb-based perovskites show extraordinary characteristics due to their high absorption all along the visible light range, extended carrier diffusion lengths, low-temperature processing, and tunable bandgap [7, 8]. The high absorption coefficient allows great light harvesting even with thin absorbing layers. They also exhibit efficient charge transport and collections due to long charge carrier diffusion lengths. Furthermore, Pb-based perovskites show high defect tolerances, so the effect of defect-induced trap states on the device performance is minimal [9]. However, the inclusion of Pb is a serious issue due to its toxicity to health and the environment [10–12]. When solar panels are exposed to the open environment, PbI_2 degrades as a substance due to moisture. That may result in serious health issues, including cardiovascular diseases and neurological and reproductive system damage [13]. In addition, lead contamination in water resources and soil has long-term impacts on humans, animals, and plants [3, 14]. These detrimental effects of Pb have opened new avenues for discovering Pb-free alternatives with suitable photovoltaic properties for solar cells.

The perovskite structure can be found in a diverse range of compounds, allowing for a number of material combinations. However, these manifold possibilities mean that a large number of materials should be investigated, making it a time-consuming process. Currently, the Density Functional Theory (DFT) based simulation is utilized to find the appropriate perovskite materials for solar cell applications [15, 16]. However, this approach is associated with high computational costs and requires extensive quantum

chemistry knowledge. Furthermore, it is common to observe differences between experimental measurements and theoretical predictions [17]. Data-driven research has gained widespread attention due to its efficiency and effectiveness [18]. In this approach, precomputed materials databases and statistical techniques effectively screen the most appropriate candidates. Machine learning (ML) is being increasingly utilized to predict crystal structures, develop predictive models for material properties, and develop interatomic potentials [19–22].

Structural formability and band gap of perovskite material plays an important role in deciding the appropriateness of perovskite materials for different optoelectronic applications, particularly solar cells. Structural formability is the ability of a material to fit into a perovskite ABX_3 crystal structure. The structural formability of the perovskite materials is characterized by the Goldsmith tolerance factor (t), which verifies the geometric compatibility between A cation and BX_6 octahedron. The band gap of the materials is prominent since it decides the photon energy range that a perovskite material can absorb with high efficiency, determining the optical and electronic characteristics of the semiconductors [3]. The optimally chosen bandgap can absorb a significant portion of the solar spectrum, hence enhancing the overall efficiency of the solar cell. The bandgap of the material can be tuned by compositional engineering by adjusting the A, B, or X sites in the perovskite structure. Therefore, while searching for new perovskite materials, the application of bandgap enables the rapid identification of suitable materials. Different machine learning models have been employed to predict the bandgap of the perovskite materials and, thereby, identify the promising candidates for solar cell applications.

In a recent study, CatBoost algorithms has been identified as the best performer in bandgap prediction, with the accuracy of 92.3% [23]. Gradient Boosted Regression Trees (GBRT) model demonstrated an accuracy of 87% in the bandgap prediction of lead-free halide double perovskites [24]. Using the RF model, 1252 perovskite materials were identified as top candidates based on the formability screening, where bandgap prediction reached an accuracy of 87% using the XG Boost regression algorithm [25]. One hundred thirty-two stable free hybrid organic-inorganic perovskites (HOIPs) materials were identified by a combined methodology of ML and DFT calculations in one such study [6]. The most accurate model of bandgap prediction was the GBR model, having an accuracy rate of 82.7%.

Despite the significant progress made in discovering perovskite materials using ML, further research is needed to improve the accuracy of the prediction model and integrate additional techniques to identify a few of the most promising materials for experimental investigations. In this study, Extreme Gradient Boost (XGB), Random Forest (RF), Gradient Boost Regression (GBR), and Ada Boost Regression (ABR) models have been used to predict the Goldschmidt factor and the band gap to identify Pb-free perovskites. Furthermore, the K-nearest Neighbors (KNN) model was applied to enhance the accuracy of band gap prediction. This hybrid approach of the band gap prediction adds new insights into ML-driven new perovskite material discovery. In the existing studies, a large number of possible perovskite materials have been screened out; however, no systematic approach has been applied to refine the selection further, except the DFT based calculations. Therefore, Principal Component Analysis (PCA), a statistical technique, was employed in the predicted data, facilitating the identification of the most appropriate nontoxic perovskite materials for photovoltaic applications.

2 Methods

The methodology followed for conducting this study has been demonstrated as a flow chart in Fig. 1.

2.1 Data processing

A well-curated HTSperoDB high-throughput perovskite database, which is well-known for its extensive computational dataset, was used as the primary data source for band gap prediction [26]. The dataset employed for Goldsmith tolerance factor prediction was based on a previously published study [27]. The categorical dataset in HTSperoDB, phase details, was converted into a binary feature using one-hot encoding method by Python programming within a Google Colab environment. Missing values were identified as null entries in the dataset and all rows containing missing values were removed. The remaining dataset was filtered to retain only structurally valid and thermodynamically stable perovskite materials by excluding non-perovskite phases and thermodynamically unstable materials based on their stability score. This data pre-processing approach filters out reliable datasets, ensuring the accuracy of the analysis.

2.2 Feature selection

This machine learning approach selected three critical target properties, including formability, structural stability, and band gap of perovskite materials, to ensure comprehensive and precise modelling. In formability and structural stability, ionic radii were selected as the key features that determine the effective formation of the perovskite materials. The model utilized the Goldschmidt tolerance factor indicates in Eq. (1).

$$t = \frac{r_A + r_X}{\sqrt{2}(r_A + r_X)} \tag{1}$$

Where r_A, r_B, and r_X, are the ionic radius of the A, B site cations, and X site anions, respectively.

The size of the A, B, and X ions determines the Goldschmidt tolerance so that ionic radii affect the structural stability of the perovskite structure. To retain the 3D perovskite structure, the value of t should lie between 0.81 and 1.00 range [28]. Deviations from this range can lead to distortions or instability in the crystal structure. Therefore, calculating the Goldschmidt tolerance factor using ionic radii makes it possible to predict the structural stability and likelihood of forming a stable perovskite crystal structure.

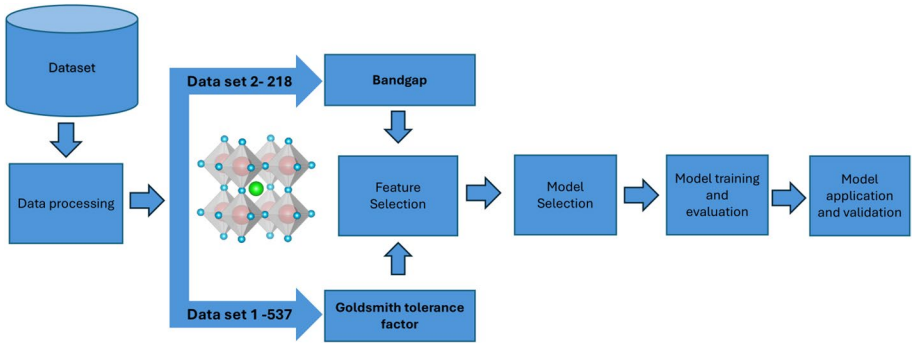


Fig. 1 Flow chart of ML framework

In the band gap model, ionic radii, electronegativity, and ionization energy were selected as primary features as they have a theoretical correlation with the electronic structure of perovskite materials. These properties primarily influence crystal geometry, bond character, and energy level distribution, which affect the band gap of the materials. Consequently, the light absorption of the perovskite material and the efficiency of the solar cell are also affected [29].

Lattice dimensions and bond angles within the ABX_3 perovskite structure are governed by the ionic radius. The size disparities of A, B, and X site ions change the volume of the unit cell and cause octahedral tilting of the BX_6 framework. These structural changes affect orbital overlaps between metal and halide ions, altering band dispersion and bandgap width. Smaller or mismatched ionic radii can distort the crystal structure, decreasing orbital coupling and broadening the bandgap, whereas perfectly matched ionic sizes tend to create favorable electronic properties [30].

In the case of electronegativity, it affects the distribution of electron density within the crystal lattice, which is crucial for forming the band structure and determining the band gap. Larger electronegativity differences between ions can result in enhanced polarization within the material. This phenomenon affects the energy level shifting, the positions of the valence band and the conduction band to the Mulliken theory in electronegativity [31]. Consequently, it influences the electronic properties and overall performances [32].

The selected third characteristic is the ionization energy, which is required to eject an electron away from an ion or an atom. Specifically, the ionization energy of the A-site cation affects the lattice strain, which impacts the bandgap indirectly. Materials with lower ionization energy tend to have more shallower valence bands, leading easier electron excitation under solar illumination. This leads to narrower band gaps and enhanced photon absorption. Conversely, high ionization energy results in deeper valence bands and wider bandgaps [33]. Therefore, regulating ionization energy is important for optimizing light absorption and charge generation in PSCs.

2.3 Model selection

In selecting the model, different ML models, including Extreme Gradient Boost (XGB), Random Forest (RF), Gradient Boost Regression (GBR) and Ada Boost Regression (ABR) were compared, and the best suitable highly accurate models were selected to be further optimized. Each model has its typical data ideal for data processing, and sampling verification methods like cross-validation and independent test results can determine each model?s accuracy. The optimal model for the data set has been selected out of these methods.

2.4 Model training

The model learns to recognize patterns within training data through an iterative process of parameter adjustment in a manner that reduces differences between observed and predicted values. Here, the data was divided into two sub-datasets; 75% of the data was taken for model training, and the remaining 25% was used to test the performance of the model. This split makes it easy for the model to be tested on unseen data, hence allowing for a better estimate of the capacity of the model to be generalized.

2.5 Model evaluation

ML model evaluation is a process of assessing the ability of the trained model to perform on novel or unseen data. This evaluation confirms that the model generalizes well beyond the trained dataset and delivers accurate and reliable predictions in real-world scenarios. The commonly available model evaluation methods are independent tests, cross validations, and bootstrapping. In this study, the generalization error of the method was evaluated using the independent test. Because the model aims to predict unknown samples, a testing set is necessary to assess the generalization ability of the model accurately. The error obtained from the testing set serves as an approximation of the generalization error. As this study involves regression tasks, regression models were utilized to evaluate the model, including Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2.

2.6 Hyperparameter selection

To optimize model performance and ensure robust generalization, hyperparameter tuning was conducted for each machine learning model using grid searching method. The optimized hyperparameters are tabulated in Table 1.

2.7 Model application

Various new perovskite materials were generated, systematically exploring the possible combinations of elements to predict the selected critical properties.

The elements highlighted in Fig. 2 indicate the candidates selected for substitution at the A, B, and X sites of the ABX_3 perovskite structure, based on established perovskite chemistry principles. A-site cations (shown in brown) are typically large monovalent ions such as K^+, Rb^+, Cs^+ These ions occupy the cuboctahedra voids in the perovskite lattice and contribute to structural stability through size matching, assessed via the Goldschmidt tolerance factor [34]. In addition to these selected A site cations, MA^+ ($CH_3NH_3^+$) and FA^+ ($HC(NH_2)_2^+$) were also selected as suitable candidates due to their proven compatibility with the perovskite crystal structure and their widespread use in high-performance hybrid PSCs.

Cations such as Bi^{3+}, Zn^{2+}, Ge^{2+}, and Sn^{2+}, which have different valence values, were chosen as the B-site (green). These cations play a direct role in defining the electronic structure and bandgap of the material. X-site anions (blue), primarily halides such as Cl^-, Br^-, and I^- bridge B-site cations and influence band edge positions, optical absorption, and ionic mobility [35]. Some elements, such as In, Ba, and Ag, have suitable ionic sizes and oxidation states that permit them to occupy either the A site or the B site in the ABX_3 perovskite structure, depending on the specific composition and crystal chemistry. Hence, they are colored with a combination of brown (A site) and green (B site) in Fig. 2 to represent their dual-site potential.

Table 1 Values of the optimized hyperparameters used in each model

Model	List of optimized hyperparameters	Values of optimized hyperparameters
GBR	n_estimators, learning rate	n_estimators=100, learning rate=0.1
XG Boost	n_estimators, learning rate	n_estimators=100, learning rate=0.1
ABR	max_depth, n_estimators	max_depth=3, n_estimators=100
RF	n_estimators, verbose	n_estimators=1000, verbose=1

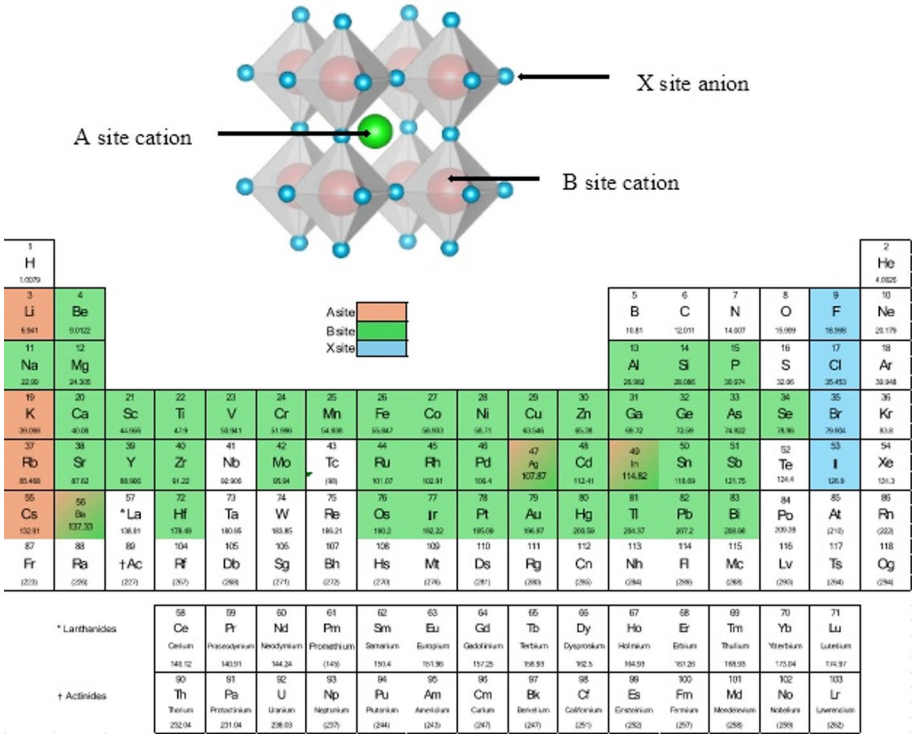


Fig. 2 The standard perovskite structure and selected elements for A, B, and X sites of the perovskites. A-site cations in brown, B-site cations in green, and X-site anions in blue

Table 2 The data filtering criteria

Goldsmith Tolerance Factor	0.81 to 1.00
Bandgap Range	1.3 to 1.6 eV

By exploring all possible combinations of these elements, 1,680 unique perovskite materials were generated through a systematic algorithm that iterates over all possible combinations of selected components. Then, the filtering criteria were applied based on the Goldsmith tolerance factor and the bandgap range (Table 2) to identify the formable perovskite materials with optimal bandgaps for solar cell applications.

The selected elements were chosen based on a combination of factors, namely formability in perovskite-type structures, compatibility with solution processing, non-toxicity, and abundance. This element selection strategy supports the generation of a diverse and realistic dataset for machine learning predictions and material screening.

The toxicity and the abundance of the elements were considered to filter out the final set of perovskite materials. To evaluate toxicity and resource abundance, we employed several indicators. For toxicity, four primary measures were utilized: the elements' radioactivity, compliance with the European Union's Restriction of Hazardous Substances (RoHS) directive, the attributes of concern listed by the European Chemical Agency (ECHA), and the compound toxicity.

To evaluate the resource abundance of the selected materials, global crustal abundance values were obtained from the United States Geological Survey. Furthermore, material criticality and supply risk information were based on the European Commission's Critical Raw Materials Report (2023). An algorithm based on these indicators,

Table 3 ML model type with their corresponding result for goldsmith tolerance factor prediction

ML model	MSE	MAE	R^2
XG Boost	0.00014	0.0070	0.9850
GBR	0.00017	0.0092	0.9821
RF	0.00036	0.0098	0.9629
ABR	0.00129	0.0272	0.8693

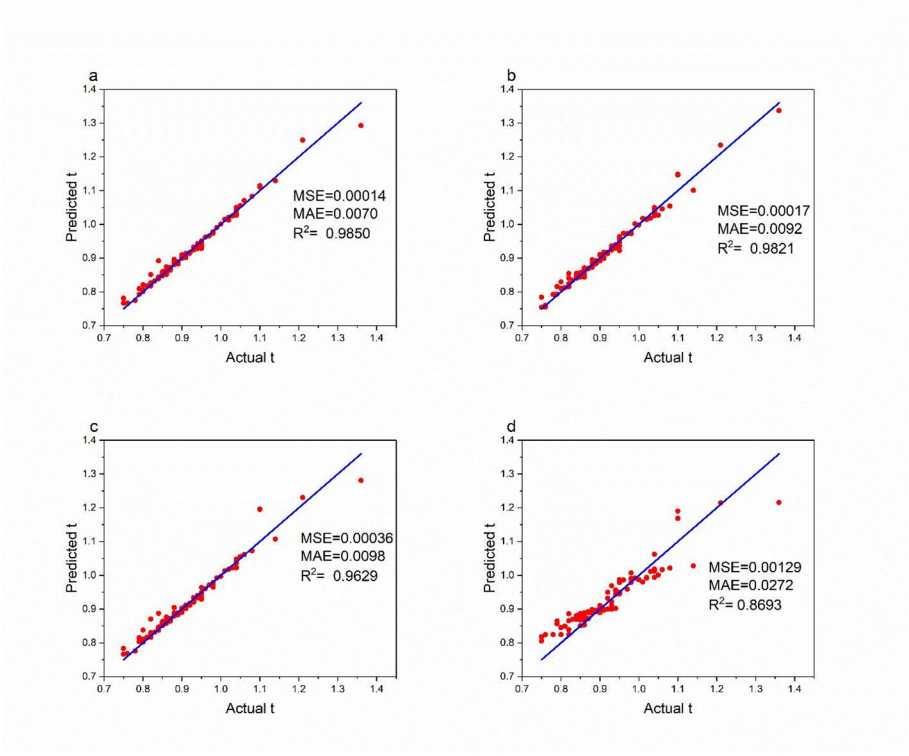


Fig. 3 Demonstration of results of different ML models used in predicting Goldsmith tolerance factor (**a**) XGBoost (**b**) GBR (**c**) RF (**d**) ABR. The red dots represent the data points, and the blue line represents the ideal case of prediction (prediction=true data)

both non-toxic and abundant perovskite materials, was selected, ensuring the safety and sustainability of solar cell applications.

3 Results and discussion

3.1 Goldsmith tolerance factor prediction

The first subset of 537 data points was built to predict the Goldsmith tolerance factor of the perovskite material, which determines its structural formability. For the selected models, the results of regression tasks are depicted in Table 3. According to the results, the XGBoost model attains the lowest MSE and MAE values of 0.00014 and 0.0070, respectively, which are nearest to zero alongside the highest R^2 value of 0.9850, approaching one.

Figure 3 presents the actual versus predicted values for the Goldschmidt tolerance factor. Notably, the XGBoost model demonstrates a concentration of data points close to the central reference line, revealing higher accuracy compared to other models. These findings exhibit that XGBoost delivers the most precise and reliable predictions over other evaluated models for Goldsmith tolerance factor prediction.

Table 4 ML model type with their corresponding result for band gap

ML model	MSE (eV^2)	MAE (eV)	R^2
GBR	0.4531	0.6169	0.7815
ABR	0.9140	1.2053	0.5733
XGBoost	0.6005	1.5574	0.4486
RF	0.6486	1.5359	0.4562

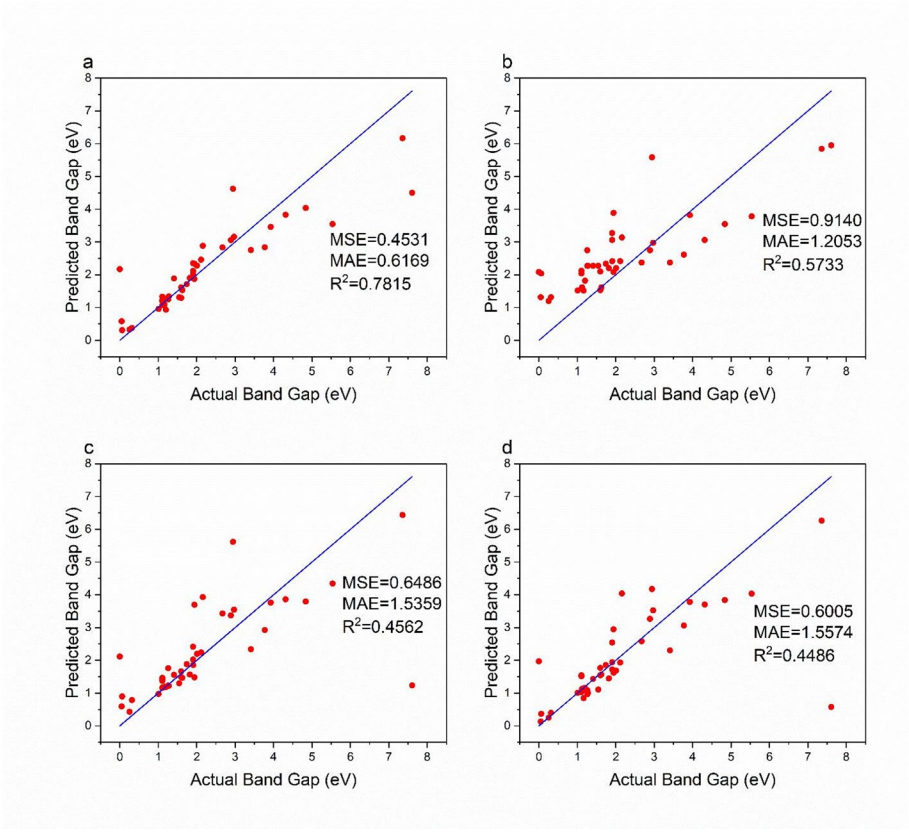


Fig. 4 Demonstration of results of different ML models used in predicting band gap (**a**) GBR (**b**) ABR (**c**) RF (**d**) XGBoost

3.2 Band gap prediction

The band gap is a crucial parameter in the exploration and development of novel perovskite materials for photovoltaic applications. Consequently, the second subset of data with 218 data points was curated to predict the bandgap of the perovskite materials.

As per the comparison of four different models indicated in Table 4, The Gradient Boost Regression (GBR) stands out as the best suited model, achieving the lowest MSE and MAE values of 0.4531 eV^2 and 0.6169 eV respectively, along with the highest R^2 value of 0.7815. Figure 4 illustrates the actual bandgap value vs. the predicted bandgap values. The tight clustering of red dots around the ideal prediction line compared to other models demonstrates the effectiveness of the GBR model in predicting the band gap values.

Nonetheless, the accuracy of the models in predicting bandgap parameters, reflected by the MAE, MSE, and R^2 values, does not reach the high levels observed for tolerance factor predictions. This may be attributed to the smaller dataset available for bandgap

predictions, which has 218 data points. Small datasets are particularly susceptible to overfitting, where models learn noise rather than the underlying data patterns. GBR can include regularization techniques that help mitigate this risk. While XGBoost also has strong regularization capabilities, its complexity can make it more problematic to tune effectively fig. 4 for smaller datasets. Further, the XG Boost cannot harness the full potential of a limited dataset without hyperparameter tuning. In RF, the model can suffer from overfitting with small-sized datasets as it relies on the aggregation of multiple decision trees.

3.2.1 Band gap model optimization

A significant deviation between the predicted and actual band gap values was observed in the selected model. The K-nearest neighbors (KNN) model was employed to address this disparity, proving the highest accuracy for this process.

Using the actual band gap values as outputs and the predicted values as inputs from the GBR model, a relationship equation between the two sets of values was generated using the KNN model. Subsequently, this equation was applied to the predicted band gap values of newly found perovskites, resulting in optimized band gap values. The scatter plot after optimization is displayed in Fig. 5.

The optimized bandgap gap values are further aligned with the ideal prediction line, exhibiting an improvement in accuracy and optimization. This is reflected in the fact that the MAE decreases from 0.4531 eV to 0.0421 eV while MSE values decrease from 0.6169 eV^2 to 0.1088 eV^2, approaching zero. In addition to that, the R^2 value has significantly enhanced from 0.7821 to 0.9873, moving closer to 1, exhibiting a strong correlation between the actual value and the optimized band gap value.

It has been reported that the highest accuracy achieved by the artificial neural network (ANN) model for bandgap prediction yielded an R^2 value of 0.9409 [36]. Our optimized

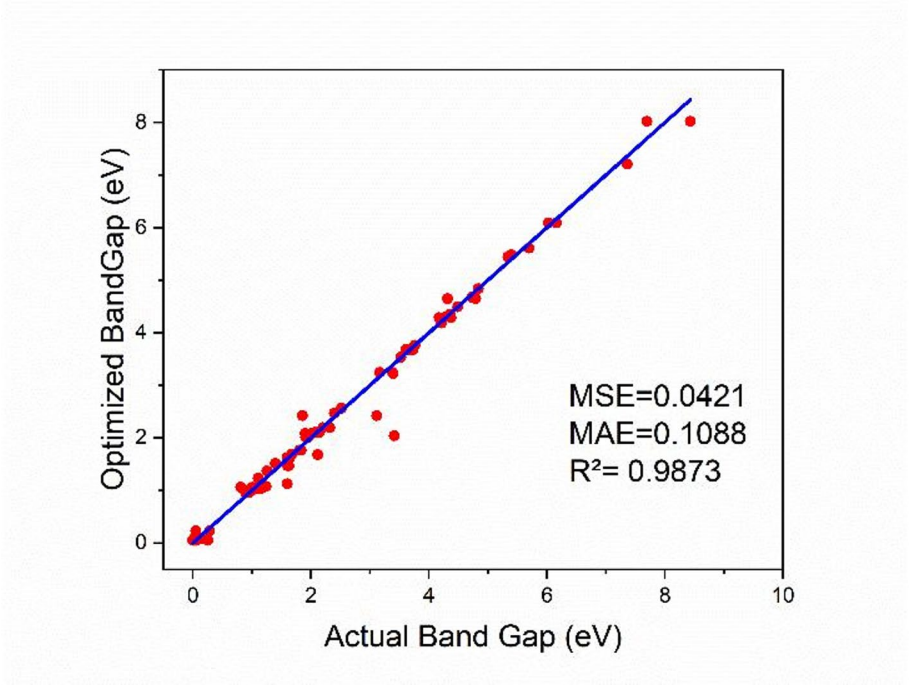


Fig. 5 Scatter plot of actual band gap vs. optimized band gap after optimization

Gradient Boosting Regression (GBR) model, combined with the K-nearest neighbors (KNN) model, outperformed this result, achieving an R^2 value of 0.9873. A more recent study by Subham and Chatterjee also reported an R^2 value of 0.9216 using the CatBoost Regression for bandgap prediction [23].

3.3 Final selection of optimal perovskite materials

Using the trained models, band gaps and Goldsmith tolerance factor values for all 1628 perovskite materials were predicted. The optimum range for band gap and Goldsmith tolerance factor values were determined based on the findings of the literature. It was identified that 0.81-1.00 is the optimal range of the Goldsmith tolerance factor for the stable formation of perovskite and the band gap within the 1.3–1.6 eV range for photovoltaic application [27]. This criterion filtered out 139 perovskite materials. Subsequently, 49 lead-free materials were selected after assessing their toxicity and abundance.

3.3.1 Principal component analysis and data cluster analysis

After identifying 49 nontoxic perovskite materials, principal component analysis (PCA) was performed to retain the essential information. PCA facilitates understanding the relationships among the different properties, reducing the complexity of the dataset and highlighting the key factors. To further refine the dataset, K-means clustering was applied to divide the materials into several groups according to their similarity in shared parameters such as band gap and structural compatibility.

This method enabled the classification of materials into distinct groups with corresponding properties, further streamlining the selection process. The combined PCA and K-means clustering approach is instrumental for narrowing down identified perovskite materials for further analysis.

As per Fig. 6; Table 5, the dataset has been divided into four distinct clusters (clusters 0, 1,2, and 3). Table 5 illustrates a detailed list of materials assigned to each cluster. According to the reports, the optimum band gap value for the best performance of the PSC is recorded as 1.53 eV to 1.56 eV for single junction solar cells [37]. Therefore, clusters 0 and 2 can be excluded as the predicted bandgaps are outside the optimum range. This exclusion narrows the focus to Cluster 1 and Cluster 3, which consists of perovskite materials with band gap values closer to the ideal range, making them more appropriate in photovoltaic applications. If the synthesis feasibility is taken into consideration, cluster 3 can be isolated as it contains Al, Bi, Zn, K, Rb, Ge, which are abundant and cost-effective.

Further, there are established methods for fabricating iodide and bromide-based perovskites, making these materials easy to incorporate into existing manufacturing systems. In cluster 1, only five perovskite materials have bandgap values that lie between the ideal bandgap range. However, four of them are Ag based, making them less feasible for large-scale applications as Ag is quite expensive. This economic limitation further reduces the practical applicability of the cluster 1 materials for photovoltaic applications. As a result of the above reasons, six materials, including $KBiBr_3$, $KZnBr_3$, $RbBiBr_3$, $RbZnBr_3$, $MAGeI_3$, $FAGeI_3$ from cluster 3 can be selected for further research.

A couple of studies have reported the synthesis of hybrid Ge iodide perovskites using complex methods such as hot co-precipitation, which is a multi-step process requiring elevated temperatures and an inert reaction atmosphere, often involving hazardous

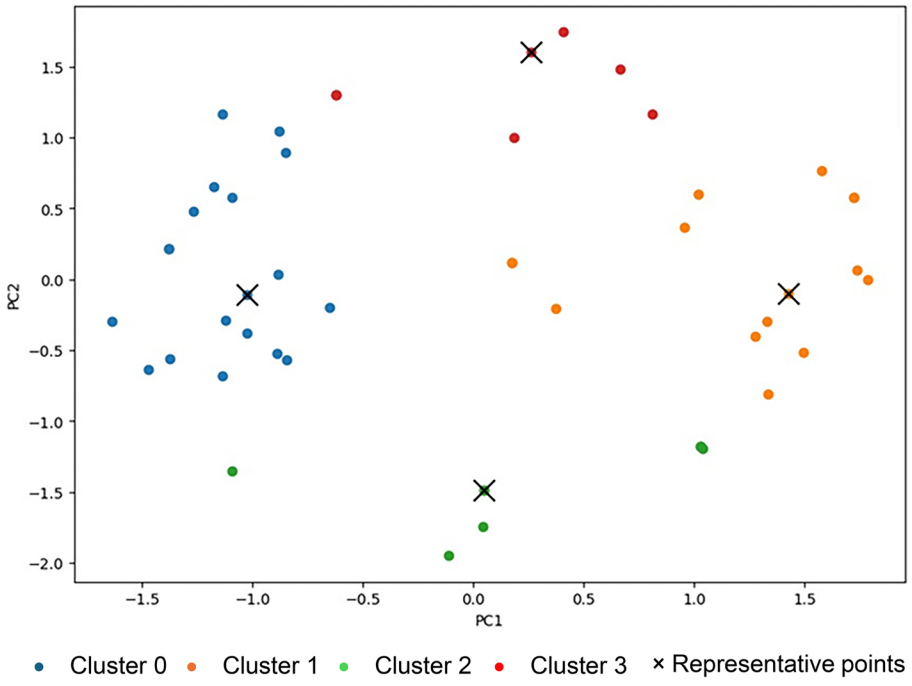


Fig. 6 Clusters and Representative points

Table 5 Cluster based material classification after PCA

Cluster 0	$AgSiBr_3$, $CsBiI_3$, $CsTiBr_3$, $CsVBr_3$, $CsYBr_3$, $CsZnI_3$, $KBiI_3$, $KGaI_3$, $KScBr_3$, $KTiBr_3$, $KVBr_3$, $KZnI_3$, $RbBiI_3$, $RbScBr_3$, $RbTiBr_3$, $RbVBr_3$, $RbZnI_3$, $MANaF_3$, $FANaF_3$
Cluster 1	$AgGaI_3$, $AgHfI_3$, $AgScBr_3$, $AgZnI_3$, $CsAgCl_3$, $CsAgBr_3$, $InSiI_3$, $KAgCl_3$, $KAgBr_3$, $RbAgCl_3$, $RbAgBr_3$, $MANaI_3$, $MASrBr_3$, $FANaI_3$, $FASrBr_3$
Cluster 2	$AgCrBr_3$, $AgFeBr_3$, $AgInBr_3$, $AgMgBr_3$, $AgScI_3$, $AgZrBr_3$, $InAlI_3$, $RbYBr_3$
Cluster 3	$AgAlI_3$, $KBiBr_3$, $KZnBr_3$, $RbBiBr_3$, $RbZnBr_3$, $MAGeI_3$, $FAGeI_3$

substances like hypophosphorous acid (H_3PO_2) and concentrated hydroiodic acid (HI) [38]. Another method, the hot-injection technique, also relies on relatively high temperatures and must be conducted under vacuum conditions [39]. In addition to that, the relatively complex solvothermal synthesis of 0D $Rb_7Sb_3Br_{16}$ has also been reported [40]. Nevertheless, Shiyu Yue and coworkers have demonstrated a simpler, room-temperature reactions using generalized and adaptable methods to synthesize Ge-based perovskites ($MAGeI_3$, and $FAGeI_3$) [41]. These findings confirm the experimental feasibility of synthesizing the identified potential materials. Therefore, future research is encouraged to explore and develop facile, scalable manufacturing techniques for these newly identified compounds.

4 Summary and conclusions

Our research addresses the significant challenges and opportunities associated with replacing lead in perovskite solar cells by applying machine learning. The findings suggest XGBoost is the most precise and reliable model in predicting the Goldsmith tolerance factor, and GBR is the best model for bandgap prediction. The accuracy of the bandgap prediction model was enhanced by utilizing the KNN model, achieving over 98% accuracy. By incorporating health and sustainability perspectives, potential candidates were effectively filtered out from the vast compositional range of perovskites.

Forty-nine ABX_3 perovskite materials were identified as promising alternatives to lead-containing perovskites. Then, utilizing PCA and K-means clustering six lead free materials, including $KBiBr_3$, $KZnBr_3$, $RbBiBr_3$, $RbZnBr_3$, $MAGeI_3$, and $FAGeI_3$ were identified as optimal candidates for photovoltaic application due to their synthesis feasibility and ideal bandgap. In conclusion, our research emphasizes the pivotal role of machine learning in identifying new materials in promoting sustainable energy technologies.

Author contributions
W.G.A.P wrote the main manuscript text, conceptualized, prepared figures, H.A.H.M.W. and M.G.M.M. K. wrote the main manuscript text, carried out machine learning calculations, D.M.C. S. data curation, validation, P.K.W. A. and G.A.S. supervision, wrote and edited manuscript text. All authors reviewed the manuscript.

Funding
The authors did not receive support from any organization for the submitted work.

Data availability
Data will be made available on request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Permission from all relevant authors and contributors obtained to publish this work.

Competing interests
The authors declare no competing interests.

Received: 8 May 2025 / Accepted: 3 July 2025
Published online: 06 July 2025

References

1. Zhang Q, Hao F, Li J, Zhou Y, Wei Y, Lin H. Perovskite solar cells: must lead be replaced–and can it be done? Sci Technol Adv Mater. 2018;19:425–42. Available from: https://doi.org/10.1080/14686996.2018.1460176
2. Kabir E, Kumar P, Kumar S, Adelodun AA, Kim KH. Solar energy: potential and future prospects. Renew Sustain Energy Rev. 2018;82:894–900.
3. Han Y, Zhao Z, Zhang Y, Yang X, Wang B, Shen Y. Developing a predictive model for the maximum power conversion efficiency of inorganic perovskites: A combined approach using density functional theory and machine learning. Comput Mater Sci. 2024;245:113325. https://www.sciencedirect.com/science/article/pii/S0927025624005469.
4. Green MA, Dunlop ED, Yoshita M, Kopidakis N, Bothe K, Siefer G, et al. Solar cell efficiency tables (Version 66). Prog Photovoltaics Res Appl. 2025;33:795–810. https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3919.
5. Rathore N, Panwar NL, Yettou F, Gama A. A comprehensive review of different types of solar photovoltaic cells and their applications. Int J Ambient Energy [Internet]. 2021;42:1200–17. https://doi.org/10.1080/01430750.2019.1592774.
6. Wang M, Wang W, Ma B, Shen W, Liu L, Cao K, et al. Lead-Free perovskite materials for solar cells. Nano-Micro Lett Springer Singap. 2021;1–36. https://doi.org/10.1007/s40820-020-00578-z.
7. Rhee S, An K, Kang K-T. Recent advances and challenges in halide perovskite crystals in optoelectronic devices from solar cells to other applications. Crystals. 2021;11:39. https://doi.org/10.3390/cryst11010039.
8. Gamage Ayomi Pabasara W, Asha Sewvandi G. Numerical simulation and optimization of stable CH3NH3PbI3-based 2D/3D mixed dimensional perovskite solar cell. Complex system research centre, ni?, serbia; mathematical Institute of the Serbian academy of sciences and arts. Serbia; 2025. https://doi.org/10.5281/zenodo.14730922.
9. Zhang Z, Liu Y, Sun Q, Ban H, Liu Z, Yu H, et al. The importance of elemental lead to perovskites photovoltaics. Chem Inorg Mater. 2023;1:100017. https://www.sciencedirect.com/science/article/pii/S2949746923000174.
10. Jayawardane ST, Akmal MD, Jayaneththi YH, Fernando TV, Hu D, Abeygunawardhana PKW, et al. Simulation-Based performance analysis of Lead-Free bismuth perovskite solar cells: A comparative study of Cs3Bi2I9 and (CH3NH3)3Bi2I9 -based perovskite solar cells. Adv Theory Simulations. 2024;7(7):2400206. https://doi.org/10.1002/adts.202400206. https://advanced.onlinelibrary.wiley.com/doi/abs/.
11. Pabasara WGA, Akmal UKDM, Wickramaarachchi WAAS, Swvandi GA. Numerical Simulation of Lead-free Bismuth Chalcohalide based Perovskite Solar Cells. In: 2024 Moratuwa Engineering Research Conference (MERCon). 2024:436–41.
12. Su P, Liu Y, Zhang J, Chen C, Yang B, Zhang C, et al. Pb-Based perovskite solar cells and the underlying pollution behind clean energy: dynamic leaching of toxic substances from discarded perovskite solar cells. J Phys Chem Lett. 2020;11:12–7. https://doi.org/10.1021/acs.jpclett.0c00503.
13. Ravi VK, Mondal B, Nawale VV, Nag A. Don't let the lead out: new material chemistry approaches for sustainable lead halide perovskite solar cells. ACS Omega. 2020;5:29631–41. https://doi.org/10.1021/acsomega.0c04599.
14. Siddiqua A, Hahladakis JN, Al-Attiya WAKA. An overview of the environmental pollution and health effects associated with waste landfilling and open dumping. Environ Sci Pollut Res. 2022;29:58514–36. Available from: https://doi.org/10.1007/s11356-022-21578-z

15. Fu F, Pisoni S, Jeangros Q, Sastre-Pellicer J, Kawecki M, Paracchino A, et al. I2 vapor-induced degradation of formamidinium lead iodide based perovskite solar cells under heat-light soaking conditions. Energy Environ Sci. 2019;3074–88. https://doi.org/10.1039/C9EE02043H.
16. Song R, Zhao R. Density functional theory study of two-dimensional hybrid organic-inorganic perovskites: frontier level alignment and chirality-induced spin splitting. AAPPS Bull. 2024;34(1). https://doi.org/10.1007/s43673-024-00125-7.
17. Shah M, Ahmad I, Hayat K, Munawar M, Mushtaq M, Ahmad W, et al. Utilizing density functional theory and SCAPS simulations for modeling High-Performance MASnI3-Based perovskite solar cells. Energy Technol. 2024;1:2301228. https://doi.org/10.1002/ente.202301228. https://onlinelibrary.wiley.com/doi/abs/.
18. Hussain W, Sawar S, Sultan M. Leveraging machine learning to consolidate the diversity in experimental results of perovskite solar cells. RSC Adv. 2023;13(32):22529–37. https://doi.org/10.1039/D3RA02305B.
19. Li J, Pradhan B, Gaur S, Thomas J. Predictions and strategies learned from machine learning to develop High-Performing perovskite solar cells. Adv Energy Mater. 2019;9:1–10. https://doi.org/10.1002/aenm.201901891.
20. Dinic F, Neporozhnii I, Voznyy O. Machine learning models for the discovery of direct band gap materials for light emission and photovoltaics. Comput Mater Sci. 2024;231:112580. https://www.sciencedirect.com/science/article/pii/S0927025623005748.
21. Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci [Internet]. 2017;129:156–63. https://www.sciencedirect.com/science/article/pii/S0927025616306188.
22. An R, Xie C, Chu D, Li F, Pan S, Yang Z. A Machine-Learning-Assisted crystalline structure prediction framework to accelerate materials discovery. ACS Appl Mater Interfaces. 2024;16:36658–66. https://doi.org/10.1021/acsami.4c10477.
23. Subba S, Chatterjee S. Machine learning-driven determination of key absorber layer parameters in perovskite solar cells. Mater Today Commun. 2025;42:111113. https://doi.org/10.1016/j.mtcomm.2024.111113.
24. Agiorgousis L, Sun M, Choe Y-Y, West D-H, Zhang D. Machine learning augmented discovery of chalcogenide double perovskites for photovoltaics. Adv Theory Simulations. 2019;2:1800173. https://doi.org/10.1002/adts.201800173. https://onlinelibrary.wiley.com/doi/abs/.
25. Touati S, Benghia A, Hebboul Z, Lefkaier IK, Kanoun MB, Goumri-Said S. Machine learning models for efficient property prediction of ABX3 materials: A High-Throughput approach. ACS Omega. 2024;9:47519–31. https://doi.org/10.1021/acsomega.4c06139.
26. Gebhardt J, Gassmann A, Wei W, Weidenkaff A, Elsässer C. Screening for sustainable and lead-free perovskite halide absorbers? A database collecting insight from electronic-structure calculations. Mater Des. 2023;234:112324. https://www.sciencedirect.com/science/article/pii/S0264127523007396.
27. Bartel CJ, Sutton C, Goldsmith BR, Ouyang R, Musgrave CB, Ghiringhelli LM, et al. New tolerance factor to predict the stability of perovskite oxides and halides. Sci Adv. 2019;5:eaav0693. https://doi.org/10.1126/sciadv.aav0693.
28. Stanley JC, Mayr F, Gagliardi A. Machine learning stability and bandgaps of Lead-Free perovskites for photovoltaics. Adv Theory Simulations. 2020;3:1900178. https://doi.org/10.1002/adts.201900178. https://onlinelibrary.wiley.com/doi/abs/.
29. Miah MH, Khandaker MU, Rahman MB, Nur-E-Alam M, Islam MA. Band gap tuning of perovskite solar cells for enhancing the efficiency and stability: issues and prospects. RSC Adv. 2024;14:15876–906. https://doi.org/10.1039/d4ra01640h.
30. Pitaro M, Tekelenburg EK, Shao S, Loi MA. Tin halide perovskites: from fundamental properties to solar cells. Adv Mater. 2022;34. https://doi.org/10.1002/adma.202105844.
31. Wu Y, Zhang J, Long B, Zhang H. The stability and electronic and photocatalytic properties of the ZnWO4(010) surface determined from first-principles and thermodynamic calculations. RSC Adv. 2021;11:23477–90. https://doi.org/10.1039/d1ra03218f.
32. Smith WA, Sharp ID, Strandwitz NC, Bisquert J. Interfacial band-edge energetics for solar fuels production. Energy Environ Sci. 2015;8(10):2851?62. https://doi.org/10.1039/C5EE01822F.
33. Zheng C, Rubel O. Ionization energy as a stability criterion for halide perovskites. J Phys Chem C. 2017;121:11977–84. https://doi.org/10.1021/acs.jpcc.7b00333.
34. Wu T, Wang J. Global discovery of stable and non-toxic hybrid organic-inorganic perovskites for photovoltaic systems by combining machine learning method with first principle calculations. Nano Energy. 2019;66:104070. Available from: https://www.sciencedirect.com/science/article/pii/S2211285519307773
35. Cai X, Zhang Y, Shi Z, Chen Y, Xia Y, Yu A, et al. Discovery of Lead-Free perovskites for High-Performance solar cells via machine learning: ultrabroadband absorption, low radiative combination, and enhanced thermal conductivities. Adv Sci. 2022;9(4):1–15. https://doi.org/10.1002/advs.202103648.
36. Li J, Pradhan B, Gaur S, Thomas J. Perovskite solar cells: predictions and strategies learned from machine learning to develop High-Performing perovskite solar cells. Adv Energy Mater. 2019;9:1970181. https://doi.org/10.1002/aenm.201970181. https://onlinelibrary.wiley.com/doi/abs/.
37. Yang X, Li L, Tao Q, Lu W, Li M. Rapid discovery of narrow bandgap oxide double perovskites using machine learning. Comput Mater Sci [Internet]. 2021;196:110528. https://www.sciencedirect.com/science/article/pii/S092702562100255X.
38. Stoumpos CC, Frazer L, Clark DJ, Kim YS, Rhim SH, Freeman AJ, et al. Hybrid germanium iodide perovskite semiconductors: active lone pairs, structural distortions, direct and indirect energy gaps, and strong nonlinear optical properties. J Am Chem Soc. 2015;137:6804–19. https://doi.org/10.1021/jacs.5b01025.
39. Wu X, Song W, Li Q, Zhao X, He D, Quan Z. Synthesis of Lead-free CsGeI3 perovskite colloidal nanocrystals and Electron Beam-induced transformations. Chem? Asian J. 2018;13:1654–9. https://aces.onlinelibrary.wiley.com/doi/abs/10.1002/asia.201800573.
40. McCall KM, Benin BM, Wörle M, Vonderach T, Günther D, Kovalenko MV. Expanding the 0D Rb7M3X16 (M?=?Sb, bi; x?=?br, I) family: Dual-Band luminescence in Rb7Sb3Br16. Helv Chim Acta. 2021;104. https://doi.org/10.1002/hlca.202000206.
41. Yue S, McGuire SC, Yan H, Chu YS, Cotlet M, Tong X, et al. Synthesis, characterization, and stability studies of Ge-Based perovskites of controllable mixed cation composition, produced with an ambient Surfactant-Free approach. ACS Omega. 2019;4:18219–33. https://doi.org/10.1021/acsomega.9b02203.

Publisher?s note