Cationic perturbation strategy to solve the information drought in material explainable machine learning

Zhengxin Chen,[1,*] Hange Wang,[1,*] Bowen Liu,[1] Hairui Zhou,[1] Yingjie Cao,[1] Yanan Wang,[1] Lin Peng,[1] Xiaolin Liu,[1] Jia Lin,[1,†] Xianfeng Chen,[3,4,‡] and Jiang Wu[2,§]

[1] *College of Mathematics and Physics, Shanghai University of Electric Power, Shanghai 200090, China*
[2] *College of Energy and Mechanical Engineering, Shanghai University of Electric Power, Shanghai 200090, China*
[3] *State Key Laboratory of Advanced Optical Communication Systems and Networks, School of Physics and Astronomy, Shanghai Jiao Tong University, Shanghai 200240, China*
[4] *Collaborative Innovation Center of Light Manipulation and Applications, Shandong Normal University, Jinan 250358, China*

(Received 13 December 2023; revised 17 January 2024; accepted 22 January 2024; published 13 February 2024)

In the field of materials research, machine learning (ML) techniques have emerged as indispensable tools. However, the opaqueness in decision making by models can compromise the trustworthiness of results, underscoring the crucial need for model interpretability. Explainable machine learning (XML) strives to augment researchers' comprehension of material properties and performance. Yet, reliance on high-quality datasets and scarcity of prior knowledge pose challenges for XML research, particularly when dealing with smaller datasets. In this study, using spinel as a representative example, we successfully addressed the data challenges in XML through a cationic perturbation strategy. We demonstrate an effective approach for handling information scarcity in small datasets, thus offering a feasible method for material research and broadening the scope of XML applications in materials science. Furthermore, our investigation successfully uncovered potential causal relationships underlying material properties and validated their consistency with physical cognition. These causal relationships can serve as experimental guides, facilitating the design and optimization of new materials. Consequently, this research holds significant scientific merit in advancing XML in the realm of materials science, while providing profound insights into material properties and fostering the development of reliable ML-based materials research.

DOI: 10.1103/PhysRevB.109.085306

I. INTRODUCTION

In order to explore new materials and reduce experimental costs and trial and error time, scientists have developed high-throughput first-principles simulation techniques [1–4]. However, due to the high computational cost [5] and the limited amount of information that human experts can handle and acquire for multivariate coupling mapping functions, there is a need for alternative approaches. Moreover, to ensure high-throughput computational efficiency, most of the data in material databases are derived from semilocal functionals and generalized gradient approximation (GGA) [2–4], which often underestimate the experimental band gap by 30%–100% [6]. Finding new techniques applicable for exploring new materials is imperative. In recent years, with the expansion of some material databases [2–4], machine learning (ML) techniques have been widely applied in the field of material design and exploration. ML can spontaneously learn hidden patterns from massive data without relying on deep domain knowledge [7–11], significantly improving the efficiency and accuracies in material research [12]. However, the complexity of ML itself and the prevalence of "black-box models" are detrimental to materials scientists' understanding of results and insight into the decision-making process. In fact, some state of the art ML models suffer from logical unreliability and exhibit poor extrapolation performance [13,14]. Therefore, materials scientists generally seek to enhance the interpretability of ML models.

Explainable machine learning (XML) [15] is primarily aimed at enhancing researchers' understanding of the relationship between new material properties and performance. By combining computer simulations, ML, and data-driven approaches, XML can extract useful information from large amounts of data, thereby accelerating the design and optimization of new materials. In XML, an important aspect is the generation of highly explainable models, enabling researchers to intuitively understand the model's prediction results and the factors it depends on. However, the application of XML still faces some challenges. (1) Most material data are scarce, limiting the predictive and expressive capabilities of the models. Hidden patterns in small data often exhibit uncertainty and specificity. (2) The uncertainty and noise inherent in the data pose challenges to the accuracy and interpretability of the models. (3) The complexity of correlations between material properties makes model fitting and interpretation extremely difficult, highlighting the importance of incorporating domain

[*] These authors contributed equally to this work.
[†] jlin@shiep.edu.cn
[‡] xfchen@sjtu.edu.cn
[§] wjcfd2002@163.com

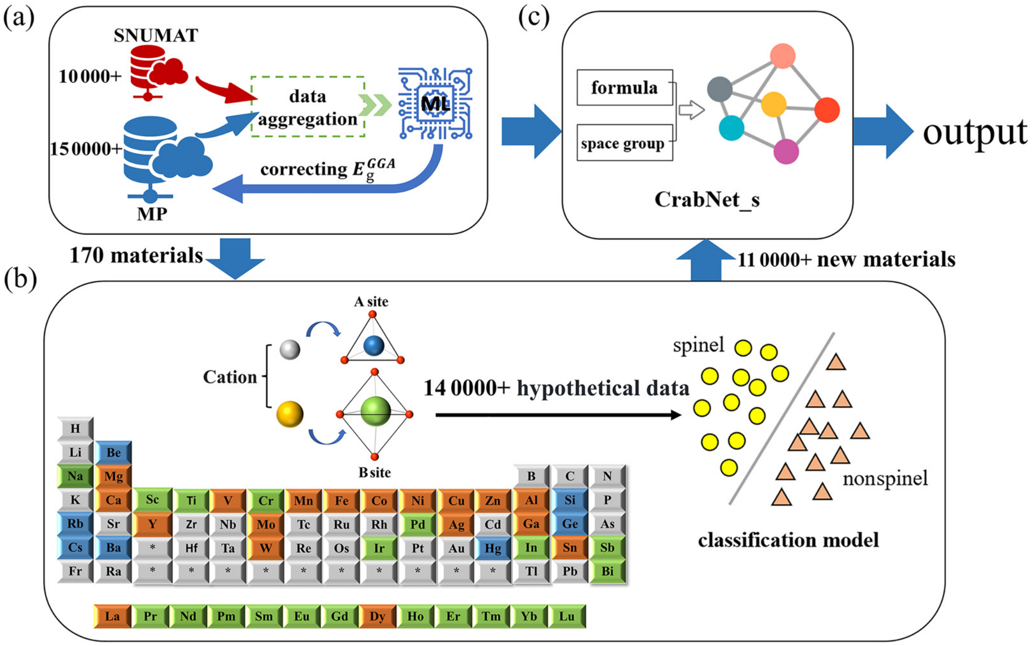

FIG. 1. Flow chart for optimizing the spinel oxide dataset. (a) E_g^{GGA} in MP is corrected by combining the SNUMAT database with MP to train ML model. (b) The 170 known spinel oxides in MP are replaced by cations at site *A* and site *B*, respectively, and are distinguished as to whether they form spinel structures or not, where blue is the optional element at site *A*, green is the optional element at site *B*, and orange is the optional element at both sites. (c) The proxy model is trained on MP_m to extract the labels of the revised dataset after cation substitution.

knowledge and designing descriptors to balance model accuracy and interpretability [16,17].

With the development of generative artificial intelligence (GAI), large-scale GAI models such as ChatGPT possess powerful capabilities for generating diverse content across multiple domains [18]. They can also be deeply integrated with AI for science, generating abundant high-quality data guided by domain knowledge, thereby greatly accelerating material property prediction and new material development processes. However, the establishment of such GAI strategies requires a foundation of domain knowledge and a certain scale of data, making it ineffective for material types with extremely scarce data. By partially replacing the original ions in materials with a large number of selectable substituting elements, the scale of generated data can be freely controlled by adjusting the substitution ratio. Importantly, when the original structure type remains unchanged under certain conditions after ion substitution, the structure remains constant while the composition becomes variable. In this case, it is only necessary to extract descriptors for the composition, significantly reducing the complexity of the model and enhancing its interpretability.

Band gap properties play a vital role in understanding the electronic and optical characteristics of materials and developing new materials with desired properties for various applications. In this study, we focused on investigating the band gap of spinel oxides as an example. However, we encountered data challenges due to the limited number of spinel oxide compounds (only 170) available in the Materials Project database (MP) [3] when searching with specific chemical formula (AB_2O_4) and space group number (227) conditions. Moreover, the majority of data in the database were derived from GGA, which constrained the study of XML in terms of data quality and quantity. To overcome these challenges, we conducted three key steps as illustrated in Fig. 1:

(i) HSE06 calculations (referred to as HSE) provide band gaps that are closer to experimental values. Therefore, we trained a ML model using the SNUMAT database [19], which contains 10 000 HSE band gaps (E_g^{HSE}), combined with the MP to correct the GGA band gaps (E_g^{GGA}), resulting in a revised dataset (MP_m) with improved accuracy closer to experimental values.

(ii) By cationic perturbation strategy, we obtained approximately 140 000 spinel oxides. Subsequently, we trained a high-precision classification model to screen for materials that maintain the spinel phase after ion substitution.

(iii) Due to the computational challenges associated with obtaining the specific structural parameters for such a large number of materials, structureless learning is more suitable for this study. However, structureless learning cannot handle the problem of multiple crystal structures corresponding to the same chemical formula [20]. Therefore, we made adjustments to the CrabNet [14] network structure (referred to as CrabNet_s) and used the CrabNet_s model to train a regression model on the MP_m dataset to predict the band gap of new spinel oxides.

Furthermore, to enhance the model's expressive power and prediction capability, we extracted atom-level physical quantities as descriptors for spinel oxides based on prior experimental experience (e.g., atomic number, electronegativity). Finally, using XML, we revealed the causal relationships behind the band gap of spinel oxides from both a global and individual perspective, with the dominant factor being the sum of valence electrons for cations. This approach can unveil the complex causal relationships underlying these material

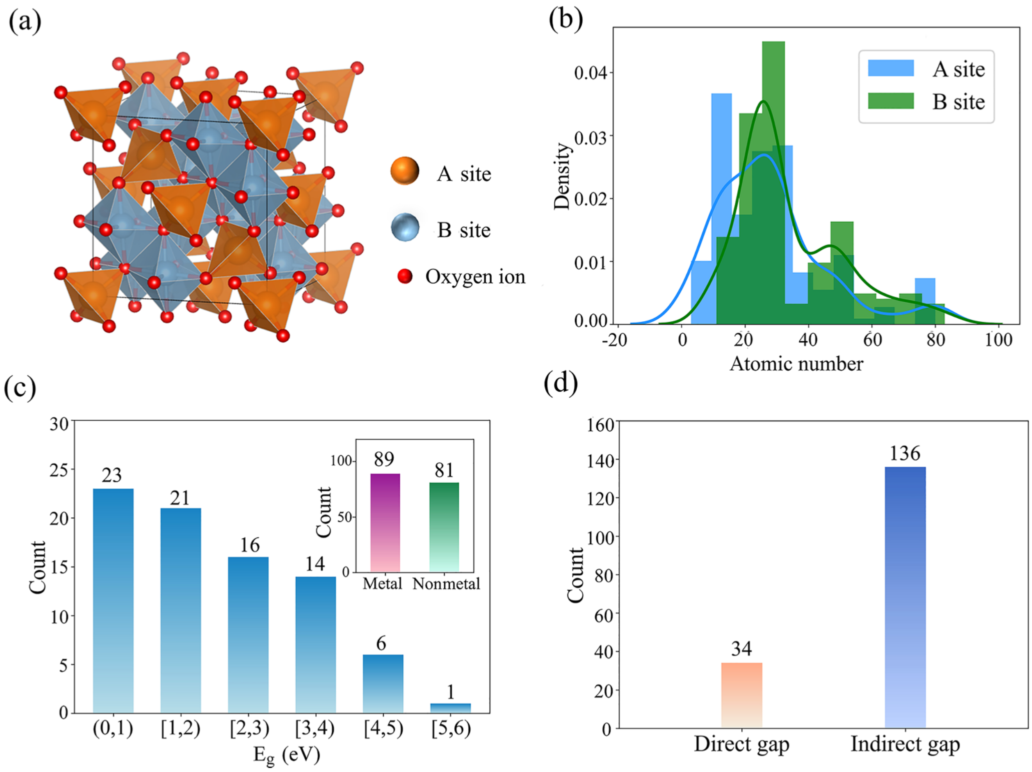

FIG. 2. Structure and data statistics of spinel oxides. (a) Spinel usually adopts a cubic phase lattice structure, in which the anion occupies the position of the body center, and the *A* site and *B* site are occupied by different metal cations, distributed in the tetrahedral and octahedral voids, respectively. (b) Data distribution of the atomic number of the *A* site and *B* site of spinel oxides in the dataset. (c) Band gap distribution of spinel oxides; metals are distinguished from semiconductors simply by whether the band gap is zero, where the main panel shows the data distribution for semiconductor materials and the inset shows the data distribution for metal and semiconductor materials. (d) Data distribution of direct and indirect band gap for spinel oxides.

properties and provides strong support for experimental and theoretical research.

II. RESULTS AND DISCUSSION

A. Initial spinel dataset

We screened out 170 spinel oxides in MP, and statistical analysis of the data was conducted. Spinel oxides with the chemical formula AB_2O_4 are significant inorganic materials that commonly exhibit a cubic structure [see Fig. 2(a)] and belong to the $Fd\bar{3}m$ space group with a space group number 227. In this structure, the *A* and *B* sites occupy 1/8 of the tetrahedral void and 1/2 of the octahedral void, respectively [21]. Figure 2(b) illustrates the distribution of atomic numbers in the dataset, where the *A* and *B* sites typically consist of transition metal elements. Figure 2(c) displays the distribution of band gaps in the dataset, where 81 materials are semiconductors and 89 materials are metals. In this study, the materials are distinguished as metallic or nonmetallic simply by whether the band gap is zero or not. Considering the band gap distribution of the semiconductor materials in the dataset, the amount of data is not sufficient to train a good regression model to accurately predict the band gaps. Furthermore, the dataset predominantly comprises materials with indirect band gaps, with only 1/5 of the materials having direct band gaps as shown in Fig. 2(d). Analysis of this dataset shows that it is difficult to train an effective interpretable model to build a bridge between cause and effect.

B. Fixing the GGA band gap in MP

As mentioned above, GGA usually underestimates the band gap, and using such data will hinder the interpretation of the model. Before continuing with the next work, our first step is to establish an effective E_g^{GGA} correction model.

The SNUMAT database provides band gap data based on hybridization generalization calculations, and E_g^{HSE} is closer to the experimental value compared to E_g^{GGA}. However, the SNUMAT database contains only about 13 000 data, which makes it difficult to train ML models with better generalization ability compared to the 150 000-volume MP database. The band gap data provided by the MP database are divided into two parts, of which 42 938 partial transition metal oxides and fluorides are calculated using GGA+U, and the remaining 111 777 data calculated using GGA [22] need to be corrected to achieve near experimental level accuracy. GGA+U is an effective optimization method [23] and, therefore, corrections are only necessary for the 111 777 data as it is essentially an approximation of the experimentally measured band gap.

Figure 3(a) shows a scatter plot of the band gap of the semiconductor material of SNUMAT, revealing that a strong linear relationship between E_g^{GGA} and E_g^{HSE} can be found:

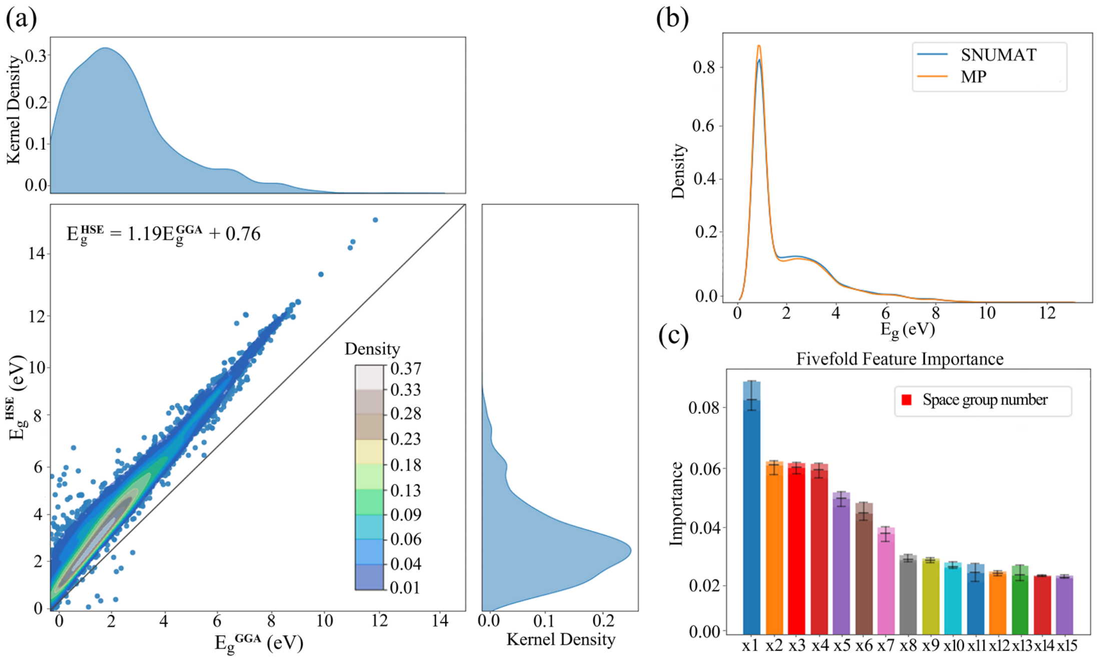

FIG. 3. Pre-data-analysis work to correct the E_g^{GGA} of MP. (a) Scatter plot and distribution plot of E_g^{HSE} and E_g^{GGA} in SNUMAT, revealing the correlation between them and the consistency of the data distribution. (b) Distribution of the E_g^{GGA} after aggregation of the SNUMAT and MP data, with overall consistency. (c) Feature importance ranking of the band gap classification model plots, with the feature importance of the space group number in red.

$E_g^{HSE} = 1.19E_g^{GGA} + 0.76$. It is feasible to use this linear relationship for the correction of the E_g^{GGA}. However, it is worth noting that this linear relationship becomes less reliable for small band gaps. To address this, richer information and an advanced integrated learning method, LightGBM [24] (LGBM), are employed to improve the accuracy of predicting E_g^{HSE}. The process of correcting the E_g^{GGA} of MP is shown in Fig. 1(a), where we first aggregate SNUMAT and MP by chemical formula and space group to obtain an intersection (for data with the same chemical formula and space group, we keep the one with the smallest absolute error between the E_g^{GGA} of MP and the E_g^{GGA} of SNUMAT). The slight error introduced by different computation methods is reasonable, and after removing a small number of outliers, the distribution of E_g^{GGA} of the two databases is nearly identical [see Fig. 3(b)]. The model is trained using the E_g^{GGA} from MP without considering the SNUMAT database, ensuring better prediction capability for unknown data by maintaining consistency in data distributions.

The E_g^{GGA} is a powerful descriptor for predicting E_g^{HSE}, and we also derived the chemical formula into 145 features related to the constituent elements [25,26] to obtain more complete information. A major drawback of component-based ML is that the chemical formulas of some compounds in the database correspond to multiple structures, making it impossible to establish a one to one mapping relationship from components to properties. The research object of this paper takes spinel as an example. Although the specific structure of each sample is different, all their space group numbers are 227. The space group represents partial properties of structures and can greatly alleviate the above shortcomings of component learning. In addition, we found that the space group feature plays a significant role in band gap prediction [see Fig. 3(c)], so it should also be taken into account. In order to avoid the "curse of dimensionality," we performed covariance feature removal and explicit forward selection of features to substantially reduce the feature size in the subsequent stages of our work. The details of features X1–X15 are shown in Table I.

The model needs to be validated for both extrapolation ability, which refers to its ability to predict data beyond the training set, and generalization ability, which pertains to its performance on arbitrary samples. In this study, 2222 data (T1) in SNUMAT, excluding the intersection set mentioned above, are completely unknown during the model's training process, allowing for a comprehensive assessment of its extrapolation ability. In addition, to verify the generalization ability of the model, we also partition the dataset into a test set (T2) at a ratio of 0.2, and the remaining part is the training set.

We adopt a hierarchical prediction approach: firstly, we use the classification model to predict whether the band gap is 0. Subsequently, a regression model is utilized to predict the non-0 band gaps. This approach offers several advantages, including the ability to distinguish between 0 band gap and non-0 band gap materials, avoiding the introduction of noise through direct regression, improving the regression performance by refining predictions for non-0-band gap materials, and effectively addressing the issue of dataset imbalance. To prevent feature redundancy and mitigate the risk of dimensionality issues, an appropriate number of features were selected for training the classification and regression models. The results of feature selection show that the performance of the model is the best when 18 features and 32 features are selected for the classification model and regression model, respectively.

TABLE I. Feature mapping table.

No.	Feature	Importance	Feature description
X1	Max MendeleevNumber	0.085	The highest Mendeleev number among the elements present in all compounds
X2	Min MeltingT	0.061	The lowest melting temperature among the compounds
X3	Space group	0.061	The space group of the structure
X4	Compound possible	0.060	A compound that may be possible or not
X5	Mode MendeleevNumber	0.049	The most frequently occurring Mendeleev number among the compounds
X6	Mean MendeleevNumber	0.044	The mean Mendeleev number among the compounds
X7	Min space groupNumber	0.037	The smallest or lowest space group number among the compounds
X8	Mean GSvolume_pa	0.027	The average ground-state volume per atom among the compounds
X9	Mean Number	0.027	The mean number among the compounds
X10	Range CovalentRadius	0.025	The difference between the maximum and minimum covalent radii among the elements or compounds
X11	Mode NUnfilled	0.022	The most frequently occurring number of unfilled electron orbitals among the compounds
X12	Mean NpUnfilled	0.022	The average number of partially unfilled electron orbitals among the compounds
X13	Min NUnfilled	0.022	The smallest number of partially unfilled electron orbitals among the compounds
X14	2-norm	0.021	The 2-norm feature of elements in material composition
X15	Range GSvolume_pa	0.021	The range of ground-state volume per atom among the compounds

Area under the receiver operating characteristic (ROC) curve (ROC_AUC) is a measure of classifier performance, which measures the magnitude of the area under the ROC curve, and is calculated in Eq. (1),

$$\mathrm{ROC\_AUC} = \int_0^1 \mathrm{TPR}[\mathrm{FPR}^{-1}(t)]dt, \tag{1}$$

where TPR represents the true positive rate (the ratio of correctly identified positive cases to the total number of positive cases), and FPR denotes the false positive rate (the ratio of negative cases incorrectly identified as positive cases to the total number of negative cases).

The coefficient of determination (R^2) is employed as an overall accuracy measure for the regression model's predictions:

$$R^2 = 1 - \frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}}. \tag{2}$$

R^2 reflects the proportion of the variance of the dependent variable that can be explained by the model, with values ranging from 0 to 1, and the closer it is to 1 indicates a better fit of the model. $\mathrm{SS_{res}}$ represents the residual sum of squares, and $\mathrm{SS_{tot}}$ denotes the total sum of squares. The sum of squared residuals quantifies the difference between actual and predicted values, while the total sum of squares captures the difference between actual values and the sample mean.

MAE (mean absolute error) is used to evaluate the average absolute error of the model's predictions. A smaller MAE indicates more accurate model predictions. The calculation formula for MAE is provided in Eq. (3): y_i represents the true value and $\hat{y}_i$ represents the predicted value.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{3}$$

According to Fig. 4, our trained ML model demonstrates excellent performance. The ROC_AUC of the classifier on T1 is 0.97, indicating strong extrapolation performance. The ROC_AUC on T2 is 0.99, indicating strong generalization ability, and the fivefold cross validation ROC_AUC, averaging around 0.99, further demonstrates the stability of the model. The regression model also exhibits good overall fit. R^2 is 0.81 and MAE is 0.44 eV on T1; R^2 is 0.96 and MAE is 0.23 eV on T2. These results indicate that the regression model possesses both extrapolation and generalization capabilities. Although a few outliers are present, the overall performance of the model is satisfactory. Finally, we obtained the improved dataset (MP_m) by using classification to predict whether the band gap in the MP database is 0 or not, and subsequently using the regression model to correct the non-0 band gaps.

We compared the improved dataset MP_m with an experimental dataset of 4604 band gap measurements provided by Zhuo *et al.* [27], which is accessible on the MATMINER [26] platform. Although this experimental dataset only includes chemical formulas and band gaps, the compositions of semiconductors significantly influence the band gaps. As the band structure and electronic properties are directly related to the composition, we can roughly estimate the approximate band gap range. To ensure accuracy and avoid errors caused by polycrystalline phenomena, we removed data points from MP_m where the chemical formula matched multiple structures. This resulted in a final dataset of 2621 data points for analysis. Comparing the predicted band gap in MP_m with the experimental band gap, we calculated the MAE to be 0.37 eV. In contrast, the MAE between the band gap calculated by GGA and the experimental band gap is 0.91 eV. In conclusion, our dataset MP_m generated in this study demonstrates closer accuracy to experimental measurements when compared to the previous dataset MP.

C. Cationic perturbation strategy

The cationic perturbation strategy involves systematically replacing the cations of the components in incremental steps to generate a large number of hypothetical materials. These materials are not necessarily physically realizable, but their purpose is to enhance the effectiveness and explanatory power of ML models. Therefore, it is not necessary to consider the

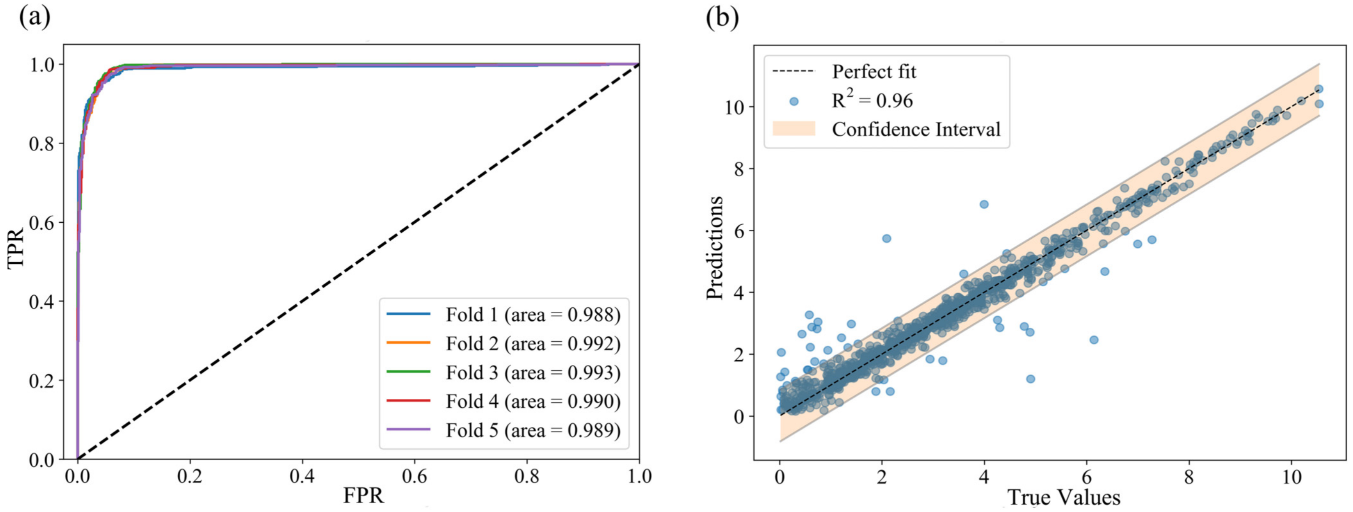

FIG. 4. Plots of model performance for predicting E_g^{HSE}. (a) Plot of the fivefold ROC curve for the classification model used to predict whether E_g^{HSE} is 0. The area below the curve is closer to 1 indicating the better classification performance of the model. (b) Plot of the performance of the regression model used to predict the specific value of E_g^{HSE}. The more the blue points converge toward the center of the best-fit line the closer the predicted value of the regression model is to the true value.

stability of these hypothetical materials; rather, the focus is on whether their components can form the space group number 227. Figure 1(b) illustrates the workflow of cation substitution in the *A* and *B* sites of the spinel structure, with the optional elements for these sites taken from the work of Wang *et al.* [28]. Since the cations in the *A* and *B* sites occupy interstitial positions within the crystal structure, we chose to replace only one cation at a time while keeping the other cation unchanged. The substitution process was carried out sequentially, with a chemical coefficient of 0.1, until complete substitution of the site was achieved. It was ensured that the same elements did not appear in both the *A* and *B* sites.

The unprocessed dataset obtained after cation substitution consists of approximately 140 000 entries. However, it is computationally challenging to calculate the crystal structure for such a large dataset, especially considering our requirement of retaining the spinel phase in the substituted material. To address this issue, we employed a suitable method based on the tolerance factor. The tolerance factor can roughly estimate whether the material can form the spinel phase [29]. However, it is difficult to estimate the ionic radius and tolerance factor (the calculation of the tolerance factor depends on the ionic radius, which is related to the oxidation valence state), particularly for spinel structures containing transition metal elements with complex and variable oxidation states. In addition, some hypothetical materials may have complex crystal structures or contain defects, which may lead to the existence of valence imbalance. To overcome this challenge, we trained a classification model to predict whether a material would form a spinel structure after cation substitution.

We created a training dataset for the classification model by selecting 170 spinel oxides as positive examples and 200 non-spinel-structured ternary and quaternary materials from the MP database as negative examples. The features used for training were derived from chemical formulas, resulting in a dataset size of 370 entries with 145 features. To construct the best ML pipeline for classification, we employed the tree-based pipeline optimization tool (TPOT). TPOT utilizes genetic algorithms to optimize ML pipelines and has shown effectiveness in solving regression and classification problems [30]. TPOT takes a lot of time and computational resources when dealing with large-scale data, but is very suitable for this dataset of only 370 data. The 370 data are divided into training and test sets in the ratio of 4:1, and then the classifier is trained on the training set using TPOT, with the population parameter set to 50 for each training, and the optimization target is the mean value of ROC_AUC for the fivefold cross validation. After ten iterations, the best model achieved a score of 0.99 on the training set and a score of 0.94 on the test set, indicating good generalization ability of the classifier. Using this classifier, we screened out 111 138 materials with spinel structure from the initial dataset of 140 000 materials.

D. Proxy model to extract tags

The 111 138 data generated after cation substitution only have information on the chemical formula and corresponding sites, while they lack the target variable (label). Therefore, it is necessary to train a generalizable model on MP_m to extract labels for this dataset. The information contained in crystal structures is richer than that in compositions, so some deep learning (DL) methods based on crystal structure as input [31–33] have achieved impressive results in predicting band gaps. However, due to the significant time and cost required for the precalculation of these 110 000 crystal structures, structure-based DL methods are dismissed. Some structure-free learning DL methods [14,29,34] have also achieved quite good results in the field of band gap prediction. In cases where the standard for model prediction accuracy can be relaxed appropriately, these structure-free DL methods are more suitable for this task.

One such structure-free learning method is CrabNet, which is a Transformer-based [35] approach for building attention graph neural networks. CrabNet achieves accurate band

TABLE II. Comparison of MAE performance of CrabNet and CrabNet_s on MP_m.

No.	MAE of CrabNet_s (eV)	MAE of CrabNet (eV)
Fold_0	0.3916	0.4126
Fold_1	0.4046	0.4286
Fold_2	0.4004	0.4193
Fold_3	0.3983	0.4186
Fold_4	0.3995	0.4178
Mean	0.3989	0.4194

TABLE III. MAE performance of CrabNet vso CrabNet_s on 170 spinel oxide datasets.

No.	MAE of CrabNet_s (eV)	MAE of CrabNet (eV)
Fold_0	0.5287	0.5641
Fold_1	0.4663	0.5856
Fold_2	0.5017	0.5617
Fold_3	0.4809	0.5484
Fold_4	0.5051	0.5761
Mean	0.4965	0.5636

gap predictions using only composition and element ratio information. It is worth noting that CrabNet fractionally encodes elemental ratio information and maps it into a high-dimensional space, making it sensitive to small changes in elements that can impact overall material properties. However, structure-free learning usually uses some methods of averaging the target variables to create a unique mapping between inputs and outputs, which may lead to incorrect predictions of compound properties. Spinel is a typical compound with polycrystalline phenomena, and failure to address this issue can lead to significant errors between the model output and the true band gap of spinel. To address this issue, we incorporate space group information, an important concept in crystallography [36]. Spinel oxides typically exhibit a specific space group symmetry (space group number 227), and within a database like MP_m, many candidate materials may have the same chemical formula but different structures and properties. Using space group features can help to accurately identify oxides with spinel structures from these candidate materials. Figure 3(c) shows the importance of space group features, and space group information can be used to guide the model to learn interactions between different elements in the crystal and capture relationships between properties, thus enhancing the representativeness of the model.

We partitioned 170 spinel oxides from MP_m as the test set, and the rest of the data are used for fivefold cross-validation. Table II shows that the performance of the revised network architecture, CrabNet_s, has improved by approximately 5% compared to the original CrabNet, with the same evaluation methodology. This improvement is attributed to the inclusion of space group information and enhanced feature representation in our model. It is worth noting here that the MAE of the CrabNet MAE on MP_m is higher than that on MP [37], which is caused by the fact that E_g^{HSE} is typically larger than E_g^{GGA}.

The test set of 170 spinel oxides was used to validate the prediction ability of the model on unknown spinel data. Table III shows that CrabNet_s achieves a significantly smaller MAE compared to CrabNet for predicting spinel oxide data, improving performance by approximately 12%. The CrabNet_s model trained by Fold_1 has the strongest prediction ability, so we use this model for the subsequent study. Among these 170 spinel oxides, 64 data have band gap records calculated by GGA. We assume that the corrected band gap in MP_m is the true value and the band gap calculated by the GGA method is the predicted value, and the MAE of these 64 data is 1.029 eV, while the MAE predicted by the Fold_1-trained CrabNet_s is 0.4663 eV. Figure 5 provides a violin plot illustrating the residual value statistics from these two methods, indicating that CrabNet_s provides more accurate and stable predictions. Therefore, the CrabNet_s model demonstrates positive prediction ability on these 170 spinel oxides. On this basis, we used the model to extract labels for the aforementioned 111 138 cation-substituted spinel oxide data.

E. Explainable machine learning

We obtained 111 138 spinel oxide data after optimizing the original set of 170 spinel oxides. This dataset now has a sufficient data size and, importantly, labels that closely correspond to experimental results. To extract features from each data point, we consider the A and B sites separately and extract difference features and summation features. Since the anion is fixed as O, these features mainly reflect the differences between cations. These features contain 12 basic physical quantities (density, dipole polarizability, covalent radius, atomic radius, first ionization, number of valence electrons, number, period, electronegativity, number of s and p electrons, number of d electrons, and Mulliken electronegativity).

The features of the substituted sites are calculated using Eq. (4):

$$X_{\mathrm{site}} = s_1E_1 + s_2E_2, \tag{4}$$

where E_1 and E_2 represent the two elements present at the site; s_1 and s_2 denote the stoichiometric numbers, respectively.

We conducted correlation analysis on the dataset to assess the associations between variables and improve the performance and reliability of the ML models. Then we selected the four features with the highest correlation with the band gap from the feature pool. The correlation coefficients between these features are visualized in the heat map shown in Fig. 6(a). The sum of the valence electron numbers of the cations (X_2^{VE}), the Mulliken electronegativity of B sites (B^{MEn}), and the sum of the densities of the cations (X_2^{D}) show a negative correlation with the band gap [see Figs. 6(b), 6(d), and 6(e)], while the average ionic properties (AIC) are positively correlated with the band gap [see Fig. 6(c)].

Before training the ML model, the data need to be analyzed and a suitable modeling strategy needs to be selected first. The dataset contains 59 601 samples with 0 band gap and 51 537 data samples with non-0 band gap. Directly predicting the band gap value by regression will cause significant errors. To address this, a hierarchical training approach is adopted,

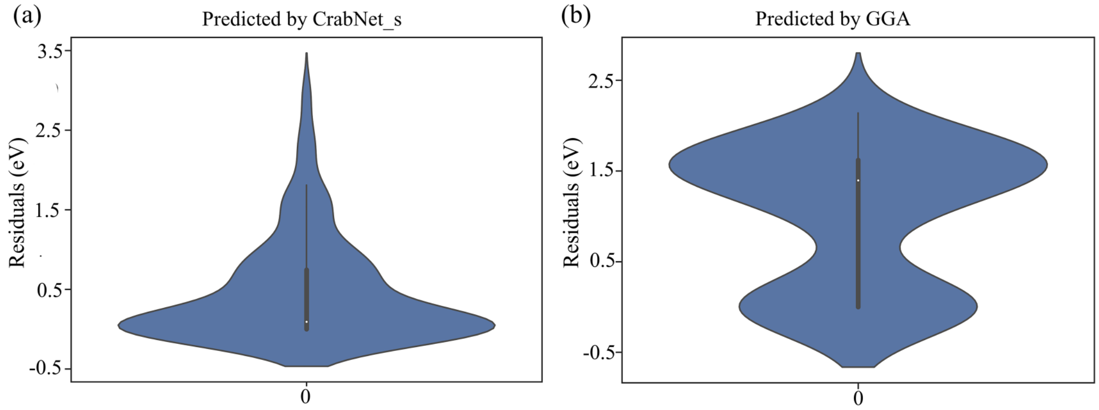

FIG. 5. The violin plot of residual statistics. (a) The violin plot shows the residual statistics of the target variable predictions by the CrabNet_s model. (b) The violin plot displays the residual statistics of the target variable predictions by the GGA method.

involving two tasks: classification and regression. Firstly, a classification model is trained to predict whether the material has a 0 band gap. Secondly, for the 35 602 data with band gap greater than or equal to 0.5 (as this study focuses on semiconductor materials that generally have a band gap greater than 0.5), a regression model is trained to predict the band gaps.

XML requires models that have good predictive performance and are not too complex. For this purpose, we employ the random forest (RF) algorithm, which is an ensemble learning algorithm based on multiple decision trees. It combines the results of each decision tree to obtain the final prediction. Random forest is simpler, easier to understand, and has good generalization ability and robustness compared to the boosting idea of the LGBM algorithm. As the amount of training data increases, the performance improves. Notably, only with a training dataset of at least 10 000 samples, the classifier can achieves good performance (ROC_AUC >= 0.85), while the original dataset only had 170 samples. This highlights the necessity of optimizing the original dataset. The random forest classifier achieves a ROC_AUC of 0.92 on the test set, while the regression task achieves an R^2 of 0.82 and a MAE of 0.38 eV. These results indicate that the model exhibits strong predictive power, and it also proves that cationic perturbation can significantly improve the performance of ML.

After confirming the model's good predictive performance, we proceeded with its interpretation using SHAP (SHapley Additive exPlanations), a method for XML models [17]. SHAP is based on Shapley values, which quantify the contribution of each feature to the model's predictions, providing a highly interpretable approach to global and local model interpretation [38]. Figure 7 shows the SHAP plot of the model. The summary_plot function from the SHAP library visualizes the importance and influence of each feature on the prediction results. As observed in the previous model-independent

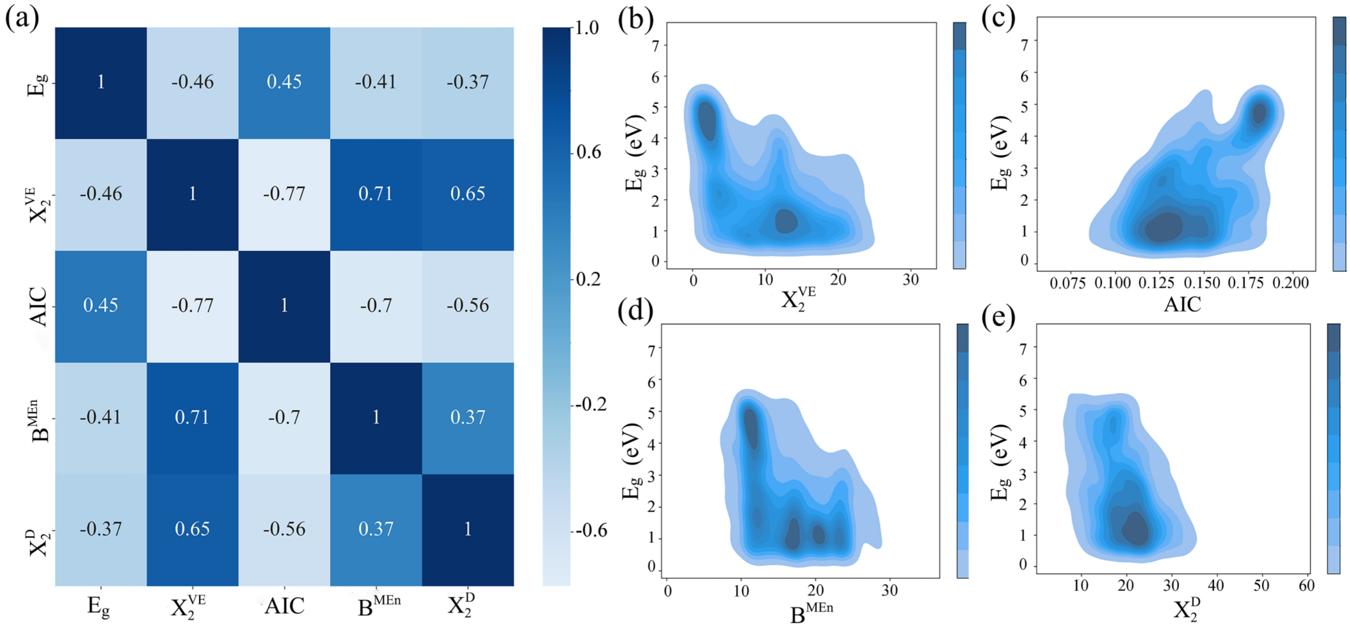

FIG. 6. Correlation analysis. (a) Heat map of these features. (b)–(e) Two-dimensional density diagrams of them with respect to the band gap. The feature naming rules are as follows: features named after elements represent the fraction of elements, *A* and *B* represent the features of corresponding sites, X_1 represents the difference of features between cations, and X_2 represents the sum of cation features.

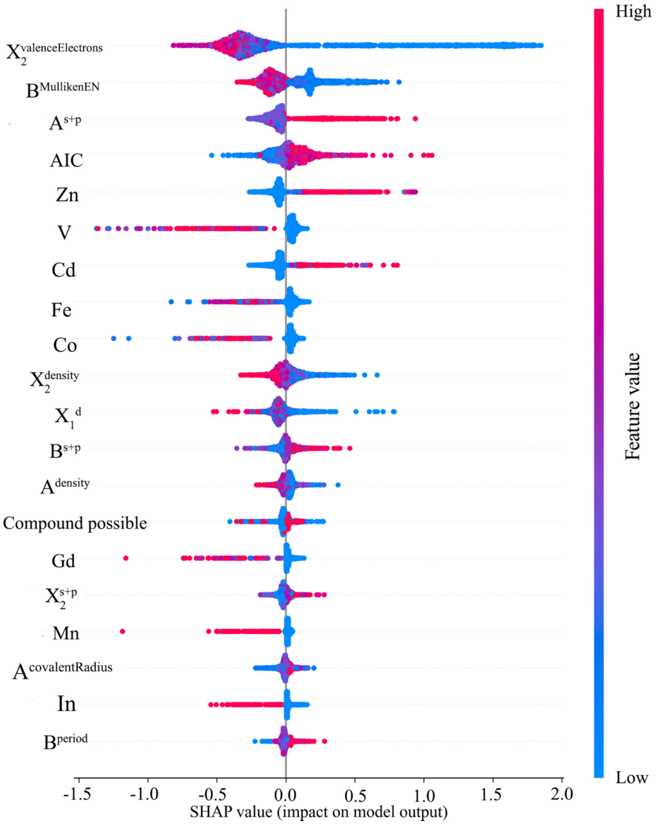

FIG. 7. SHAP plot of the model. In SHAP plots, blue points typically represent a negative impact on the model's prediction for a certain class, while red points indicate a positive impact on the prediction.

interpretation (Fig. 6), a lower number of valence electrons in the cation and a lower Mulliken electronegativity of the *B* site correspond to a larger band gap.

In solid-state band theory, the band gap is directly correlated with a material's ionicity [39]. A key parameter for measuring ionicity is the AIC. Our data mining reveals that an increase in AIC usually corresponds to a wider band gap, confirming the established correlation between band gap and ionicity. This is further supported by the concept that ionicity can be inferred from the electronegativity difference between constituent atoms of the material. Typically, a larger difference in electronegativity leads to a wider band gap [40–42].

However, our XML data analysis has unveiled a unique trend: in most samples, as the electronegativity of *B*-site atoms increases, the band gap decreases. This is likely due to the reduced electronegativity difference between these *B*-site atoms and oxygen, altering the band gap. Another critical factor is the valence electron count of cations. In spinel oxides, an increase in valence electrons, which fills the *d* orbitals more, tends to reduce the band gap. This reduction results from the *d*-orbital energy levels splitting due to crystal field effects and intensified electron-electron interactions.

Overall, these effects could lead to a reduction in the band gap, but this is not a fixed rule. Instead, it is subject to the interplay of several factors, including the specific properties of the material, electronic configuration, and crystal field effects. Therefore, understanding and predicting changes in the band gap requires a comprehensive consideration of these complex interactions by XML. These findings adhere to the fundamental laws of physics and are vitally important in materials science. They relate directly to the electronic properties and potential applications of materials, offering valuable insights into their functionality and uses.

In addition, from Fig. 7, we can see that the introduction of various transition metals affects the band gap of the material. To investigate this effect, we selected several transition metals for study. $ZnBi_2O_4$, a material of interest in photocatalytic applications, has an experimentally measured band gap ranging from 2.2 to 3 eV [43]. Figures 8(b) and 8(c) show the band gap regulation by cation substitution at the *A* and *B* sites. It can be observed that the introduction of all transition metals, except Cd, leads to a smaller band gap than the initial value (2.25 eV). This is probably due to Cd being in the last group of transition metals, where all electrons are already paired, resulting in lower electron activity, consistent with the finding in Fig. 7. The introduction of other transition metals reduces the band gap, potentially because the position of the valence band top (VBM) in spinel is often determined by the *d* orbitals of transition metals. As the number of electrons filled in the e_g orbitals increases, the Coulomb interaction between the *d* orbitals and neighboring orbitals strengthens, leading to a narrowing of the band gap. In Fig. 8(a), features that increase the prediction are represented in red, while features that decrease the prediction are represented in blue. The length of the lines indicates the magnitude of the feature's impact on the output. By examining the scale values on the *x* axis, we can observe the amount of increase or decrease in the influence. It can be concluded that the higher the valence electron count of the cations, the smaller the band gap, which is consistent with the previous analysis, further confirming the inference made in Fig. 6(b). It also can be seen that Zn promotes the band gap value more than Fe does, accounting for the smaller band gap in Figs. 8(b) and 8(c) when Fe is introduced.

Understanding the material's stability is crucial for predicting its performance in practical applications. The "energy above hull" (ΔEh) is a common and effective method for measuring stability. A material is considered thermodynamically stable if its ΔEh is small, typically less than 0.2 eV/atom. Consequently, we trained a ML model with ΔEh as the target, achieving a MAE of 0.029 eV/atom. Subsequently, we analyzed the feature importance of the model and found that the electronegativity of the *B* site is the most significant factor affecting stability. As the electronegativity of the *B* site increases, the material's band gap tends to decrease, but its ΔEh increases, thereby destabilizing the material. This could be due to the fact that an increase in the cation's electronegativity reduces the difference in electronegativity between the cations and anions, leading to weaker ionic bonds, thus destabilizing the material.

III. METHODS

The model was trained on Nvidia GeForce RTX 4080, and the neural network was built based on TORCH1.12.0+CU113. The version of CrabNet is 2.0.8, and all parameters are set by

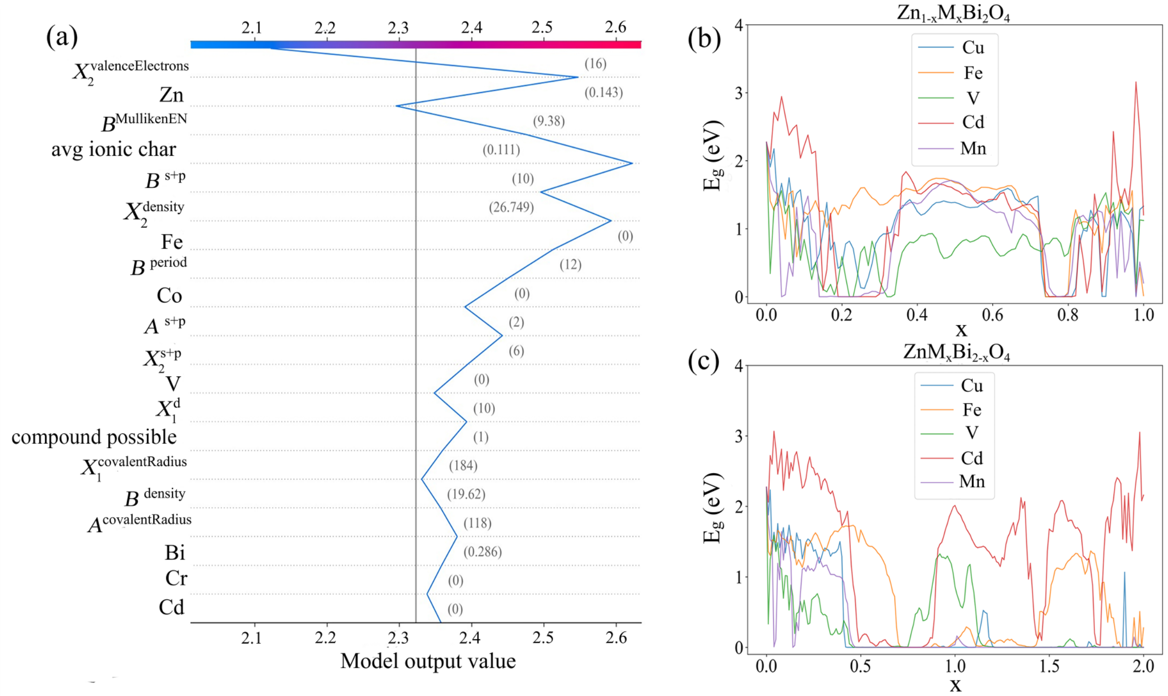

FIG. 8. Cation substitution and predicted band gap for $ZnBi_2O_4$, with M as transition metal element. (a) Feature decision map for predicting $ZnBi_2O_4$. (b) Ratio between A-site substitution and band gap. (c) Ratio between B-site substitution and band gap.

default. We adapted the network architecture of CrabNet by introducing the space group information. To prevent overfitting caused by having too many layers in the neural network, we have set the number of layers in the TransformerEncoderLayer to 2, the dropout ratio to 0.2, the output dimension of the fully connected layer to 1024, and the rest of the parameters remain unchanged, and the model parameter scale is 11992839.

In this work, we utilized a graph neural network (GNN) with an incorporated attention mechanism. The details of the model are described as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (5)$$

$$h_i^t = \sigma\left[\sum_{j\in N_i} \alpha\left(h_i^{t-1}, h_j^{t-1}\right) W^{t-1} h_j^{t-1}\right]. \qquad (6)$$

In Eq. (6), α represents an attention function that adaptively controls the contribution of neighboring node j to node i. In order to learn the attention weights for different subspaces, GNN can also employ multiple attention mechanisms:

$$h_i^t = ||_{k=1}^K \sigma\left[\sum_{j\in N_i} \alpha\left(h_i^{t-1}, h_j^{t-1}\right) W_k^{t-1} h_j^{t-1}\right]. \qquad (7)$$

By incorporating the attention mechanism, the CrabNet model can optimize feature representation and improve performance and accuracy. However, the original CrabNet model lacks representation of spatially symmetric information despite its optimized feature representation for component information in the final output of the attention layer. Therefore, we added the space group information to CrabNet and modified the original network architecture. The inputs to the model consist of both chemical composition and symmetry information. The chemical composition related inputs include matrices derived from atomic numbers and stoichiometry. We changed the operation rule of these two matrices in the original work from addition to multiplication, to weight the elemental information by stoichiometric numbers. The input related to the symmetric information is unique in this study and is represented by the matrix derived from the space group number (SDM). Since the space group number information is one dimensional, and in order to align it with the dimensionality of the component information, we first map it to a high-dimensional space using a fully connected layer. Then we apply an attention layer to optimize its feature representation and obtain the input matrix. The process of obtaining the input matrix is shown in Fig. 9.

Detailed information about the revised network architecture (CrabNet_s) is in Fig. 10. The composition information and symmetry information belong to the same layer. Therefore, after obtaining the final feature representation of the element information (EDM′), we concatenate the space group information (SDM) to it to obtain the global information representation of the material (GDM). The self-attentive layer is then repeated N times to optimize the global information feature representation and obtain the final feature input (GDM′). After multiple Transformer layers, the purpose is to extract features that are closer to objective physical laws, thereby obtaining a higher level of model performance.

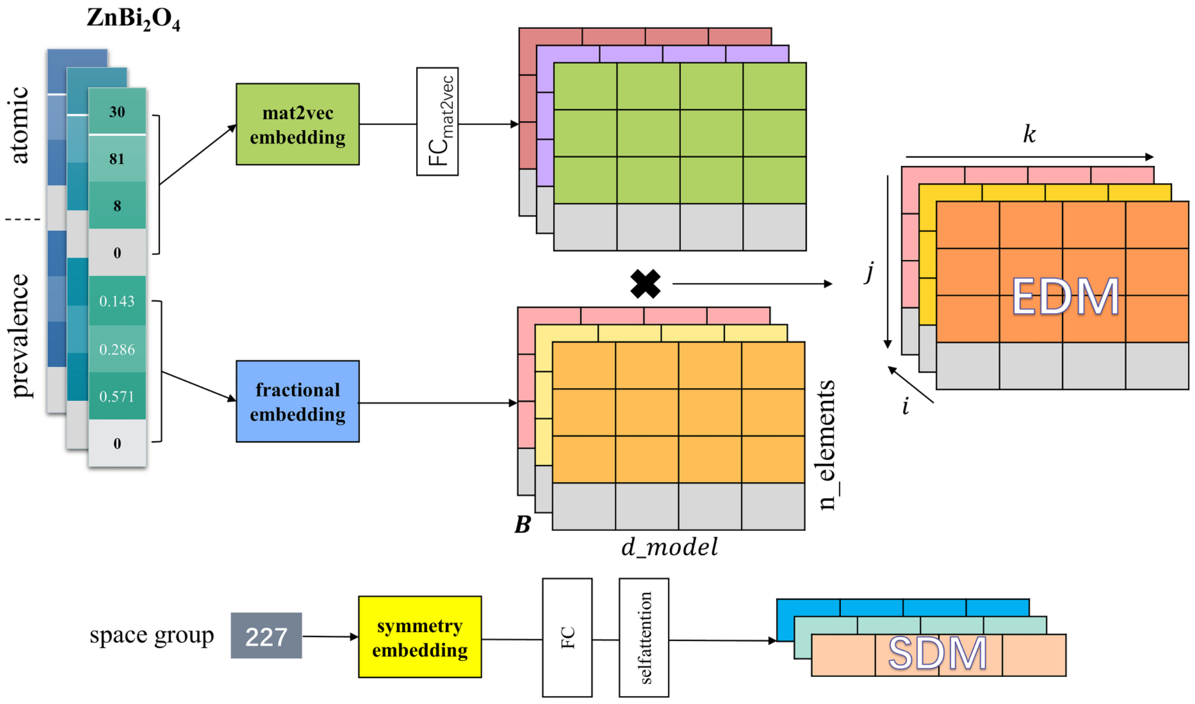

FIG. 9. Schematic representation of the derivative matrix of $ZnBi_2O_4$, where B denotes the batch, *d_model* denotes the number of elemental characteristic dimensions, and *n_elements* denotes the number of elements.

IV. SUMMARY

In materials science, we often face challenges related to poor data quality and limited size. This study successfully overcomes these challenges. Our method for generating large-scale HSE level data is not limited to spinel oxides; it is also applicable to other material systems with fixed formulas and structures, such as perovskite halides, which follow the formula ABX_3.

To tackle the quality issues in the original dataset, we expanded the initial 170 entries using a cation perturbation method, along with HSE level predictions. This approach generated 111 138 valid entries, establishing a strong

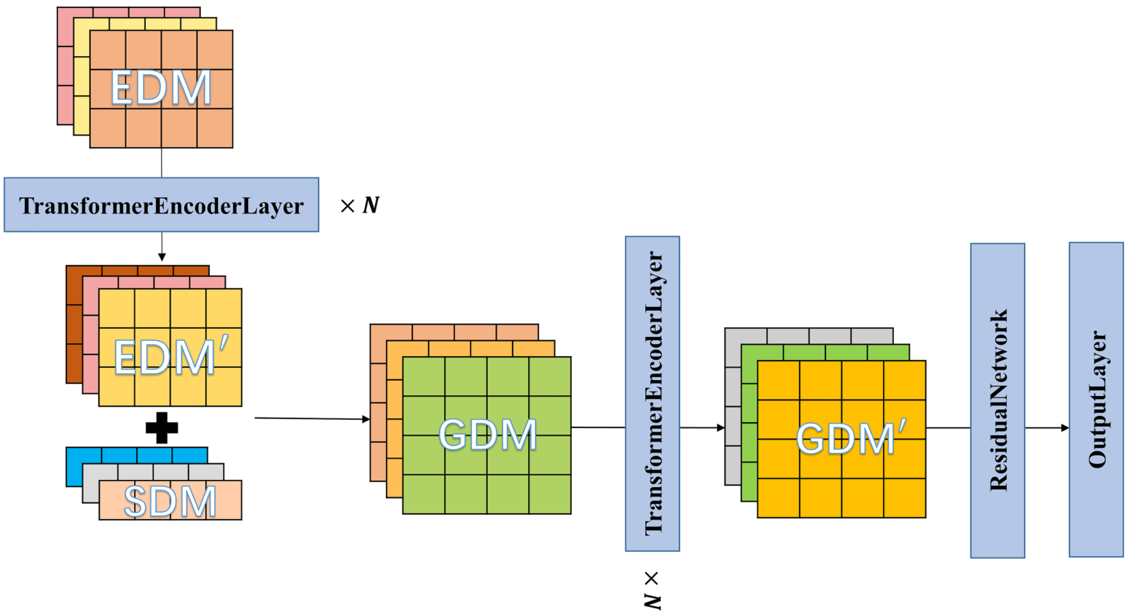

FIG. 10. Schematic diagram of CrabNet_s structure, including input EDM, self-attentive layer (repeated *N* times), updated element representation EDM′, space group input SDM, global information input (GDM), self-attentive layer (repeated *N* times), updated global information representation GDM′, residual network, and the final model output.

foundation for future XML studies. Calculations at the HSE level are usually time consuming. Processing 111 138 materials quickly using traditional methods is impractical. Therefore, our method is notably efficient.

During our research, we created a dataset, MP_m, which nearly matches experimental accuracy. We improved handling polycrystalline structures in structure-free learning, enhancing the CrabNet model's performance. We identified factors affecting the band gap in spinel oxides. For example, we quantified the impact of cations' valence electron count and the electronegativity of *B*-site elements. These findings align with physical intuition. Our work also shows how changing elements and their amounts affects material properties. Once validated, these causal relationships could guide improvements in material performance in experimental settings.

The dataset and code during the current study are available in the GitHub repository [44].

ACKNOWLEDGMENTS

This work was partially sponsored by National Natural Science Foundation of China (Grants No. 52076126 and No. 12174246), National Key Research and Development Program (Grant No. 2021YFE0109800), Key Laboratory of Clean Power Generation and Environmental Protection Technology in Mechanical Industry, Science and Technology Commission of Shanghai Municipality (Grant No. 21010501300), and the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Grant No. 22SG51).

[1] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), JOM **65**, 1501 (2013).

[2] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies, npj Comput. Mater. **1**, 15010 (2015).

[3] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[4] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli *et al.*, Aflowlib.org: A distributed materials properties repository from high-throughput *ab initio* calculations, Comput. Mater. Sci. **58**, 227 (2012).

[5] M. Yu, S. Yang, C. Wu, and N. Marom, Machine learning the Hubbard U parameter in DFT+U using Bayesian optimization, npj Comput. Mater. **6**, 180 (2020).

[6] J. Snoek, H. Larochelle, and R. P. Adams, in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Red Hook, NY, 2012).

[7] X. Chen, X. Liu, X. Shen, and Q. Zhang, Applying machine learning to rechargeable batteries: From the microscale to the macroscale, Angew. Chem. Int. Ed. Engl. **60**, 24354 (2021).

[8] M. Kim, M. Y. Ha, W. B. Jung, J. Yoon, E. Shin, I. D. Kim, W. B. Lee, Y. Kim, and H. T. Jung, Searching for an optimal multi-metallic alloy catalyst by active learning combined with experiments, Adv. Mater. **34**, 2108900 (2022).

[9] C. D. Lv, X. Zhou, L. X. Zhong, C. S. Yan, M. Srinivasan, Z. W. Seh, C. T. Liu, H. G. Pan, S. Z. Li, Y. G. Wen *et al.*, Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries, Adv. Mater. **34**, e2101474 (2022).

[10] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. **29**, 5090 (2017).

[11] S. Takamoto, C. Shinagawa, D. Motoki, K. Nakago, W. W. Li, I. Kurata, T. Watanabe, Y. Yayama, H. Iriguchi, Y. Asano *et al.*, Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements, Nat. Commun. **13**, 2991 (2022).

[12] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, Machine learning in materials informatics: Recent applications and prospects, npj Comput. Mater. **3**, 54 (2017).

[13] A. Nguyen, J. Yosinski, and J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2015), p. 427.

[14] A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, Compositionally restricted attention-based network for materials property predictions, npj Comput. Mater. **7**, 77 (2021).

[15] X. T. Zhong, B. Gallagher, S. S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y. J. Han, Explainable machine learning in materials science, npj Comput. Mater. **8**, 204 (2022).

[16] H. Nori, S. Jenkins, P. Koch, and R. Caruana, InterpretML: A unified framework for machine learning interpretability, arXiv:1909.09223.

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

[18] Y. Liu, Z. W. Yang, Z. Y. Yu, Z. T. Liu, D. H. Liu, H. L. Lin, M. Q. Li, S. C. Ma, M. Avdeev, and S. Q. Shi, Generative artificial intelligence and its applications in materials science: Current situation and future perspectives, J. Materiomics **9**, 798 (2023).

[19] S. Kim, M. Lee, C. Hong, Y. Yoon, H. An, D. Lee, W. Jeong, D. Yoo, Y. Kang, Y. Youn *et al.*, A band-gap database for semiconducting inorganic materials calculated with hybrid functional, Sci. Data **7**, 387 (2020).

[20] Z. Song and Q. Liu, Tolerance factor and phase stability of the normal spinel structure, Cryst. Growth Des. **20**, 2014 (2020).

[21] Y. Li, B. Xiao, Y. Tang, F. Liu, X. Wang, F. Yan, and Y. Liu, Center-environment feature model for machine learning study of spinel oxides based on first-principles computations, J. Phys. Chem. C **124**, 28458 (2020).

[22] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, A high-throughput

infrastructure for density functional theory calculations, Comput. Mater. Sci. **50**, 2295 (2011).

[23] Á. Morales-García, R. Valero, and F. Illas, An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations, J. Phys. Chem. C **121**, 18862 (2017).

[24] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017).

[25] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. **2**, 16028 (2016).

[26] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla *et al.*, MATMINER: An open source toolkit for materials data mining, Comput. Mater. Sci. **152**, 60 (2018).

[27] Y. Zhuo, A. M. Tehrani, and J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. Lett. **9**, 1668 (2018).

[28] Z. Wang, H. Zhang, and J. Li, Accelerated discovery of stable spinels in energy systems via machine learning, Nano Energy **81**, 105665 (2021).

[29] A. Ihalage and Y. Hao, Formula graph self-attention network for representation-domain independent materials discovery, Adv. Sci. **9**, 2200164 (2022).

[30] T. T. Le, W. Fu, and J. H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector, Bioinformatics **36**, 250 (2020).

[31] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. **120**, 145301 (2018).

[32] K. Choudhary and B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Comput. Mater. **7**, 185 (2021).

[33] R. Ruff, P. Reiser, J. Stühmer, and P. Friederich, Connectivity optimized nested graph networks for crystal structures, arXiv:2302.14102.

[34] R. E. A. Goodall and A. A. Lee, Predicting materials properties without crystal structure: Deep representation learning from stoichiometry, Nat. Commun. **11**, 6280 (2020).

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017).

[36] D. Hestenes and J. W. Holt, Crystallographic space groups in geometric algebra, J. Math. Phys. **48**, 023514 (2007).

[37] R. Jacobs, T. Mayeshiba, B. Afflerbach, L. Miles, M. Williams, M. Turner, R. Finkel, and D. Morgan, The Materials Simulation Toolkit for Machine Learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research, Comput. Mater. Sci. **176**, 109544 (2020).

[38] Y. Zhang, X. He, Z. Chen, Q. Bai, A. M. Nolan, C. A. Roberts, D. Banerjee, T. Matsunaga, Y. Mo, and C. Ling, Unsupervised discovery of solid-state lithium ion conductors, Nat. Commun. **10**, 5260 (2019).

[39] J. C. Phillips, Ionicity of the chemical bond in crystals, Rev. Mod. Phys. **42**, 317 (1970).

[40] Y. Makino, Interpretation of band gap, heat of formation and structural mapping for *sp*-bonded binary compounds on the basis of bond orbital model and orbital electronegativity, Intermetallics **2**, 55 (1994).

[41] F. Di Quarto, C. Sunseri, S. Piazza, and M. C. Romano, Semiempirical correlation between optical band gap values of oxides and the difference of electronegativity of the elements. Its importance for a quantitative use of photocurrent spectroscopy in corrosion studies, J. Phys. Chem. B **101**, 2519 (1997).

[42] J. A. Duffy, Trends in energy gaps of binary compounds: An approach based upon electron transfer parameters from optical spectroscopy, J. Phys. C: Solid State Phys. **13**, 2979 (1980).

[43] V.-H. Nguyen, M. Mousavi, J. B. Ghasemi, Q. V. Le, S. A. Delbari, A. Sabahi Namini, M. Shahedi Asl, M. Shokouhimehr, and M. Mohammadi, Novel $p-n$ heterojunction nanocomposite: TiO_2 QDs/$ZnBi_2O_4$ photocatalyst with considerably enhanced photocatalytic activity under visible-light irradiation, J. Phys. Chem. C **124**, 27519 (2020).

[44] See https://github.com/ccv81121/CationPerturbationML.git