ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Machine Learning Models for Efficient Property Prediction of ABX_3 Materials: A High-Throughput Approach

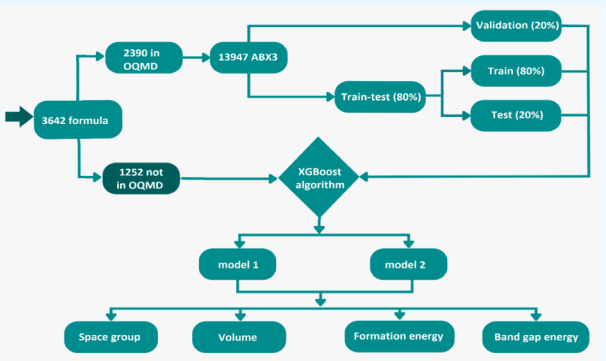
Soundous Touati, Ali Benghia, Zoulikha Hebboul, Ibn Khaldoun Lefkaier, Mohammed Benali Kanoun, and Souraya Goumri-Said*

Cite This: *ACS Omega* 2024, 9, 47519−47531

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Recently, ABX_3 materials have garnered significant attention due to their diverse applications in photovoltaics, catalysis, and optoelectronics as well as their remarkable efficiency in energy conversion. However, progress has been somewhat slow due to the high expenses of the experiment or the time-consuming density functional theory (DFT) calculation. In this study, we utilized the extreme gradient boosting (XGBoost) algorithm to facilitate the discovery and characterization of ABX_3 compounds based on vast data sets generated by DFT calculations. While the XGBoost algorithm provides a powerful tool for accelerating the discovery of ABX_3 compounds, it is crucial to acknowledge that different DFT approximation levels can significantly impact the predicted band gaps, potentially introducing discrepancies when compared with experimental values. In the first step, we predict the space group of 13947 oxides and halides using the Open Quantum Materials Database and elemental features. Our analysis yields classification accuracies ranging from 82.39% to 99.14% across these materials. Following this, XGBoost regression algorithms are employed to interrogate the data set, enabling predictions of volume (achieving an optimal accuracy of 98.41%, with a mean absolute error (MAE) of 2.395 $Å^3$ and a root-mean-square error (RMSE) of 4.416 $Å^3$), formation energy (an optimal accuracy of 97.36%, with an MAE of 0.075 eV/atom and an RMSE of 0.132 eV/atom), and band gap energy (an optimal accuracy of 87.00%, an MAE of 0.391 eV, and an RMSE of 0.574 eV). Finally, these prediction models are employed to identify the possible space groups for each of the 1252 new ABX_3 formulas. Then, we predict the volume, the formation energy, and the band gap energy for each candidate space group. Through these predictive models, machine learning accelerates the exploration of new materials with enhanced performance and functionality.



1. INTRODUCTION

Machine learning has become an effective tool for accelerating the discovery of new ABX_3 materials, revolutionizing research in materials science.[1−4] ABX_3 compounds, which have perovskite crystal structures, are applicable in various fields, including catalysis,[5−7] light-emitting diodes,[8−10] superconductivity,[11−13] piezoelectricity,[14,15] ferromagnets,[16] ferroelectrics,[14] energy storage, and solar cells.[17−20] Traditional methods of material discovery often rely on onerous trial-and-error experiments, aiming to find a material possessing desired properties, or density functional theory (DFT), which offers a computationally expensive and time-consuming method to predict material behavior accurately. By contrast, machine learning uses complex systems or vast data sets to predict material properties, significantly speeding the discovery process and unlocking the full potential of ABX_3 compounds in an advancing technological field.

To understand the ABX_3 material's performance, first, predicting the space group aids in understanding crystal material properties. Li et al.[21] utilized a random forest model with data from the Material Project Database to predict space groups based on crystal material formulas. They achieved performance rates ranging from 67% to 92% across the 14 space groups examined. Nomura et al.[22] used machine learning models to predict the space groups for $Ba(Ce_{0.8-x}Zr_x)Y_{0.2}O_3$ perovskite, achieving 94% accuracy across the space groups considered. However, these models determine only the most probable space group for each formula. Second, in crystallography, the lattice constant significantly influences material identification. X-ray diffraction is often a straightforward method for determining it with high accuracy and an expensive cost. To obtain the lattice constant or volume without using experimental methods, researchers utilize data mining

Received: July 2, 2024
Revised: October 24, 2024
Accepted: October 29, 2024
Published: November 18, 2024

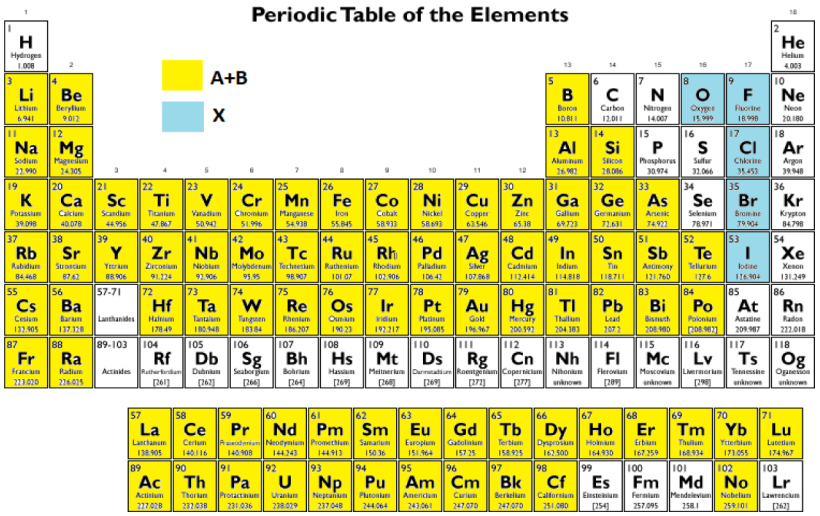


Figure 1. A and B cations and X anions of ABX_3 compounds.

techniques because they are among the most extensively utilized methods in the world of research for extracting information from a large amount of data and organizing it for greater use. Majid et al.[23] utilized support vector regression and neural networks to predict the lattice constant of perovskites within two crystal systems, monoclinic and cubic. They achieved performance rates ranging from 97.1% to 99.8%. Li et al.[24] utilized the random forest model with a novel descriptor to predict lattice constants specifically for cubic crystals, achieving a performance accuracy of 97.3%. However, when data from other crystal systems (orthorhombic, tetragonal, and trigonal) were aggregated, the overall accuracy decreased to 69.9%. That indicates a reduction in the model accuracy when combining space groups. Within our model, we accurately forecasted the volume across 103 space groups with a precision exceeding 97%.

Formation energy (E_f) is generally used to select thermodynamically stable materials with desirable properties.[25,26] Stability conditions also aid in identifying compounds that are more resistant to phase transitions or degradation, ensuring reliability and long-term performance. The band gap energy (E_g) is an essential property of materials; it affects their electronic behavior and makes them applicable in a wide range of fields, including photovoltaics,[27] optoelectronics,[28] and semiconductors,[29] because E_g directly influences the light wavelength range that can absorb and control the solar cell efficiency. However, predicting the band gap energy effectively poses a challenge due to the complicated electrical interactions and structural intricacies within the materials. Thus, many researchers were interested in E_f and E_g. For example, Im et al.[30] predicted the formation energy for 540 hypothetical double perovskites with an RMSE of 0.021 eV/atom to get the solar cell perovskites used. Li et al.[31] use ML models (GBR, bagging, SVR, and RF) to predict the formation and band gap energies for 758 perovskites. Zhang et al.[32] use random forest regression to predict the band gap of 1306 double perovskites with an accuracy of 85.6% and MSE = 0.64 eV. Gao et al.[33] employ three machine learning models (XGBR, ANN, and SVR) to predict the band gap of 745 inorganic double perovskites. While previous studies have applied machine learning to predict individual properties of materials, our work addresses the gap in comprehensive, high-throughput prediction of multiple critical properties for ABX_3 materials. This approach enables rapid screening of a vast compositional space, accelerating the discovery process in a way that traditional DFT-based methods cannot match in terms of speed and scale.

In this investigation, we address existing challenges by developing XGBoost models to predict the critical parameters of ABX_3 materials. Our approach first predicts all feasible space groups for each chemical formula, followed by the prediction of volume (V), formation energy (E_f), and band gap energy (E_g) based on the chemical formulas. We use data from the Open Quantum Materials Database (OQMD) for training our models, without performing additional DFT calculations. Furthermore, we apply these predictive models to elucidate the potential space groups for each of 1252 unexplored ABX_3 formulas and to predict the properties of each candidate space group. This method allows for rapid screening of a large number of potential materials without the need for time-consuming DFT calculations.

2. MATERIALS AND METHODS

2.1. Machine Learning. Machine learning (ML) integrates computer science, mathematics, statistics, and engineering, revolutionizing data analysis by uncovering hidden relationships without human programming. Various machine learning techniques, including supervised and unsupervised learning, assist in proficiently managing data and identifying descriptors associated with targeted attributes. This is especially valuable in rapidly identifying potential solar cell materials and accurately predicting material band gaps.[34−36] Crafting models through supervised learning allows precise anticipation of values for unexplored materials, advancing material discovery. This pursuit of understanding material behavior through ML transcends predictive accuracy, aiming to catalyze developments in science and technology.

2.2. XGBoost. Extreme gradient boosting (XGBoost) is a powerful and widely used distributed gradient boosting tool proposed by Chen and Guestrin.[37] It is an accessible tool for building predictive models for regression and classification tasks.[38]

The core idea of boosting is to iteratively create more accurate models by combining multiple low-accuracy trees and generating a prediction by summing the previous output, where the model is trained to rectify the prediction errors

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

Table 1. Possible Oxidation States of Periodic Table Atoms from the Shannon Ionic Radii Database[42]

oxidation states	1	2	3	4	5
Periodic table elements	Ag, Au, Cs, Cu, Fr, Hg, K, Li, Na, Pd, Rb, Tl	Ag, Am, Ba, Be, Ca, Cd, Co, Cr, Cu, Dy, Eu, Fe, Ge, Hg, Mg, Mn, Nd, Ni, No, Np, Pb, Pd, Pt, Ra, Sm, Sr, Ti, Tm, V, Yb, Zn	Ac, Al, Am, Ag, As, Au, B, Bi, Bk, Ce, Cf, Cm, Co, Cr, Cu, Dy, Er, Eu, Fe, Ga, Gd, Ho, In, Ir, La, Lu, Mn, Mo, Nb, Nd, Ni, Np, Pa, Pd, Pm, Pr, Pu, Rh, Ru, Sb, Sc, Sm, Ta, Tb, Ti, Tl, Tm, U, V, Y, Yb	Am, Bk, Ce, Cf, Cm, Co, Cr, Fe, Ge, Hf, Ir, Mn, Mo, Nb, Ni, Np, Os, Pa, Pb, Pd, Po, Pr, Pt, Pu, Re, Rh, Ru, Si, Sn, Ta, Tb, Tc, Te, Th, Ti, U, V, W, Zr	As, Au, Bi, Cr, Ir, Mn, Mo, Nb, Np, Os, Pa, Pt, Pu, Re, Rb, Ru, Sb, Ta, Tc, U, V, W

made by prior trees. The sum of the classification and regression trees in CARTs yields the final prediction $\hat{y}_i$:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

where $f_k(x_i)$ refers to the output of one tree, K is the tree's number, and F denotes the set of all CART's potential.

XGBoost is a widely used algorithm for classification and regression problems due to its exceptional performance. Also, it is known for its efficient memory usage, which makes it an attractive choice for data science professionals seeking to develop high-performance machine learning models.

2.3. SHapley Additive exPlanations. In data science, SHapley Additive exPlanations (SHAP) are widely used to simplify the interpretation of machine learning model outputs. SHAP values are derived from cooperative game theory, and each feature is assigned an importance value for a particular prediction by considering its contribution across all possible combinations of features. This fair distribution of "credit" among features allows for consistent and theoretically grounded interpretation of model predictions. The SHAP graph visually represents this by illustrating the significance of each feature in a given data set, with descriptors displayed along the horizontal axis and the vertical axis representing a normalized measure of importance, ranging from 0 to 1, to facilitate comparison. To enhance clarity, we incorporate this explanation into the SHAP section, providing readers with a deeper understanding of its fundamental concepts and applications. The feature importance graph arranges all features based on estimating the individual contributions of each one across the model's trees. As a result, this graph aids in providing valuable insights into the importance of different variables in influencing the target variable or overall model performance.[39−41] Using SHAP, data scientists and professionals can explain the predictions made by machine learning models in a human-understandable way by identifying the most significant features and understanding how the model made its final decision.

2.4. Data Collection. In pursuit of enumerating all possible ABX_3 crystal structures from the elements of the periodic table within the oxide and halide families (where X = Br, F, Cl, I, and O), we fill the A and B positions with 81 semimetal or metal atoms (see Figure 1).

Obtaining neutral ABX_3 compounds requires specific oxidation states for the oxides (A-B cation pair: $A^{+1}B^{+5}$,$A^{+2}B^{+4}$,$A^{+3}B^{+3}$,$A^{+4}B^{+2}$,$A^{+5}B^{+1}$) and halides (A-B cation pair: $A^{+1}B^{+2}$,$A^{+2}B^{+1}$) to maintain charge balance with anions. For this purpose, we searched for all possible oxidation states for the yellow atoms, utilizing the Shannon database (see Table 1). Based on these criteria, it is possible to find 1448 halides and 2194 oxides, constituting 3642 chemical formulas. In our study, Shannon ionic radii, introduced by R.D. Shannon in 1976, are employed to describe the effective size of ions within a crystal lattice. These radii, widely used across chemistry, materials science, and crystallography, account for variations based on the ion's charge, coordination number, and oxidation state. This approach provides a standardized measure of ionic size that is critical for accurate modeling and analysis of ionic interactions and crystal structures in our research.

A comprehensive search within the OQMD (Open Quantum Materials Database)[43,44] yielded 2390 of these formulas, which can exist in multiple space groups, resulting in

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

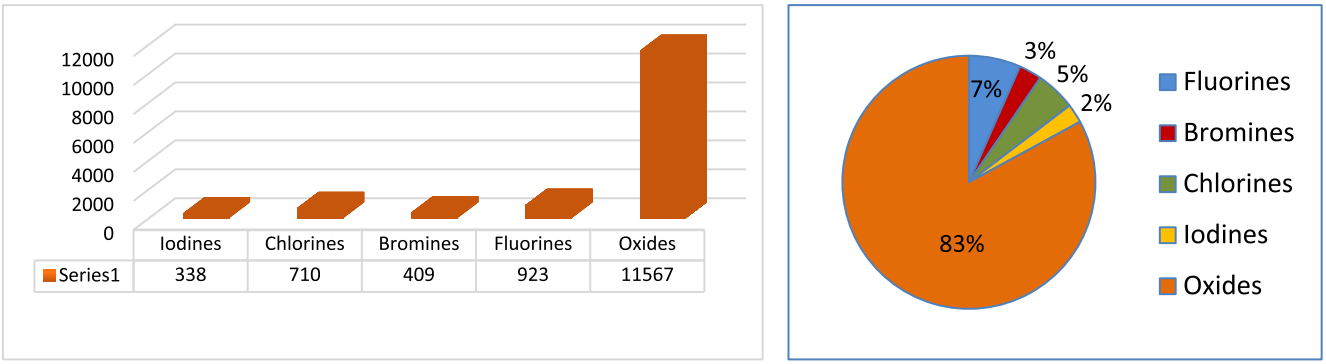


Figure 2. Distribution of 13947 ABX_3 in oxides, fluorines, bromines, chlorines, and iodines.

Table 2. Distribution of Space Groups in the ABX_3 Train-Test Data[a]

Crystal System	Space Group	Occurrences	Crystal System	Space Group	Occurrences	Crystal System	Space Group	Occurrences
Cubic	*Pm*-3m	1805	Tetragonal	*P4mm*	863	Orthorhombic	*Fmmm*	6
	Fd$\bar{3}$*m*	32		*I4/mcm*	462		*Pbcn*	6
	Ia$\bar{3}$	25		*P4/mmm*	360		*Pba2*	5
	Fm$\bar{3}$*m*	18		*P4/mbm*	278		*Fddd*	4
	I213	14		*I4/mmm*	76		*Immm*	3
	Im$\bar{3}$	14		P41	51		*Pmmn*	3
	Ia$\bar{3}$*d*	9		*I4/m*	16		*Ama2*	2
	Pn$\bar{3}$*m*	9		*p*-421c	6		*Cmm2*	2
	P213	3		P42/nmc	3		*Aba2*	1
	Pn$\bar{3}$	5		*I4mm*	2		*Cmca*	1
	*I*23	1		*P*$\bar{4}$*m2*	2		*Pcc2*	1
	Pa$\bar{3}$	1		P42/mcm	2		Pca21	1
Hexagonal	P63/mmc	190		P42/n	1		*Pnnm*	1
	P63 cm	16		*P*$\bar{4}$*b2*	1		Pmn21	1
	P63mc	10		I41/a	1		*Pmm2*	1
	P$\bar{6}$2*m*	6		I41/amd	1		*Ibam*	1
	P6322	4		*P4/nmm*	1	Monoclinic	P21/m	227
	P63	1	Orthorhombic	*Pnma*	1447		P21/c	217
	P63/mcm	1		*Imma*	411		*C2/m*	296
	P6522	1		*Amm2*	274		*C2/c*	97
Trigonal	*R*$\bar{3}$*c*	1080		*Cmcm*	211		*Cm*	38
	R$\bar{3}$	937		*Cmmm*	149		C2	25
	R$\bar{3}$*m*	259		*Ima2*	146		*Pc*	20
	*R*3*c*	107		*Pmmm*	136		*Pm*	16
	*R*3*m*	74		Cmc21	54		*Cc*	9
	P$\bar{3}$*m*1	28		C2221	32		P21	9
	R3	19		*Pbcm*	28		*P2/c*	2
	P$\bar{3}$1*c*	13		P212121	21	Triclinic	*p*-1	201
	*P*321	13		Pna21	19		P1	73
	P$\bar{3}$1*m*	6		*Pnna*	13			
	*P*3*m*1	4		*Pcca*	12			
	p-3	2		Pmc21	12			
	P3121	1		*Fdd2*	11			
	P31	1		*Pbam*	11			
	*P*31*c*	1		*Pmma*	9			
	*R*32	1		*Pccn*	7			
	P32	1		*Pbca*	7			

[a]We split the 11117 ABX3 with a train/test ratio of 80/20, and then, we built the XGBoost classification utilizing both model 1 and model 2.

13947 ABX_3 compounds. The remaining 1252 new formulas, however, are not present in the OQMD database and are the subject of our interest. Figure 2 provides the distribution of these 13947 ABX_3 categorized by their anionic parts: oxides, fluorines, bromines, chlorines, and iodines.

The'oxides' category has the highest number of compounds, with a count of 11567 (83%), because they have multiple oxidation states, while'halides' have the fewest compounds, with only 2380 (17%) of the data set.

2.5. Features Generation. To determine the properties of ABX_3 compounds, the first model employs **80** initial elemental

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

Table 3. Evaluation of Space Group Models in ABX_3 Compounds

		Model 1				Model 2			
Crystal System	Space Group	Train	Train + 10 cv	Test	Validation	Train	Train +10 cv	Test	Validation
Cubic	*Pm*-3m	96.53	95.67	93.46	93.29	96.99	95.21	93.98	93.71
Hexagonal	P63/mmc	97.51	94.69	95.29	95.60	100	93.31	92.41	94.34
Trigonal	$R\bar{3}c$	100	88.07	89.79	86.16	99.87	87.54	87.96	86.58
	$R\bar{3}$	99.80	88.07	90.05	84.70	100	85.96	84.55	83.86
	$R\bar{3}m$	100	96.72	97.12	95.39	99.74	94.42	93.19	93.50
Tetragonal	*P*4*mm*	99.02	87.61	84.82	87.00	100	83.35	86.39	87.00
	*I*4/*mcm*	100	91.74	90.58	93.50	100	87.81	89.01	89.94
	*P*4/*mmm*	99.93	93.18	91.10	89.94	99.34	90.75	90.58	89.73
	*P*4/*mbm*	96.53	93.00	91.62	89.31	100	91.80	91.88	91.19
Orthorhombic	*Pnma*	100	86.36	82.46	83.65	100	85.44	88.22	82.39
	Imma	100	95.54	96.60	95.39	100	90.82	90.84	91.00
	*Amm*2	100	96.52	95.81	96.65	99.93	91.28	89.53	90.57
	Cmcm	92.66	90.95	88.74	89.10	93.51	90.95	89.27	88.47
	Cmmm	100	97.37	93.72	94.34	100	95.93	95.03	95.39
	*Ima*2	99.87	99.14	97.12	98.11	100	96.85	96.86	96.02
	Pmmm	99.93	95.87	95.55	95.18	100	95.67	95.03	94.97
Monoclinic	P21/m	98.17	91.08	91.62	89.31	100	90.56	90.31	86.58
	P21/c	89.52	89.12	87.70	85.95	100	90.29	86.39	86.16
	*C*2/*m*	96.46	91.74	93.46	91.40	100	91.47	92.15	89.73
Triclinic	*p*-1	95.15	90.82	90.05	87.21	100	90.63	90.58	86.79

features for the A, B, and X atoms. Of these, **72** are sourced from the Python Materials Genomics library (Pymatgen).[45] These features include the valence, period, group, atomic number, molar volume, atomic mass, number of s, p, d, and f electrons, ionic radius, atomic radius, van der Waals radius, covalent radius, melting point, boiling point, electron affinity, electron negativity, thermal conductivity, electrical resistivity, T curie, first ionization energy, second ionization energy, and enthalpy of fusion. For the polarizabilities of these **3** sites, we selected them from.[46] The remaining **5** features are the types of compounds (oxides, fluorines, bromines, chlorines, and iodines).

To alleviate the computational difficulty,[47−50] we introduce a second model that utilizes the mean of the elemental features from A, B, and X sites, resulting in **25** novel features. The remaining **5** features are the types of compounds (oxides, fluorines, bromines, chlorines, and iodines). The total of these variables is **30**. We collected the ABX_3 properties from the OQMD database: volume (V), space group, band gap energy (E_g), and formation energy (E_f).

3. RESULTS AND DISCUSSION

3.1. Space Group Prediction. *3.1.1. Data Preprocessing (Removing Highly Correlated Features).* To predict the preferred space group for 1252 new ABX_3 halides and oxides solely from their chemical formulas, we utilized the 13947 known ABX_3 compounds with their respective space groups. We generated a correlation matrix using Python codes for both the elemental features of the first model and the mean of the elemental features of the second model to remove highly correlated features when the coefficient between them exceeds 0.95. Following removal, we obtained 60 features in the first model and 27 in the second model.

3.1.2. Data Splitting. We divided the 13947 ABX_3 compounds into a train-test set, comprising 11117 materials (80%), and a validation set, comprising 2830 materials (20%). Table 2 represents the distribution of the 11117 ABX_3 train-test, where there are 7 crystal systems with space group sets,

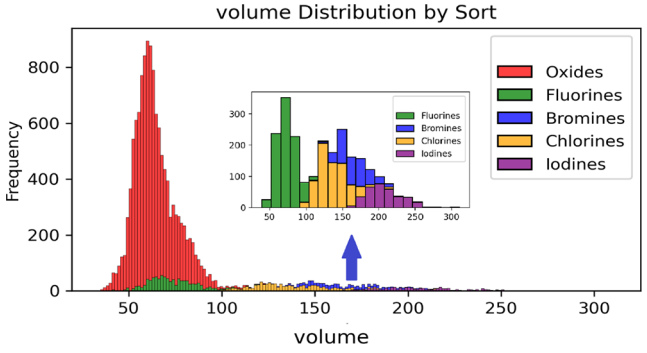


Figure 3. Volume distribution of 13947 ABX_3 in oxides, fluorines, bromines, chlorines, and iodines.

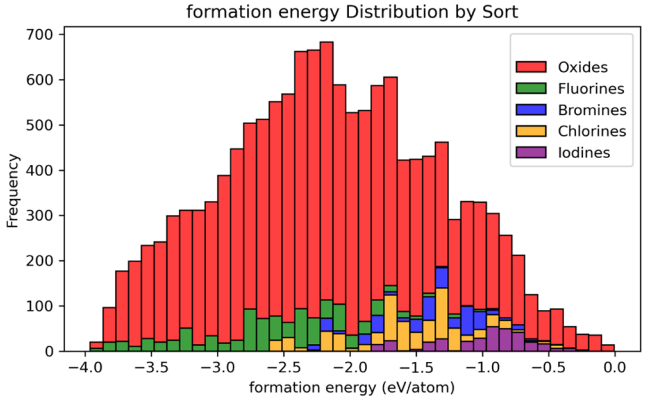


Figure 4. Formation energy distribution of 13947 ABX_3 in oxides, fluorines, bromines, chlorines, and iodines.

and the counts beside each space group indicate the number of its occurrences in the data set. For example, the space group "*Pm*-3m" has 1805 occurrences, which is the most common structural symmetry; "*Pnma*" has 1447; and so on.

3.1.3. Model Evaluation. For each space group, we need to build an XGBoost classification model that predicts "yes" or

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

Table 4. Optimal Hyper-Parameters for XGBoost Regression Models

	Model 1		Model 2	
The hyper-parameters	Volume (Å^3)	Formation energy (eV/atom)	Volume (Å^3)	Formation energy (eV/atom)
colsample_bytree	0.7	0.7	0.7	0.5
learning_rate	0.1	0.1	0.1	0.1
max_depth	5	5	7	7
min_child_weight	3	3	3	5
n_estimators	1000	1000	1000	1000
objective	reg:squarederror[a]	reg:squarederror	reg:squarederror	reg:squarederror
Subsample	0.7	0.7	0.7	0.7

[a]reg:squarederror: regression tasks where the model predicts a continuous value.

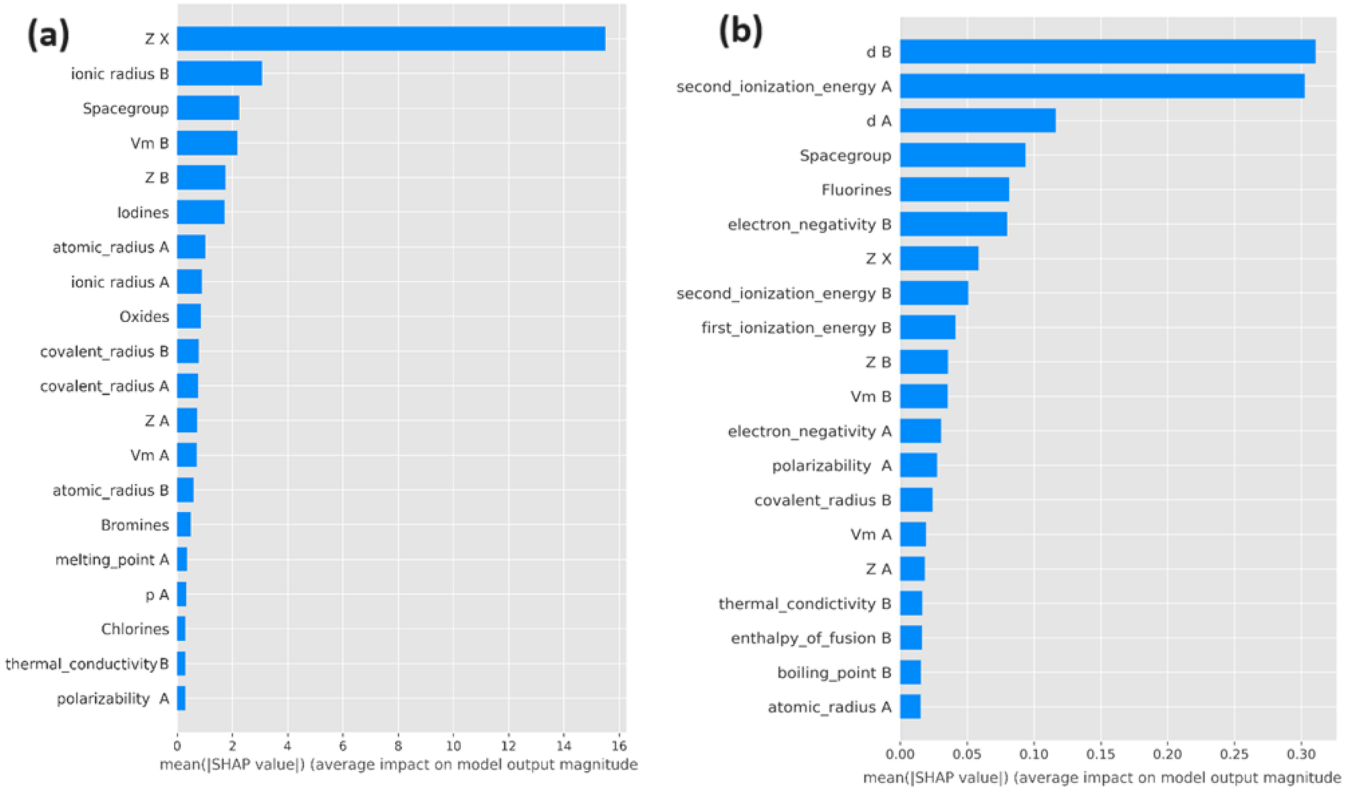


Figure 5. Feature importance plot. (a) The SHAP plot of the volume (V); (b) the SHAP plot of the formation energy (E_f).

Table 5. Relationship between d B and E_f Clusters

E_f (eV/atom)		−4 to −3.3	−3.3 to −2.7	−2.7 to −2	−2 to −1.5	−1.5 to −1.2	−1.2 to 0
d B	fluorines	0−2	0−5	0−10	5−10	6−10	10
	chlorines	/	/	0	0−10	5−10	5−10
	bromines	/	/	0	0−2	3−10	
	iodines	/	/	/	0	0	
	oxides	0−2	0−5	0−10	5−10	5−10	

″no″ values. However, some of these groups contain a small number of compounds (for example, *Ama*2 appears only once), which hinders the construction of a strong model. Therefore, we chose only the first 20 groups that had more than 100 occurrences in the data. These are the most common space groups found in perovskite structures and have technological applications in various fields. Table 3 summarizes the results that yield classification accuracies ranging from 82.39% to 97.12% across these materials in train, train + 10 cross-validation (cv), test, and validation data. In the test set, models 1 and 2 achieve the best accuracies of 97.12% and 96.86%, respectively, corresponding to the space group ″*Ima*2″, and their lowest accuracies of 82.46% and 88.22%, respectively, linked with the space group ″*Pnma*″, In the validation set, models 1 and 2 achieve their highest accuracies of 98.11% and 96.02%, respectively, associated with the space group ″*Ima*2″, and their lowest accuracies of 83.65% and 82.39%, respectively, tied to the space group ″*Pnma*″.

3.2. Volume (V) and Formation Energy (E_f) Prediction.

3.2.1. Data Distribution. Figure 3 illustrates the volume distribution of oxides and four halide types, indicating a range of behaviors. Chlorines and oxides demonstrate a narrow and high peak, suggesting a precise volume measurement with less variability (less than 140). Iodines exhibit a broad distribution, signifying a high degree of variability in volume, which may imply diverse physicochemical properties. Also, the biggest volumes are iodines (more than 220 Å^3). The overlap between bromines and chlorines indicates similarities in volume within specific ranges. Figure 4 illustrates the formation energy distribution of the previous five types. Oxides dominate the distribution, with a prominent symmetrical peak centered

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

Table 6. Evaluation of Formation Energy (E_f) and Volume (V) Models in the ABX_3 Compounds

		Model 1			Model 2		
		Accuracy (%)	MAE (eV/atom)	RMSE (eV/atom)	Accuracy (%)	MAE (eV/atom)	RMSE (eV/atom)
V	train	99.59	1.343	2.123	99.85	0.843	1.272
	train + 10 cv	98.00	2.400	4.255	98.00	2.890	5.000
	test	98.41	2.395	4.416	98.06	2.748	4.878
	validation	97.85	2.588	4.955	97.02	3.427	5.829
E_f	train	99.42	0.039	0.062	99.70	0.028	0.044
	train + 10 cv	98.00	0.070	0.140	97.00	0.080	0.150
	test	97.36	0.075	0.132	96.93	0.083	0.143
	validation	96.67	0.086	0.140	94.60	0.125	0.179

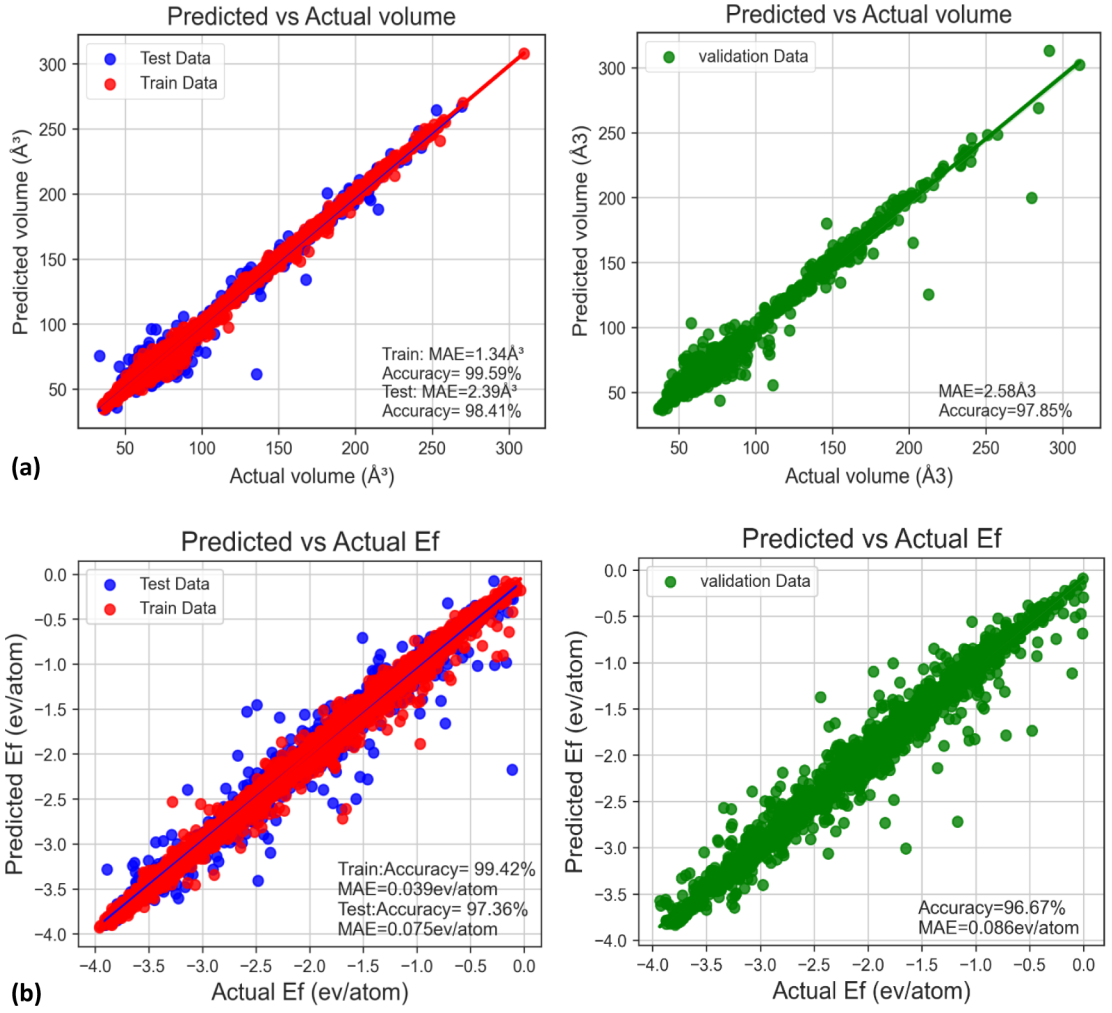


Figure 6. Prediction results of volume and formation energy for ABX_3 materials in train, test, and validation data. (a) Parity plot of the volume (V); (b) parity plot of the formation energy (E_f). Note: "actual" refers to reported values in the OQMD.

Table 7. Evaluation of Metal or Nonmetal State Models in the ABX_3 Compounds[a]

	Model 1				Model 2			
	train	train + 10 cv	test	validation	train	train + 10 cv	test	validation
Metal or nonmetal model accuracy (%)	98.91	90.47	91.17	88.63	99.92	89.75	90.58	86.26

[a]If the compound exhibits metallic properties, indicating a band gap energy of zero, otherwise, XGBoost regression models can predict the gap energy for nonmetallic compounds (Table 7).

around −2.0. The other halides exhibit significantly lower frequencies, with bromines and iodines showing a relatively narrow distribution, whereas fluorines and chlorides display broader spreads, indicative of higher variability. It can also be noted that the most stable compounds for the phase transition are oxides and fluorines (less than −2.7 eV/atom).

3.2.2. Data Splitting and Removing Highly Correlated Features. To predict the volume (V) and formation energy

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

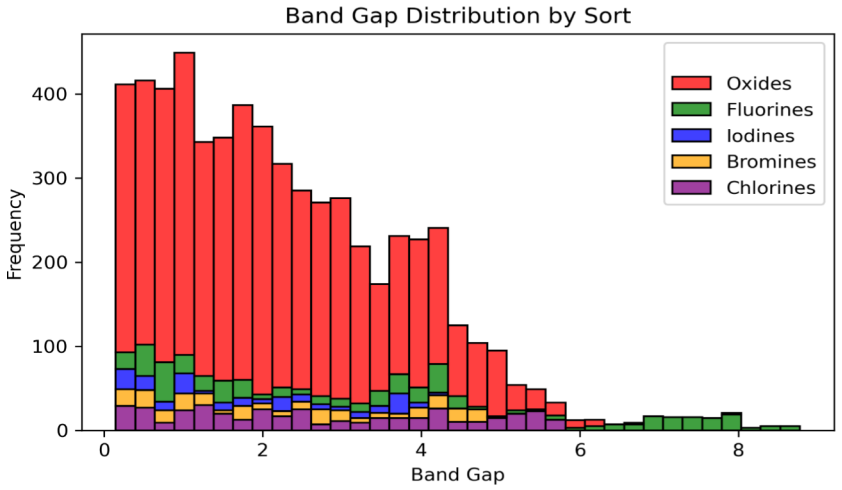


Figure 7. Band gap energy distribution of 13947 ABX_3 in oxides, fluorines, bromines, chlorines, and iodines.

Table 8. Optimal Hyper-Parameters for XGBoost Regression Models

	Model 1	Model 2
The hyper-parameters	E_g	E_g
colsample_bytree	0.5	0.5
learning_rate	0.01	0.01
max_depth	10	10
min_child_weight	5	5
n_estimators	1000	1000
objective	reg:squarederror	reg:squarederror
Subsample	0.7	0.7

(E_f) of 1252 new ABX_3 halides and oxides just from the chemical formula and proposed space group, we selected the 13947 known ABX_3 compounds with their space groups. We generate a correlation matrix using Python codes for both the elemental features of the first model and the mean of the elemental features of the second model to remove highly correlated features when the coefficient between them exceeds 0.95. Following removal, we obtained 60 features in the first model and 27 in the second model.

We divided the 13947 ABX_3 compounds into a train-test set, comprising 11117 materials (80%), and a validation set, comprising 2830 materials (20%). Subsequently, we divided these 11158 train-test ABX_3 compounds with a train-test ratio of 80/20, upon which we constructed our XGBoost regression models.

3.2.3. Hyper-Parameters Optimization. Before the prediction processes begin, we need to optimize the hyper-parameters to enhance the performance of our models and attain heightened accuracy. Table 4 represents the optimal hyper-parameters for XGBoost regression utilizing both model 1 and model 2.

3.2.4. Features Importance Plot. By examining the SHAP plots presented in Figure 5, we can discern the influence of these features on our prediction. In (a), the primary factors affecting the prediction of V are the atomic number of the X atom, the ionic radius of the B atom, and the space group (Figure 5a). Also, the most important family is iodines, which confirms the distribution in Figure 3, because the largest volumes in our data are iodines. In contrast, the top features in predicting E_f are the number of d electrons in the B atom (d B), the second ionization energy of the A atom, and the number of d electrons in the A atom (d A). The most important family is fluorines, because most of them are stable for phase transition, which confirms the distribution in Figure 4. We can explain the E_f clusters in Figure 5b with the SHAP plot of the formation energy (E_f) that confirms the importance of d B as in Table 5.

3.2.5. Model Evaluation. Table 6 represents the model performance for train, train + 10 cross-validation (cv), test, and validation data, which summarizes the results that yield regression accuracies greater than 96% across these materials. For the volume, both models 1 and 2 achieve a test set accuracy of 98% and a validation set accuracy of 97%. In the formation energy, model 1 achieves an accuracy of 97.36% and model 2 achieves 96.93% in the test set, while in the validation set, model 1 achieves 96.67% accuracy and model 2 achieves 94.60% accuracy. Figure 6 presents the parity plot of the volume and formation energy for ABX_3 materials in model 1. The training data set is indicated by red circles, the test data set is indicated by blue circles, and the validation data set is also illustrated by green circles.

3.3. Band Gap Energy (E_g) Prediction. The prevalence of zero values in the band gap target in the regression models greatly hinders prediction accuracy. This challenge arises because these models struggle with the imbalance between zeros and nonzeros, leading to inadequate performance.

To construct a precise predictive model, we must remove the zero values of the band gap energy, which constitute 57% of the data. Therefore, we propose the development of a new classification model termed "metal or nonmetal", aimed at discerning between metallic ($E_g = 0$) and nonmetallic states ($Eg \neq 0$).

3.3.1. Metal or Nonmetal State Prediction. Among the 13947 ABX_3, 27 energy gaps are not mentioned in the OQMD, so we divide the 13927 ABX_3 compounds into an 80% train-test set and a 20% validation set. Subsequently, we divided these train-test ABX_3 compounds with a train-test ratio of 80/20, upon which we constructed our XGBoost classification models Table 7.

3.3.2. Band Gap Energy (E_g) Prediction for Nonmetallic ABX_3 Compounds. 3.3.2.1. Data Distribution. Figure 7 illustrates the gap energy distribution of five types. Oxides with lower band gap energies are more common, so they are relevant in applications such as metals or semiconductors. The other types, including fluorines, iodines, bromines, and

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

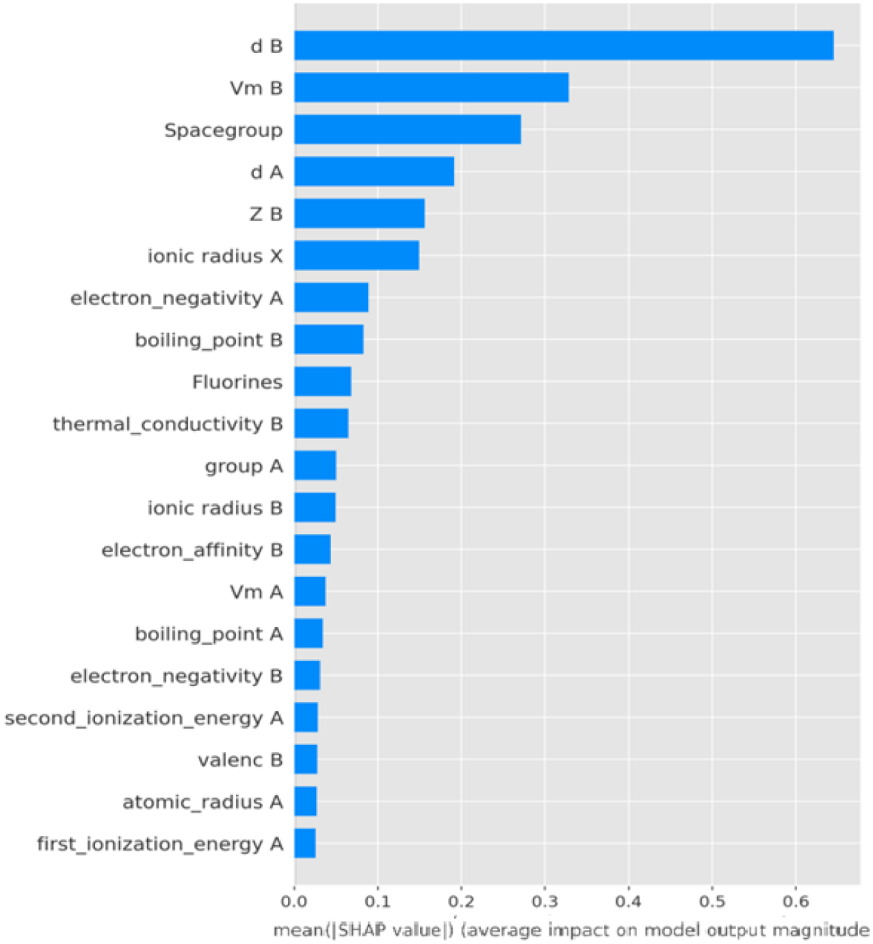


Figure 8. SHAP plot of the band gap energy (E_g).

Table 9. Relationship between d B and E_f Clusters

E_g (eV)		6−9	3−6	1.5−3	0−1.5	0
d B	fluorines	0	0, 10	0, 5, 10	0, 5−10	0, 5−10
	chlorines	/				
	bromines	/				
	iodines	/	0			
	oxides	0	0−2, 10	0−5, 10	0−10	0−10

chlorines, display broader distributions, suggesting a greater variety of band gap energies, which could reflect diversity in the electronic or structural properties of the materials studied. Fluorines have the highest band gaps (ranging from 6 to 8 eV). The distribution occurs within three energy ranges, suggesting the existence of separate energy groups or clusters within these materials.

3.3.2.2. Removing Highly Correlated Features. To predict the band gap energy (E_g) of 1252 new ABX_3 halides and oxides just from the chemical formula and proposed space group, we selected the 7164 nonmetallic ABX_3 compounds with their space groups. We generate a correlation matrix using Python codes for both the elemental features of the first model and the mean of the elemental features of the second model to remove highly correlated features when the coefficient between them exceeds 0.95. Following removal, we obtained 60 features in the first model and 27 in the second model.

We divided the 7164 ABX_3 compounds into a train-test set, comprising 5731 materials (80%), and a validation set, comprising 1433 materials (20%). Subsequently, we divided these 5731 train-test ABX_3 compounds with a train/test ratio of 80/20, upon which we constructed our XGBoost regression models.

3.3.2.3. Hyper-Parameter Optimization. Before the prediction processes begin, we need to optimize the hyperparameters to enhance the performance of our models and attain heightened accuracy. Table 8 represents the optimal

Table 10. Evaluation of Band Gap Energy Models in the ABX_3 Compounds

		Model 1			Model 2		
		Accuracy (%)	MAE (eV/atom)	RMSE (eV/atom)	Accuracy (%)	MAE (eV/atom)	RMSE (eV/atom)
E_g	train	97.15	0.18	0.26	97.32	0.17	0.25
	train + 10 cv	85.00	0.41	0.60	83.01	0.44	0.63
	test	85.17	0.40	0.60	83.00	0.45	0.64
	validation	87.00	0.39	0.57	85.46	0.42	0.60

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

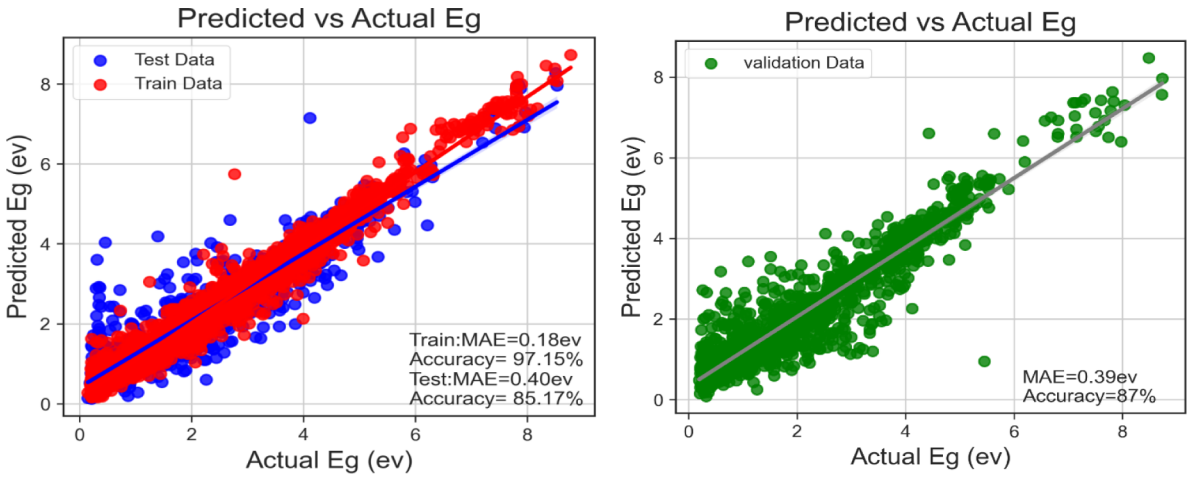


Figure 9. Prediction results of band gap energy (E_g) for ABX_3 materials in training, testing, and validation data. Note: "actual" refers to reported values in OQMD.

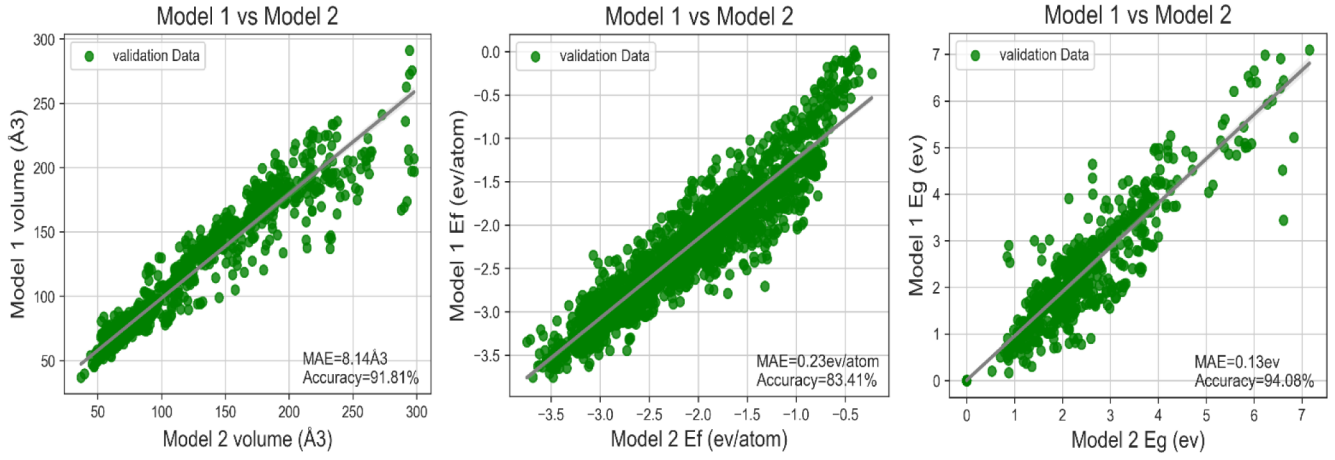


Figure 10. Comparison of volume, formation energy, and band gap energy results for new ABX_3 materials: models 1 vs 2.

hyper-parameters for XGBoost regression utilizing both model 1 and model 2.

3.3.2.4. Features Importance Plot. By examining the SHAP plots presented in Figure 8, we can discern the influence of these features on our prediction. The primary factors affecting the prediction of E_g are the number of d electrons in the B atom (d B), the molar volume of the B atom, and the space group. The most important family is fluorines, because they have the highest band gaps (ranging from 6 to 8 eV), which confirms the distribution in Figure 7. We can explain the E_g clusters in Figure 8 with the SHAP plot of the band gap energy (E_g) that confirms the importance of d B as in Table 9.

3.3.2.5. Model Evaluation. Table 10 represents the model performance for train, train + 10 cross-validation, test, and validation data, which summarizes the results that yield regression accuracies greater than 86% across these materials. For the band gap energy, model 1 achieves an accuracy of 85.17% and model 2 achieves 83.00% in the test set, while in the validation set, model 1 achieves 87.00% accuracy and model 2 achieves 85.46% accuracy. Figure 9 presents the parity plot of the band gap energy for ABX_3 materials in model 1. The training data set is indicated by red circles, the test data set is indicated by blue circles, and the validation data set is also illustrated by green circles.

3.4. Generalization of Models to Encompass 1252 New Formulas. We use the previous XGBoost models with 1252 new ABX_3 to find the possible space groups for each formula (using the 20 models we built previously), and then, we predict the volume and the formation energy for each space group (see Supporting Information). Furthermore, we test whether a compound is a metal; we identify instances where the presence of a metal indicates a zero-band gap energy, while, otherwise, we predict the band gap energy in addition to the energy gaps not mentioned in the OQMD. Models 1 and 2 detect numerous ABX_3 compounds, and when we consider the space group intersection between them, we obtain a set of 1836 materials.

For example, we consider the compound $AcAmO_3$. Both models 1 and 2 suggest two space groups with the volume, formation, and band gap energies at slightly different values.

Model 1: $AcAmO_3$

$$Pm\bar{3}m\text{: } V = 86.68\ \text{A}^3,\ Ef = -3.04\ \text{eV/atom},\ Eg = 0\ \text{eV}$$

$$R\bar{3}\text{: } V = 89.75\ \text{A}^3,\ Ef = -3.39\ \text{eV/atom},\ Eg = 2.72\ \text{eV}$$

Model 2: $AcAmO_3$

$$Pm\bar{3}m\text{: } V = 86.58\ \text{A}^3,\ Ef = -3.05\ \text{eV/atom},\ Eg = 0\ \text{eV}$$

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

$R\bar{3}$: $V = 86.57$ Å^3, $Ef = -3.38$ eV/atom, $Eg = 2.44$ eV

To clarify the compatibility between the two models, we compare them in Figure 10. This figure presents the parity plot for ABX_3 materials, showing that model 1 and model 2 achieve accuracies of 91.81%, 83.41%, and 94.08% for volume, formation energy, and band gap energy, respectively. The mean absolute errors (MAEs) for volume, formation energy, and band gap energy are 8.14 Å^3, 0.23 eV/atom, and 0.13 eV, respectively. These values indicate reasonable convergence among the models.

4. CONCLUSION

This work presents a comprehensive application of machine learning (ML) techniques for the discovery and design of ABX_3 perovskite materials. By leveraging the XGBoost algorithm, we developed predictive models capable of identifying new potential ABX_3 formulas and estimating their fundamental properties solely on the basis of their chemical compositions. First, we employed an XGBoost classification model to predict the space group symmetry of known oxide and halide ABX_3 compounds from the OQMD database. The model achieved remarkable accuracies ranging from 82.39% to 99.14%, demonstrating its ability to capture the intricate relationships between chemical formulas and crystal structures. Subsequently, we utilized XGBoost regression models to predict three crucial material properties: volume (V), formation energy (E_f), and band gap energy (E_g). These properties govern the stability, structural characteristics, and electronic behavior of perovskite materials, making them essential for assessing their suitability for various applications. The volume prediction model, trained on elemental features, exhibited an impressive accuracy of 98.41% with a mean absolute error (MAE) of 2.395 Å^3 and a root-mean-squared error (RMSE) of 4.416 Å^3. Similarly, the formation energy model achieved an accuracy of 97.36%, with an MAE of 0.075 and an RMSE of 0.132, indicating its proficiency in estimating the thermodynamic stability of these materials. Furthermore, we developed a classification model to distinguish between metallic and nonmetallic compounds, as the electronic properties of these two classes differ fundamentally. For nonmetallic compounds, an XGBoost regression model was employed to predict the band gap energy, a crucial parameter governing the optical and electronic behavior of semiconductors and insulators. This model achieved an accuracy of 87.00%, with an MAE of 0.391 and an RMSE of 0.574, demonstrating its reliability in estimating this critical property. By combining the predictions from these models, we identified a set of 1836 potential new ABX_3 formulas with estimated properties, paving the way for further exploration and experimental validation of these promising materials. The findings presented in this paper highlight the power of machine learning techniques in accelerating the discovery and design of novel perovskite materials. By leveraging the ability of ML models to capture complex patterns and relationships within material data, we can efficiently navigate the vast chemical space and identify promising candidates for targeted synthesis and characterization.

■ ASSOCIATED CONTENT

Data Availability Statement

Data sharing is not applicable to this article.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c06139.

60 features model 1 and 27 features model 2 (PDF)
corr_matrix model 1(XLSX)
corr_matrix model 2 (XLSX)
Detailed list of the 1836 potential new ABX_3 formulas (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Souraya Goumri-Said − *College of Science and General studies, Department of Physics, Alfaisal University, Riyadh 11533, Saudi Arabia;* orcid.org/0000-0002-9333-7862; Email: sosaid@alfaisal.edu

Authors

Soundous Touati − *Laboratoire de Physique des Matériaux, Université Amar Telidji de Laghouat, Laghouat 03000, Algeria; Laboratory of Applied Sciences and Didactic, Higher Normal School of Laghouat, Laghouat 03000, Algeria*

Ali Benghia − *Laboratoire de Physique des Matériaux, Université Amar Telidji de Laghouat, Laghouat 03000, Algeria*

Zoulikha Hebboul − *Laboratoire Physico-Chimie des Matériaux (LPCM), Université Amar Telidji de Laghouat, Laghouat 03000, Algeria*

Ibn Khaldoun Lefkaier − *Laboratoire de Physique des Matériaux, Université Amar Telidji de Laghouat, Laghouat 03000, Algeria*

Mohammed Benali Kanoun − *Department of Mathematics and Sciences, College of Humanities and Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia;* orcid.org/0000-0002-2334-7889

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c06139

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

S. T. acknowledges the computing resources provided on the HPC CUMULUS operated by the Plateforme Technologique Calcul Intensif (PTCI) of Amar Telidji University of Laghouat. Z.H. from Université Amar Telidji was supported by the Algerian Ministry of Higher Education and Scientific Research through Projects de Recherche Formation-Universitaire PRFU under grant B00L01UN030120220002. S. Goumri-Said thanks the office of research at Alfaisal University in Saudi Arabia for funding this research work through internal project number 24407. M. B. Kanoun would like to thank Prince Sultan University for their support.

■ REFERENCES

(1) Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W.-J. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat. Commun.* **2020**, *11*, 3513.

(2) Pilania, G.; Balachandran, P. V.; Kim, C.; Lookman, T. Finding new perovskite halides via machine learning. *Front. Mater* **2016**, *3*, 19.

(3) Kumar, S.; et al. Accelerated discovery of perovskite materials guided by machine learning techniques. *Mater. Lett.* **2023**, *353*, 135311.

(4) Gómez − Peralta, J. I.; Bokhimi, X. Discovering new perovskites with artificial intelligence. *J. Solid State Chem.* **2020**, *285*, 121253.
(5) Jacobs, R.; Liu, J.; Abernathy, H.; Morgan, D. Machine Learning Design of Perovskite Catalytic Properties. *Adv. Energy Mater.* **2024**, *14*, 2303684.
(6) Li, Z.; Achenie, L. E. K.; Xin, H. An Adaptive Machine Learning Strategy for Accelerating Discovery of Perovskite Electrocatalysts. *ACS Catal.* **2020**, *10*, 4377−4384.
(7) Shambhawi, S.; Csányi, G.; Lapkin, A. A. Active Learning Training Strategy for Predicting O Adsorption Free Energy on Perovskite Catalysts using Inexpensive Catalyst Features. *Chem. Methods* **2021**, *1*, 444−450.
(8) Zhang, L.; Li, N.; Liu, D.; Tao, G.; Xu, W.; Li, M.; Chu, Y.; Cao, C.; Lu, F.; Hao, C.; et al. Deep Learning for Additive Screening in Perovskite Light-Emitting Diodes. *Angew. Chem. Int. Ed.* **2022**, *61*, No. e202209337.
(9) Ji, K.; et al. Self-supervised deep learning for tracking degradation of perovskite light-emitting diodes with multispectral imaging. *Nat. Mach. Intell.* **2023**, *5*, 1225−1235.
(10) Fakharuddin, A.; et al. Perovskite light-emitting diodes. *Nat. Electron.* **2022**, *5*, 203−216.
(11) Varignon, J. Origin of superconductivity in hole doped SrBiO3 bismuth oxide perovskite from parameter-free first-principles simulations. *Npj Comput. Mater.* **2023**, *9*, 30.
(12) Zhang, M.; et al. Superconductivity in Perovskite $Ba_{1-x}Ln_x(Bi_{0.20}Pb_{0.80})O_{3-\delta}$ (Ln = La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu). *Inorg. Chem.* **2018**, *57*, 1269−1276.
(13) Kim, M.; et al. Superconductivity in (Ba,K)SbO3. *Nat. Mater.* **2022**, *21*, 627−633.
(14) Kuroiwa, Y.; Kim, S.; Fujii, I.; Ueno, S.; Nakahira, Y.; Moriyoshi, C.; Sato, Y.; Wada, S. Piezoelectricity in perovskite-type pseudo-cubic ferroelectrics by partial ordering of off-centered cations. *Commun. Mater.* **2020**, *1*, 71.
(15) Wu, H.-S.; Murti, B. T.; Singh, J.; Yang, P.-K.; Tsai, M.-L. Prospects of Metal-Free Perovskites for Piezoelectric Applications. *Adv. Sci.* **2022**, *9*, 2104703.
(16) Shukla, A.; Kumar, A.; Pathak, K. Ferromagnetism in LaMnO3-LaFeO3-LaCoO3 mixed spin perovskite oxide solid solution. *Ceram. Int.* **2023**, *49*, 12680−12686.
(17) Bansal, N. K.; Mishra, S.; Dixit, H.; Porwal, S.; Singh, P.; Singh, T. Machine Learning in Perovskite Solar Cells: Recent Developments and Future Perspectives. *Energy Technol.* **2023**, *11*, 2300735.
(18) Parikh, N.; et al. Is machine learning redefining the perovskite solar cells? *J. Energy Chem.* **2022**, *66*, 74−90.
(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*, 1901891.
(20) Watson, H.; et al. Performance analysis of perovskite solar cells in 2013−2018 using machine-learning tools. *Prostaglandins, Leukotrienes Essent. Fatty Acids* **2016**, *115*, 60−66.
(21) Li, Y.; Dong, R.; Yang, W.; Hu, J. Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors. *Comput. Mater. Sci.* **2021**, *198*, 110686.
(22) Nomura, K.; et al. Machine learning based prediction of space group for Ba(Ce0.8-xZrx)Y0.2O3 perovskite-type protonic conductors. *Ceram. Int.* **2023**, *49*, 5058−5065.
(23) Majid, A.; Khan, A.; Javed, G.; Mirza, A. M. Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression. *Comput. Mater. Sci.* **2010**, *50*, 363−372.
(24) Li, Y.; Yang, W.; Dong, R.; Hu, J. Mlatticeabc: Generic Lattice Constant Prediction of Crystal Materials Using Machine Learning. *ACS Omega* **2021**, *6*, 11585−11594.
(25) Emery, A. A.; Wolverton, C. High-Throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO 3 perovskites. *Sci. Data* **2017**, *4*, 170153.
(26) Talapatra, A.; Uberuaga, B. P.; Stanek, C. R.; Pilania, G. A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. *Chem. Mater.* **2021**, *33*, 845−858.
(27) Idrissi, S.; Labrim, H.; Bahmad, L.; Benyoussef, A. Study of the solar perovskite CsMBr3 (M = Pb or Ge) photovoltaic materials: Band-gap engineering. *Solid State Sci.* **2021**, *118*, 106679.
(28) Idrissi, S.; Mounkachi, O.; Bahmad, L.; Benyoussef, A. Study of the electronic and opto-electronic properties of the perovskite KPbBr3 by DFT and TDDFT methods. *Comput. Condens. Matter* **2022**, *33*, No. e00617.
(29) Sebastian, M.; Peters, J. A.; Stoumpos, C. C.; Im, J.; Kostina, S. S.; Liu, Z.; Kanatzidis, M. G.; Freeman, A. J.; Wessels, B. W. Excitonic emissions and above-band-gap luminescence in the single-crystal perovskite semiconductors CsPbB r3 and CsPbC l3. *Phys. Rev. B* **2015**, *92*, 235210.
(30) Im, J.; Lee, S.; Ko, T.-W.; Kim, H. W.; Hyon, Y.; Chang, H. Identifying Pb-free perovskites for solar cells by machine learning. *Npj Comput. Mater.* **2019**, *5*, 37.
(31) Li, C.; et al. A progressive learning method for predicting the band gap of ABO3 perovskites using an instrumental variable. *J. Mater. Chem. C* **2020**, *8*, 3127−3136.
(32) Zhang, J.; Li, Y.; Zhou, X. Machine-Learning Prediction of the Computed Band Gaps of Double Perovskite Materials. *arXiv*, **2023**. .
(33) Gao, Z.; et al. Screening for lead-free inorganic double perovskites with suitable band gaps and high stability using combined machine learning and DFT calculation. *Appl. Surf. Sci.* **2021**, *568*, 150916.
(34) Guo, Z.; Lin, B. Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells. *Sol. Energy* **2021**, *228*, 689−699.
(35) Rath, S.; Sudha Priyanga, G.; Nagappan, N.; Thomas, T. Discovery of direct band gap perovskites for light harvesting by using machine learning. *Comput. Mater. Sci.* **2022**, *210*, 111476.
(36) Hu, W.; Zhang, L. High-Throughput Calculation and Machine Learning of Two-Dimensional Halide Perovskite Materials: Formation Energy and Band Gap. *Mater. Today Commun.* **2023**, *35*, 105841.
(37) Chen, T.; Guestrin, C.XGBoost: A scalable tree boosting system*Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*ACM2016785−794
(38) Zhang, L.; Zhan, C. Machine Learning in Rock Facies Classification: An Application of XGBoost. In *International Geophysical Conference*, Society of Exploration Geophysicists and Chinese Petroleum Society: 2017, 1371−1374. .
(39) Zhang, K.; Zhang, Y.; Wang, M.*A Unified Approach to Interpreting Model Predictions Scott*Nips2012*16*426−430
(40) Lundberg, S. M.; Erion, G. G.; Lee, S.-I. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*, **2018**.
(41) Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584.
(42) Shannon, R. D. *Database of Ionic Radii*. http://abulafia.mt.ic.ac.uk/shannon/radius.php.
(43) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *Npj Comput. Mater.* **2015**, *1*, 15010.
(44) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501−1509.
(45) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(46) Schwerdtfeger, P.; Nagle, J. K. 2018 Table of Static Dipole Polarizabilities of the Neutral Elements in the Periodic Table. *Mol. Phys.* **2019**, *117*, 1200−1225.

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531

(47) Djeradi, S.; Dahame, T.; Fadla, M. A.; Bentria, B.; Kanoun, M. B.; Goumri-Said, S.High-Throughput Ensemble-Learning-Driven Band Gap Prediction of Double Perovskites Solar Cells Absorber. *MAKE* **2024**, *6*, 435.
(48) Alhashmi, A.; Kanoun, M. B.; Goumri-Said, S. Machine Learning for Halide Perovskite Materials ABX3 (B = Pb, X = I, Br, Cl) Assessment of Structural Properties and Band Gap Engineering for Solar Energ. *Materials* **2023**, *16*, 2657.
(49) Ben Kamri, A. L.; Fadla, M. A.; Lefkaier, I. k.; Ben Messaoud, C. l.; Kanoun, M. B.; Goumri-Said, S. AI-Driven Ensemble Learning for Accurate Seebeck Coefficient Prediction in Half-Heusler Compounds Based on Chemical Formulas. *Computational Condensed Matter* **2024**, *40*, e00923.
(50) Touati, S.; Benghia, A.; Hebboul, Z. I.Kh. Lefkaier, M. B. Kanoun and S. Goumri-Said, Predictive machine learning approaches for perovskites properties using their chemical formula: Towards the discovery of stable solar cells materials. *Neural Comput. Appl.* **2024**, *36*, 16319−16329.

https://doi.org/10.1021/acsomega.4c06139
ACS Omega 2024, 9, 47519−47531