Physica B 717 (2025) 417821

Contents lists available at ScienceDirect

Physica B: Condensed Matter

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

Machine learning prediction of bandgap and formation energy in two-dimensional metal oxides

Wen Yao [a], Wanli Jia [a,*], Ruofan Shen [a], Jiayao Wang [a], Lin Zhang [a], Xinmei Wang [a,b,**]

[a] School of Science, Xi'an University of Technology, Xi'an 710048, China
[b] School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China

ARTICLE INFO

Keywords:
Two-dimensional metal oxides
Perovskite oxides
Machine learning
Energy property prediction
Data-driven
Feature engineering

ABSTRACT

Two-dimensional (2D) transition metal oxides (TMOs) including perovskite oxides with tunable band gaps offer promising opportunities in optoelectronics, energy storage, catalysis, and sensing applications. In this work, we propose a machine learning (ML)-based framework for the accurate prediction and analysis of the band gap and formation energy of 2D TMOs. A comprehensive feature engineering strategy was employed to construct 120 physical descriptors, followed by feature selection using Pearson correlation coefficients and feature importance rankings. We evaluated seven machine learning algorithms across six prediction tasks encompassing various material types, scales, and target properties. Among them, eXtreme Gradient Boosting (XGBoost) and Gradient Boosting Decision Tree (GBDT)—implemented via Gradient Boosting Classifier for classification tasks and Gradient Boosting Regressor for regression tasks—consistently exhibited superior performance. In the classification of electronic band types, XGBoost achieved an accuracy of 95.4 %, while the Gradient Boosting Classifier reached 92.3 %. For the regression prediction of band gaps and formation energies, both XGBoost and Gradient Boosting Regressor attained coefficients of determination (R^2) close to 0.90. Furthermore, SHapley Additive exPlanations (SHAP) analysis provided interpretability by identifying dominant features influencing each property. The bandgap was primarily governed by the average number of d-orbital valence electrons, the proportion of s-orbital valence electrons, oxygen content (variable only in 2D oxides), and average atomic mass. In contrast, formation energy exhibited strong correlations with the electronegativity range, oxygen content in 2D oxides, and average d-orbital valence electron count. This study offers a robust and interpretable predictive approach for accelerating the screening and rational design of 2D TMOs, potentially reducing computational costs in high-throughput materials discovery workflows.

1. Introduction

Two-dimensional (2D) transition metal oxides (TMOs) have attracted considerable attention due to their unique physicochemical properties, such as tunable bandgaps, chemical stability, and diverse charge transport characteristics. These materials have been widely explored for applications in photocatalysis, gas sensing, battery electrodes, and catalysis, especially in green technologies like environmental purification and clean energy [1–3]. While conventional 2D materials such as graphene, MoS_2, and h-BN have been extensively studied, systematic investigations into 2D TMOs remain relatively underdeveloped. In particular, the structure-property relationship involving electronic (e.g., bandgap) and thermodynamic (e.g., formation energy) characteristics has not been fully established [2–4]. Traditional trial-and-error experimental synthesis and first-principles computational screening based on density functional theory (DFT) computational screening are both resource-intensive and time-consuming, posing challenges for large-scale materials discovery. Although first-principles calculations have been employed to predict the properties of certain 2D TMOs [3–5], their high computational cost limits their scalability. In contrast, machine learning (ML), as a data-driven approach, offers a more efficient alternative by leveraging existing data to predict unknown properties and underlying trends, thereby accelerating the exploration of promising materials.

Recent advances have demonstrated the potential of ML in predicting key properties like bandgaps, formation energies, and thermal

* Corresponding author.
** Corresponding author. School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China
E-mail addresses: jiawanli@xaut.edu.cn (W. Jia), wangxinmei@xaut.edu.cn (X. Wang).

https://doi.org/10.1016/j.physb.2025.417821
Received 13 May 2025; Received in revised form 1 September 2025; Accepted 18 September 2025
Available online 19 September 2025
0921-4526/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

conductivities for materials such as metal-organic frameworks, transition metal dichalcogenides, and perovskite oxides [6–10]. Several studies have employed machine learning (ML) techniques to predict key properties of perovskite-type and binary oxides, including bandgap, thermodynamic stability, and optical characteristics. A range of models—such as Support Vector Machines (SVM), Random Forests (RF) [11], and gradient boosting algorithms—have been utilized. For example, Priyanga et al. [12] used elemental descriptors to classify the bandgap nature (direct vs. indirect) of ABO_3 perovskite oxides. Mbiti and Titus [13] proposed a geometry- and topology-based framework for analyzing oxide structures. Smarak et al. [14] applied ML methods to identify direct bandgap perovskites for photovoltaic applications. However, to the best of our knowledge, these studies are often constrained by limited dataset sizes, insufficient incorporation of structure-sensitive features, and narrow compositional diversity—factors that limit their scalability and generalizability.

One major limitation is the lack of a systematic feature engineering strategy tailored to the unique structural and compositional characteristics of 2D TMOs, which compromises the accuracy of property predictions. Additionally, the absence of high-throughput predictive frameworks built upon large-scale DFT datasets restricts the efficient screening of candidate materials. Many existing models are confined to narrow material classes or small datasets, resulting in limited generalizability to previously unseen or chemically diverse compounds. Moreover, prior studies often treat bandgap and formation energy as independent targets, neglecting potential correlations between these properties and failing to leverage unified multi-task learning strategies. These gaps highlight the need for more comprehensive and scalable approaches to accelerate the discovery and design of functional 2D MOs [15–21].

In this work, we address these challenges by developing a comprehensive, interpretable, and high-accuracy ML framework for predicting both bandgap and formation energy of 2D MOs and perovskite oxides. Unlike previous approaches, our method integrates both element- and structure-based descriptors and applies rigorous feature selection and model interpretation using SHAP analysis. We systematically benchmark seven ML algorithms and demonstrate that ensemble methods such as XGBoost and GBDT outperform traditional models across classification and regression tasks. Importantly, our framework is scalable and validated on a large dataset of over 20,000 previously uncharacterized perovskite oxides, demonstrating strong generalization and predictive reliability [22,23].

To enhance prediction accuracy and extend applicability, we propose an ensemble-based workflow for material property prediction. This framework incorporates universal descriptors and high-precision ML algorithms, demonstrating strong performance in classification and regression tasks. The method is further applied to perovskite oxide systems, where random sampling verification confirms the high reliability of the predictions [24]. These findings provide a solid theoretical and computational foundation for accelerating the discovery and design of 2D metal oxide materials.

The remainder of this article is structured as follows: Section II details the feature engineering methodology, model training procedures, and evaluation metrics; Section III presents the predictive performance of the ML models, discusses key findings, and evaluates the implications of the proposed approach; Section IV summarizes the conclusions and outlines prospective research directions.

2. Computational details

2.1. Data set and ML models

All computations and analyses in this study were performed within a Python environment, utilizing essential libraries including Matminer [25], Pymatgen [26], Scikit-Learn [27], XGBoost [28], SHapley Additive exPlanations (SHAP) [29], Pandas [30], and NumPy [31]. The overall workflow (Fig. 1) consists of data acquisition and preprocessing, feature engineering, model development and optimization, and performance evaluation.

The dataset utilized in this study is primarily sourced from the Materials Project [32], JARVIS-DFT [33], and relevant literature [34]. All raw data underwent rigorous preprocessing, including the removal of entries lacking bandgap values, deduplication, and the exclusion of entries with weakly correlated features. Based on chemical composition and structural attributes, multiple datasets were systematically curated to facilitate feature extraction and model training. Specifically, the datasets were designated as follows: DataBase_T_direct (DB_{Td}, 572 entries) for the bandgap properties of TMOs; DataBase_M_bandgap (DB_{Mb}, 5,642 entries) and DataBase_M_formation (DB_{Mf}, 5,642 entries) for the bandgap and formation energy of 2D TMOs (excluding perovskites). For both 2D TMOs and perovskite oxides, the formation energy values were obtained directly from the Materials Project database. In MP, the formation energy per atom is defined as the energy difference between the compound and its constituent elemental reference states, normalized by the total number of atoms in the formula unit. All values are reported in units of eV/atom. In this study, the term "perovskite oxides" specifically refers to two-dimensional ABO_3-type compounds. Their representative atomic structures are illustrated in Fig. 2 for clarity. The corresponding datasets include DataBase_P_bandgap (DB_{Pb}, 3,744 entries), DataBase_P_formation (DB_{Pf}, 4,914 entries), and DataBase_P_direct (DB_{Pd}, 2, 472 entries), representing bandgap, formation energy, and bandgap type, respectively. Additionally, perovskite oxides with undefined properties were categorized under DataBase_unknown (DB_{un}, 21,887 entries). The systematic construction of these datasets establishes a robust foundation for subsequent investigations.

Furthermore, the DB_{un} dataset was derived from literature-reported perovskite data [35], encompassing over 60,000 entries with clear perovskite categorization. This dataset includes both confirmed and newly proposed perovskite materials, which were filtered by excluding compositions containing non-metallic elements such as Cl, Br, and F. As a result, 21,887 perovskite oxide entries were selected and designated as DB_{un}, providing extensive data support for the prediction tasks in this study. The names of all datasets, together with their corresponding entry counts and class distributions, are presented in Table 1.

Subsequently, to ensure data accuracy and usability, the preprocessing workflow was streamlined as follows. First, a component analysis based on the chemical formula (general formula MO_2, as illustrated in Fig. 4 for selected transition metals) was performed. Compounds containing elements other than transition metals and oxygen were excluded, and duplicate entries—identified by comparing bandgap values and properties—were removed, while retaining polymorphic information. This process resulted in the creation of a TMO dataset (DB_{Td}) comprising 572 bandgap entries, with indirect bandgap entries labeled as "0" and direct bandgap materials labeled as "1".

2.2. Structural descriptors

To extract comprehensive material descriptors, we utilized Matminer—a Python library seamlessly integrated with Pandas—to generate feature descriptors from both composition and crystal structure. Using the MPRester module from Pymatgen, structure objects, including cell parameters and atomic positions, were constructed for each material. Subsequently, nine structural descriptors were generated through the Structural Heterogeneity module, and an additional 120 elemental descriptors were derived using the“Meredig”module, covering properties such as atomic fraction, average atomic number, atomic weight, radius, electronegativity, and valence electron counts [36]. These features facilitated the creation of the 2D metal oxide dataset (DBMb), which includes material formulas and corresponding bandgap values, along with associated datasets: DB_{Pb}, DB_{Pf}, and DB_{Pd}.

For the datasets of 2D MOs and perovskite oxides, an element-based characterization method was adopted. Element descriptors were

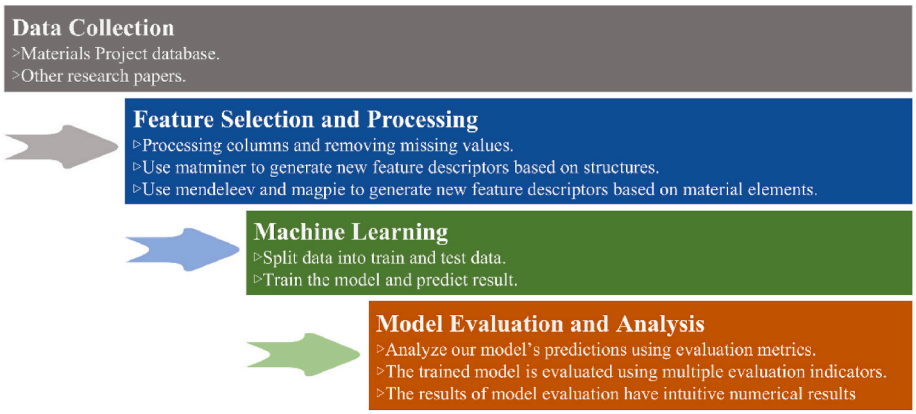

Fig. 1. Schematic representation of the machine learning workflow. The diagram outlines the sequential steps in the ML-driven methodology, including data acquisition, feature engineering, model training, evaluation, and predictive analysis.

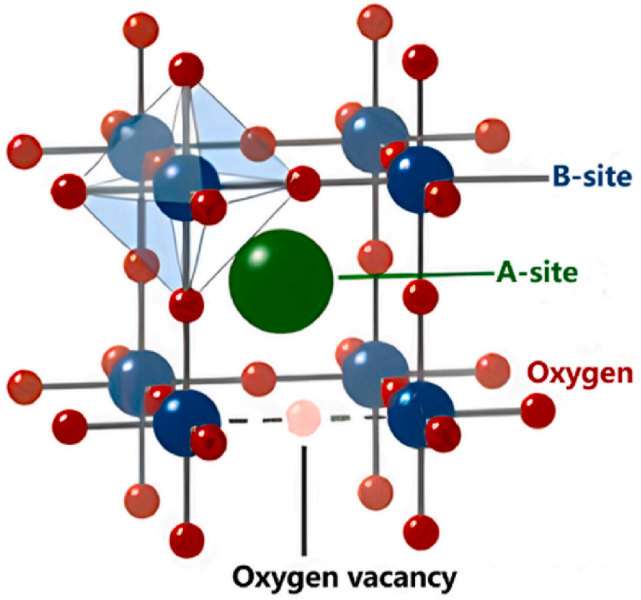

Fig. 2. Schematic diagram of perovskite oxide (ABO_3) structure.

Table 1
Overview of datasets used in this study. Both the full dataset name and the abbreviation used throughout the manuscript are listed. The total number of entries for each dataset is provided, along with the class distribution for direct ("1") and indirect ("0") bandgap materials where applicable. Regression datasets (DB_{Mb}, DB_{Mf}, DB_{Pb}, DB_{Pf}, DB_{un}) do not include categorical labels.

Dataset Full Name	Abbreviation	Entries	Class Distribution
DataBase_T_direct for bandgap properties of TMOs	DB_{Td}	572	Indirect "0": 545; Direct "1": 27
DataBase_P_bandgap for bandgap of 2D ABO_3 perovskites	DB_{Pd}	2472	Indirect "0": 2,038; Direct "1": 434
DataBase_M_bandgap for bandgap of 2D TMOs (excluding perovskites)	DB_{Mb}	5642	N/A
DataBase_P_bandgap for bandgap of 2D ABO_3 perovskites	DB_{Pb}	3744	N/A
DataBase_M_formation for formation energy of 2D TMOs (excluding perovskites)	DB_{Mf}	5642	N/A
DataBase_P_formation for formation energy of 2D ABO_3 perovskites	DB_{Pf}	4914	N/A
DataBase_unknown for perovskite oxides with undefined properties	DB_{un}	21887	N/A

generated using Matminer's Meredig module, the Magpie package [37], and the Mendeleev method [38], capturing key information on chemical stoichiometry, elemental properties, electronic structure, and ionic characteristics (e.g., element content, atomic properties, valence electron counts, ionic radii, and electronegativity).

To ensure model robustness and avoid overfitting, feature selection was conducted in three stages: (i) Pearson correlation analysis was used to filter out features with very low correlation to the target variable ($|r| < 0.1$). The Pearson coefficient r was calculated using sample covariance and standard deviation [39,40]. Although Pearson filtering may overlook non-linear or categorical relationships [41], it provides a fast and widely adopted approach in materials informatics for eliminating weakly relevant descriptors [42]. (ii) Highly collinear features with strong inter-feature correlation ($|r| > 0.8$) were removed to reduce redundancy. (iii) Feature importance rankings derived from tree-based models were used to retain features contributing cumulatively to 85 % of total importance. Steps (i)–(iii) were executed sequentially within the workflow, and thus only the overall numbers of features before and after the complete selection process can be reported. Specifically, DB_{Pd} decreased from 249 to 114 features, DB_{Td} from 119 to 114, DB_{Pf} from 249 to 163, DB_{Pb} from 249 to 160, DB_{Mb} from 252 to 146, and DB_{Mf} from 252 to 152. This multi-stage process ensured that the final feature matrix was both informative and non-redundant. Among the features, oxygen content is only variable in the 2D transition metal oxide dataset (e.g., DB_{Mb}, DB_{Mf}), while in ABO_3 perovskites it is fixed by stoichiometry. Therefore, oxygen content contributes to feature differentiation only in the former case.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (1)$$

Based on this optimized matrix, we initially employed seven ML algorithms—including XGBoost, RF [11], Logistic Regression (LR) [43], Support Vector Classifier (SVC) [44], MLPClassifier [45], Decision Tree Classifier (DTC) [46], and Gradient Boosting Decision Tree (GBDT) [47]—to predict the bandgap properties of monolayer transition MOs. The MLPClassifier, with its nonlinear approximation capabilities, was utilized to handle both binary and multi-class tasks. To ensure robust performance, 10-fold cross-validation combined with RandomizedSearchCV [48] was applied for hyperparameter optimization. Moreover, to address class imbalance, the dataset was resampled to achieve an approximately equal distribution of direct and indirect bandgap materials, followed by a 7:3 split into training and testing sets. Model performance was evaluated using confusion matrices and

Receiver Operating Characteristic (ROC) curves. For generating a ROC curve, the True Positive Rate (TPR) is plotted against the False Positive Rate (FPR) at multiple threshold values. TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN}$$
$$FPR = \frac{FP}{FP + TN} \quad (2)$$

where TP is True Positive, FP is False Positive, TN is True Negative, FN is False Negative [49].

The overall efficiency was quantified by the Area Under Curve (AUC) metric. An AUC of 0.5 corresponds to a purely random classifier, while an AUC of 1 indicates a perfect classifier [50]. SHAP analysis was further conducted to interpret the contribution of key features to the predictions.

Using the same predictive analysis framework, we developed several ensemble prediction tasks for both bandgap and formation energy of 2D MOs and perovskite oxides. For the bandgap prediction tasks, two datasets—DB_{Mb} and DB_{Pb}—were employed. Five ML algorithms, including XGBoost, RF, GBR, AdaBoost [51], and LightGBM [52], were used to model these datasets. Performance comparisons among these algorithms were conducted to select the most suitable model.

Feature correlation analysis was performed by randomly selecting 20 features (Fig. 3) to assess data quality. The correlation values for both 2D MOs and perovskite oxides were largely distributed within ±0.75, with DB_{Mb} exhibiting an even distribution concentrated around ±0.50, indicating reduced data redundancy and higher dataset quality for subsequent predictions.

After feature selection and cleaning, we found that although both datasets were used to predict bandgap values, the top 20 important features were not exactly the same. This indicates that, despite the models having the same task, ML is able to identify subtle differences for different input materials and establish prediction models that are better aligned with the characteristics of those materials.

A similar process was applied to the formation energy prediction task using feature matrices from the DB_{Mf} and DB_{Pf} datasets. As shown in Fig. 5, the correlation distributions for the 20 selected features were more uniform (ranging from −0.60 to 0.60), further confirming the robustness of our feature selection. These refined feature matrices were then used to train the same five ensemble algorithms, and their performance was compared in large-scale classification tasks.

For model training, each dataset was randomly split into training and test sets using a 7:3 ratio, followed by stratified 10-fold cross-validation on the training set to ensure statistical consistency. Hyperparameter optimization was performed using RandomizedSearchCV, with 20 iterations (n_iter = 20) and 10-fold cross-validation (cv = 10). For example, the XGBoost hyperparameter search space included max_depth ∈ {3, 5, 6, 10, 15, 20}, learning_rate ∈ {0.01, 0.1, 0.2, 0.3}, n_estimators ∈ {100, 300, 500, 1000}, and subsample ∈ {0.1, 0.5, 1.0}. Similar search grids were used for other tree-based models. These settings ensured reliable and well-optimized performance across all tasks.

3. Results and discussion

3.1. Classifying bandgap types of TMO monolayers

After constructing and optimizing the parameters of the predictive models for each task, we evaluated the performance of these models. For the bandgap property prediction of TMO monolayers, the TMO dataset (DB_{Td}), containing 572 bandgap entries, was divided into a training set and a testing set at a ratio of 7:3. To address class imbalance, resampling was performed to equalize the number of indirect and direct bandgap materials within the training set. Upon model training, ten-fold cross-validation was employed, and the classification accuracy was calculated based on the cross-validation results. The results are presented in Table 2.

Based on the data in Table 2, we can see that XGBoost, SVC, and DTC algorithms achieved an accuracy of 95.4 %, indicating their high effectiveness in classifying direct and indirect bandgap samples. In contrast, the accuracy of the Artificial Neural Network (ANN) was only 80 %. However, the primary goal of this study is not only to distinguish between direct and indirect bandgaps but also to accurately predict the category of each sample through feature engineering. Therefore, while the accuracy of ANN is lower, its role in classifying unknown samples should not be underestimated, as ANN has strong nonlinear fitting capabilities and can capture complex patterns in the data, which is especially important for classifying unknown samples. Furthermore, ANN can adapt to different datasets by adjusting its network structure and parameters, providing flexibility and scalability for classifying unknown samples [53].

Interpretability refers to whether the model's decision-making process and results can be understood and explained by humans, which is crucial for trust and transparency in scientific research and real-world applications [54]. Model complexity involves the number of parameters and structure of the algorithm. Overly complex models may lead to

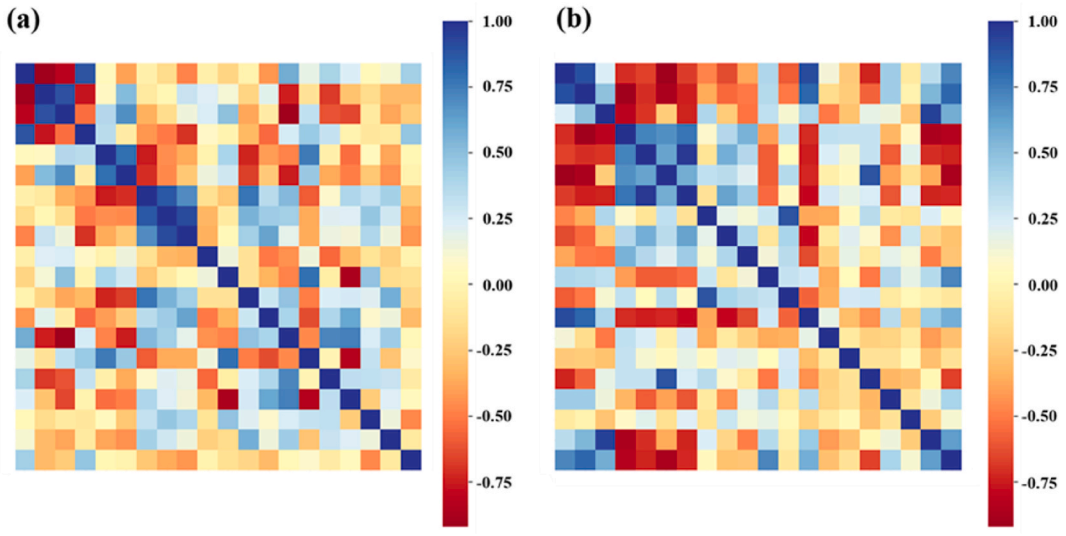

Fig. 3. Heat maps of feature correlations related to bandgap prediction for (a) DB_{Mb} and (b) DB_{Pb} datasets.

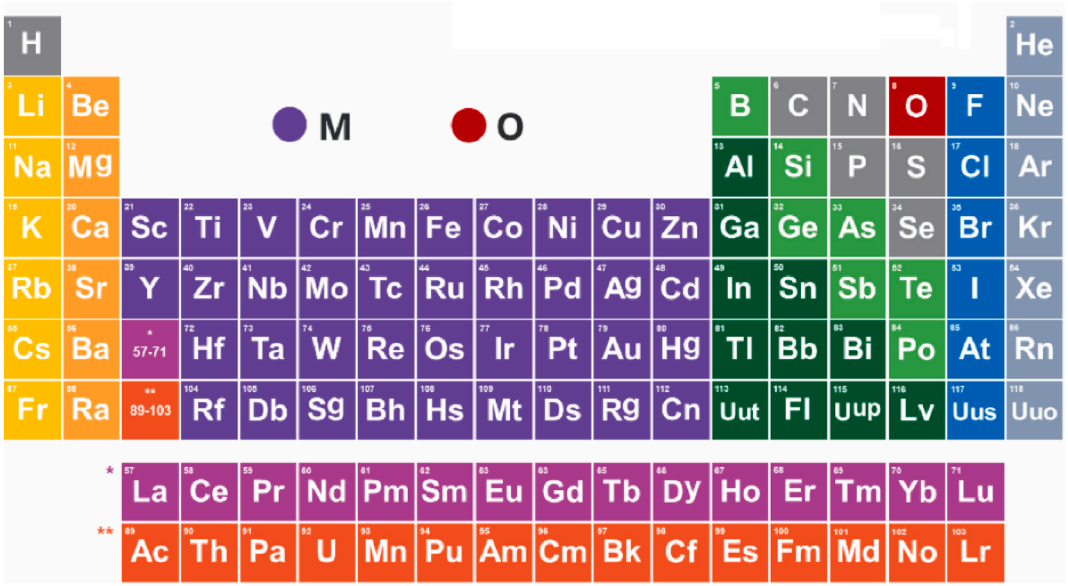

Fig. 4. Schematic diagram of atomic selection for TMO dataset.

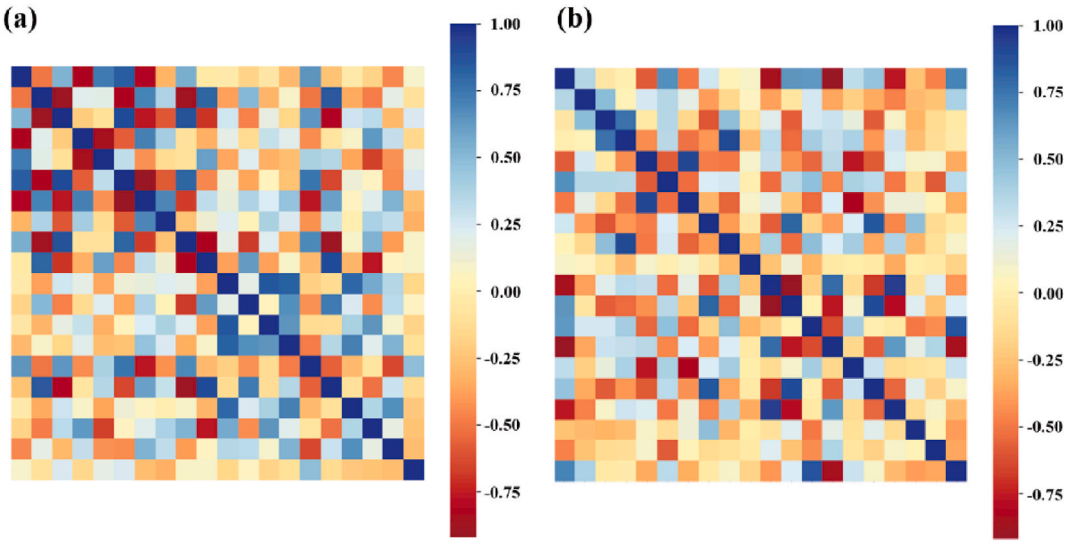

Fig. 5. Feature correlation heatmap for formation energy prediction. (a) Heat map of DB_{Mf} feature correlation and (b) Heat map of DB_{Pf} feature correlation.

Table 2
Average classification accuracy of 7 algorithms after tenfold cross-validation.

Algorithm	Average classification accuracy
XGBoost	95.4 %
RF	95.2 %
LR	89.0 %
SVC	95.4 %
ANN	80.0 %
DTC	95.4 %
GBC	92.3 %

overfitting, while overly simple models may fail to capture the complexity of the data, so a balance between model complexity and performance is necessary [55]. Computational efficiency refers to the computational resources and time required for model training and prediction, which is particularly important for large-scale datasets and real-time applications [56]. Therefore, when selecting the appropriate algorithm, these factors must be considered comprehensively to ensure that the model is both effective and practical.

The confusion matrix is crucial for evaluating algorithm performance, as it illustrates the effectiveness of these algorithms by providing detailed counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which can be used for a comprehensive analysis of each algorithm's ability to correctly identify direct bandgap materials. The accuracy of predicting direct bandgap materials is measured by the proportion of correctly identified samples among all known direct bandgap samples. Specifically, the diagonal elements in each matrix are key indicators for evaluation: the element in the top-left corner represents the number of correctly predicted direct bandgap materials, while the element in the bottom-right corner represents the number of correctly predicted indirect bandgap materials. Fig. 6 shows the confusion matrix results for the seven algorithms used in the bandgap property prediction of monolayer transition MOs. In Fig. 6(c), the ANN algorithm successfully identified 91 direct bandgap samples, which is the largest number among all algorithms. This result indicates that ANN shows a relative advantage in recognizing direct bandgap

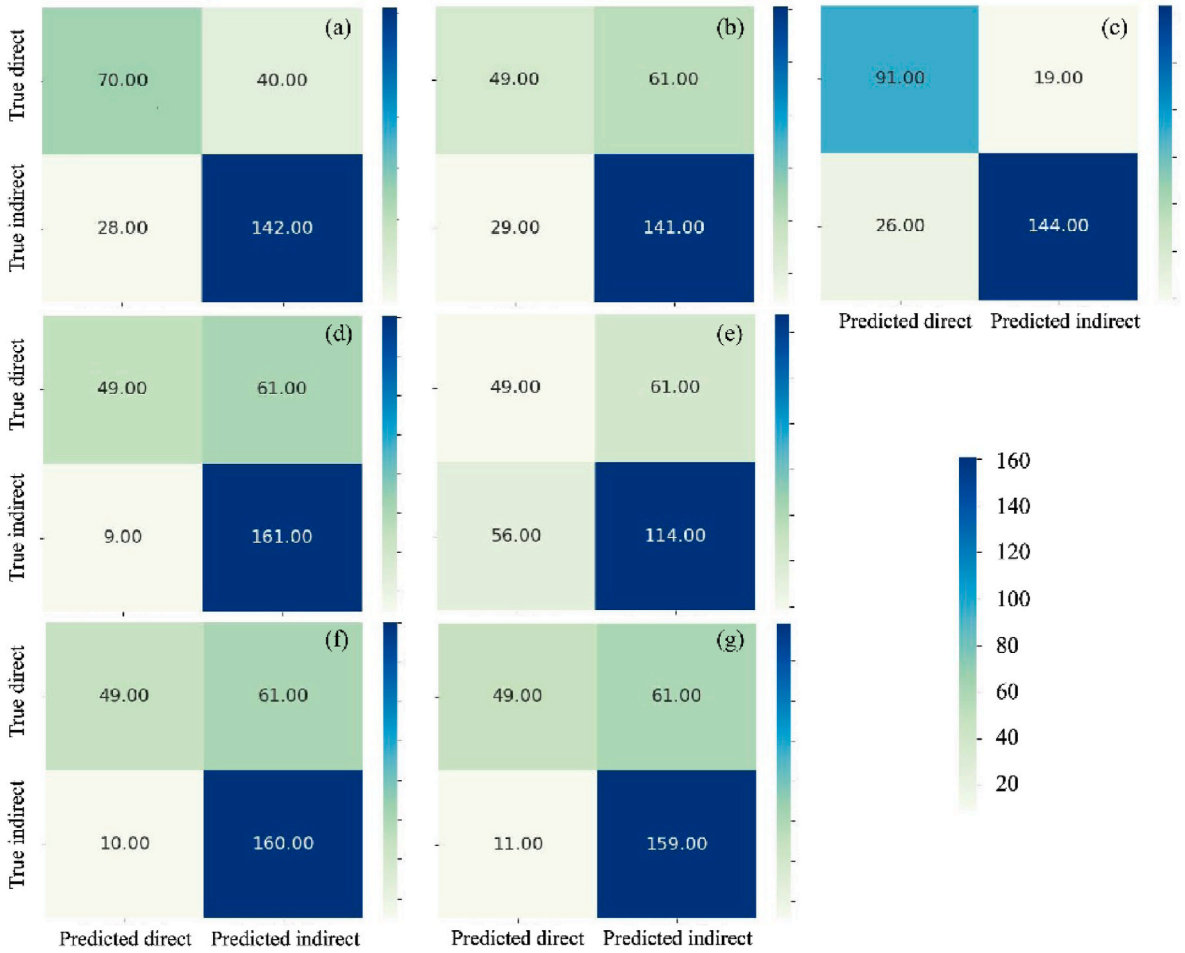

Fig. 6. Confusion matrix for seven classification algorithms. (a) LR, (b) DTC, (c) ANN, (d) XGBoost, (e) SVC, (f) GBC, and (g) RF.

materials, although the overall best performance is achieved by XGBoost and GBDT. Although the algorithms shown in Fig. 6(d), (f), and 6(g) also correctly predicted about 160 indirect bandgap samples, the core objective of this study is to improve the accuracy of predicting direct bandgap materials. Therefore, while the prediction results for indirect bandgap materials are also of some reference value, they are not the main focus of this study.

Table 3 provides a quantitative comparison of the algorithm performance in predicting the bandgap properties of monolayer transition MOs, revealing differences in prediction accuracy among the various algorithms. The analysis shows that the ANN model excels at predicting direct bandgap prediction task, achieving an accuracy rate of 0.88, the highest among all algorithms. Additionally, its F1 score of 0.80 further confirms its accuracy and reliability in predicting direct bandgap materials. Furthermore, Fig. 7 and Table 4 display the ROC curves, accuracy, and AUC values for the seven algorithms. Among these evaluation

Table 3
Predictive accuracy, recall and f1 scores for the test set.

Algorithm	Accuracy		Recall		F1-score	
	direct	indirect	direct	indirect	direct	indirect
XGBoost	0.73	0.84	0.95	0.45	0.82	0.58
RF	0.72	0.82	0.94	0.45	0.82	0.58
LR	0.78	0.71	0.84	0.64	0.81	0.67
SVC	0.65	0.47	0.67	0.45	0.66	0.46
ANN	0.78	0.88	0.85	0.83	0.86	0.80
DTC	0.70	0.63	0.83	0.45	0.76	0.52
GBC	0.72	0.83	0.94	0.45	0.82	0.58

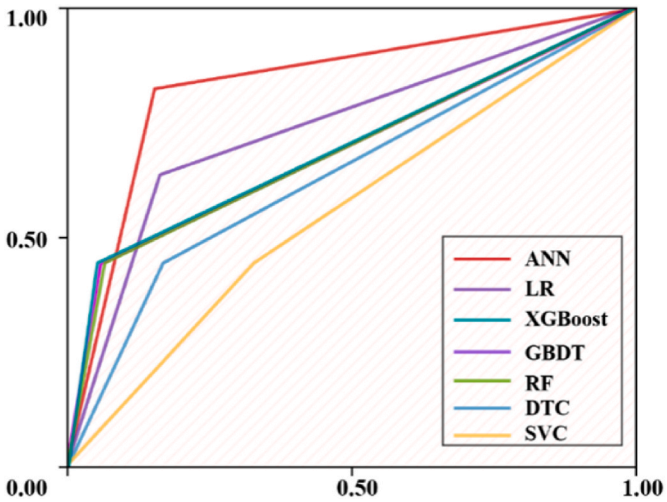

Fig. 7. ROC curves for seven classification algorithms used to classify bandgap types.

metrics, the ANN algorithm has the highest AUC value of 0.84, ranking first among all algorithms and being the only one with an AUC value exceeding 0.80. This indicates that the ANN algorithm has a significant performance advantage in predicting the bandgap characteristics of materials. Based on these results, it can be clearly concluded that, among the seven classifiers tested in this study, the ANN algorithm is the most

Table 4
Predictive accuracy and AUC values of the seven algorithms used to classify bandgap types.

Algorithm	Accuracy	AUC value
XGBoost	75 %	0.70
RF	74 %	0.69
LR	76 %	0.74
SVC	58 %	0.56
ANN	84 %	0.84
DTC	68 %	0.64
GBC	75 %	0.69

effective in predicting direct bandgap materials. At the same time, other algorithms such as XGBoost, LR, RF, and GBDT also demonstrated good performance in predicting direct bandgaps, with accuracy rates exceeding 0.80. In contrast, SVC and DTC showed suboptimal performance in terms of prediction accuracy, F1 score, and AUC value, and therefore, they are not the optimal algorithm choices in this study. These comparative results not only highlight the advantages of the ANN algorithm but also point out the potential and limitations of other algorithms in specific domains.

After a comprehensive analysis of the algorithm results, we can draw conclusions regarding algorithm performance: Ensemble learning algorithms, including RF, XGBoost, and GBDT, outperform individual ML algorithms. Ensemble learning methods enhance the generalization ability and stability of predictions by combining the advantages of multiple models, significantly improving prediction accuracy and overall performance. This approach effectively overcomes the limitations that may exist in a single model, leading to better prediction results. Among all the algorithms, the ANN algorithm demonstrates exceptional capability in predicting material bandgap characteristics. To further explore the prediction mechanism of the ANN model, we applied SHAP analysis. SHAP analysis provides insights into the model's decision-making process, revealing which features have the most significant impact on the model's prediction outcomes. This analysis not only enhances the interpretability of the model but also provides valuable information for further optimization. Fig. 8 displays the SHAP summary plot for the ANN model, listing the top 20 most important features for identifying the type of bandgap.

Through SHAP analysis, we can identify and rank the key features that influence the model's predictions. The most influential features, ranked by impact, include range number, average atomic mass, and the elemental contents of Co, Zr, and Ti. Other important features also include the average valence electron count in the f-orbital, range and maximum change in nearest-neighbor distance, Se content, and group averages. Fig. 8 illustrates the impact of various features on the prediction of four-dimensional samples. (The names of relevant features and their corresponding physical meanings are in Supplemental Material Table S1.) Each point represents a data instance, with the vertical axis representing the feature's importance, where higher values indicate greater importance; the horizontal axis displays the SHAP values, reflecting the feature's impact on the model's output. Positive values on the horizontal axis correspond to class 1, while negative values correspond to class 0. The color of the points represents the feature value, with red indicating high values and blue indicating low values; the size of the points is proportional to the probability of the feature belonging to the respective class, with larger points indicating a higher probability. Through analyzing the summary plot, we can infer the following:

(1) **Average number of s-valence electrons**: This is a key determinant of TMO direct bandgap characteristics. The lower the average number of s-valence electrons in the compound, the more likely it is to exhibit a direct bandgap. This average is calculated by dividing the total number of s-valence electrons in the unit by the total number of atoms. Transition metals typically have fewer s-valence electrons, making them more likely to display direct bandgap characteristics;

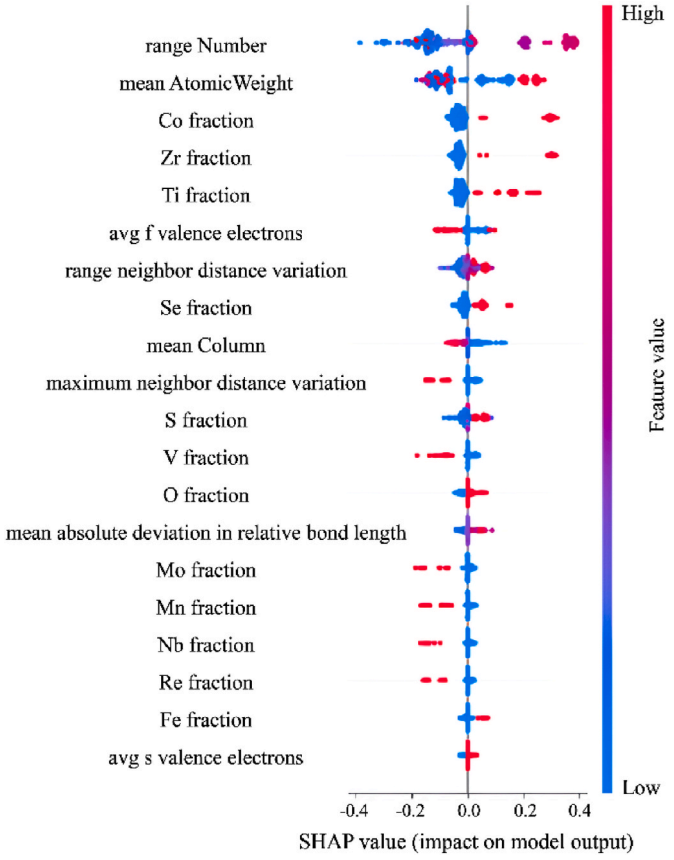

Fig. 8. SHAP summary plot illustrating feature contributions in the ANN model used to classify bandgap types.

(2) **Oxygen (O) content**: The elemental composition significantly affects the bandgap characteristics. The higher the oxygen content, the larger the direct bandgap;
(3) **Average atomic mass**: The bandgap characteristics are influenced by the compound's average atomic mass. After reaching a certain threshold, the larger the average atomic mass, the higher the probability of forming a direct bandgap;
(4) **Atomic number range**: The broader the atomic number range of the elements in the compound, the more likely it is to exhibit direct bandgap characteristics;
(5) **Content of Co, Zr, and Ti**: The concentrations of Co (Cobalt), Zr (Zirconium), and Ti (Titanium) are crucial. Higher concentrations of these elements favor the formation of direct bandgaps, while lower concentrations tend to promote indirect bandgaps.

These findings are consistent with recent SHAP-based materials studies, which emphasized that descriptors such as electronegativity differences, ionic radii, and local coordination environments play a decisive role in governing bandgap and stability, thereby supporting the physical relevance of the identified descriptors in this work [57,58].

3.2. *Predicting bandgap values of 2D TMOs*

We predicted the bandgap of 2D TMOs and analyzed the performance of ensemble algorithms. A variety of error and correlation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), were used for quantitative analysis. Table 5 displays the prediction results of five ensemble algorithms on the bandgap datasets of 2D MOs

Table 5
Evaluation metrics of five integrated algorithms for bandgap prediction on two test sets.

Algorithm	Assess							
	DB_{Mb}				DB_{Pb}			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
XGBoost	0.11	0.09	0.30	0.95	0.30	0.33	0.58	0.87
RF	0.32	0.25	0.50	0.85	0.41	0.41	0.64	0.84
GBR	0.11	0.09	0.29	0.95	0.30	0.33	0.57	0.87
AdaBoost	0.79	1.05	1.02	0.39	0.93	1.26	1.12	0.49
LightGBM	0.50	0.48	0.69	0.61	0.38	0.37	0.61	0.76

and perovskite oxides. Overall, for both datasets, the Gradient Boosting Regressor (GBR) model was consistently the best choice, with the prediction error of the XGBoost model is almost identical to that of the GBR model. In contrast, the AdaBoost model performed the worst in this task. These evaluation metrics reveal the differences between algorithms in terms of prediction accuracy and stability, providing a basis for selecting the best model.

By comparing the prediction results of the two datasets in Table 5, we gain a deeper understanding of the factors influencing algorithm performance. It is evident that larger datasets typically yield better prediction results. We draw the following conclusions from the analysis: First, from a functional perspective, XGBoost and GBR models outperform Random Forest (RF) and AdaBoost models, which is closely related to their algorithmic principles. The residuals in GBR and the regularization in XGBoost ensure superior prediction performance. In contrast, AdaBoost is more sensitive to noise and anomalies, while RF's parallel ensemble method may lead to suboptimal performance in this task. Lastly, ML is fundamentally a data-driven approach. DB_{Mb} not only has a larger dataset but also exhibits a more uniform distribution of feature correlations, representing better data quality, which results in better prediction performance.

To further elucidate the decision-making process of the GBR model, we analyzed the importance of its features using SHAP values. Fig. 9 presents the correlation heatmap between the important features of the GBR model and their SHAP values. (The shortened names of relevant features and their corresponding physical meanings are provided in Table S2 of the supplementary materials.) The color blocks in the figure represent the magnitude of the SHAP values, with the top f(x) indicating the cumulative SHAP value of all features, the gray dashed line representing the SHAP baseline, and each row corresponding to the SHAP values of different samples for the features. It is clear from the figure that the overall SHAP value of the model is primarily contributed by a few key features, particularly the portions in Fig. 9(a) and (b) where f(x) is below the baseline. This indicates that the ML algorithm assigns different weights to features, with more important features receiving higher weights, visually demonstrating the core idea of ML algorithms.

Fig. 10 displays the top 10 important features and their contributions during the bandgap prediction of both 2D MOs and perovskite oxides. The x-axis represents the output value, i.e., the sign and magnitude of the bandgap, with the color bar indicating the size of the feature values. The circles represent individual samples. Comparing Fig. 10(a) and (b), it is evident that, although both models use GBR, the ranking of feature importance is entirely different, and the contribution of the same feature varies significantly. The top 10 most important features in the two models are not exactly the same, and their order differs. This suggests that although perovskite oxides are a type of metal oxide, the formation of their bandgap is influenced by unique features, which may be the fundamental reason for their diverse photoelectric properties. Such consistency between SHAP-identified features and established physicochemical mechanisms highlights the robustness of the interpretability framework [57,58].

Common features between the two models include the average number of d orbital valence electrons (Dve_a), the percentage of s orbital valence electrons (Sve_f), and the average number of unfilled p orbitals (Puf_a). The first two features rank within the top five in both models, indicating that the valence electrons in the d and s orbitals have a significant impact on the formation of bandgaps in metal oxide systems. Notably, while the average number of unfilled p orbitals is important for the bandgap formation of both materials, its contribution is entirely different. A high value of this feature tends to widen the bandgap in MOs, while it pushes the bandgap value of perovskite oxides closer to zero. This reveals the unique mechanisms of bandgap formation in different metal oxide systems, providing new insights for material design.

Finally, we analyze the impact of these features on the bandgap of MOs and perovskite oxides separately to draw more specific conclusions. From Fig. 10, we can conclude that for MOs, the average number of d orbital valence electrons is negatively correlated with the bandgap

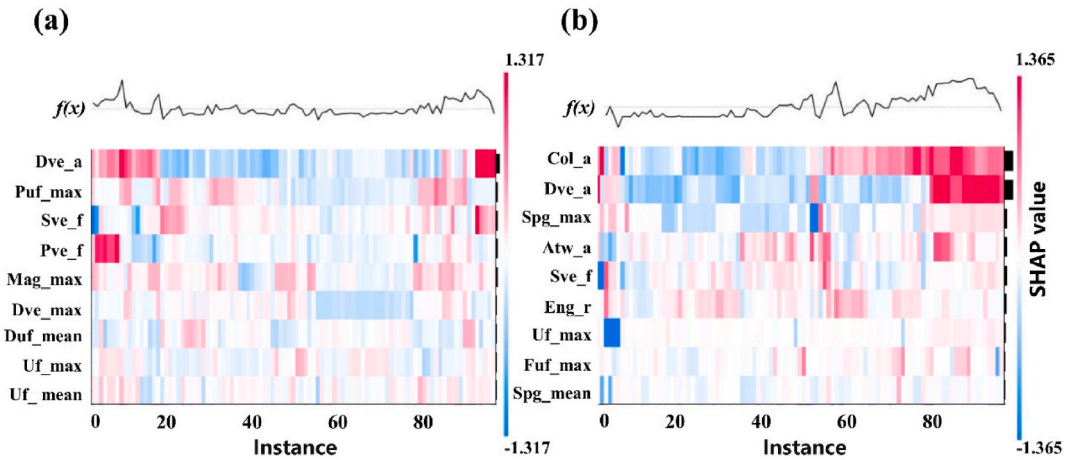

Fig. 9. Heat map of correlation between features and SHAP values for bandgap prediction on two test sets. (a) DB_{Mb} and (b) DB_{Pb}.

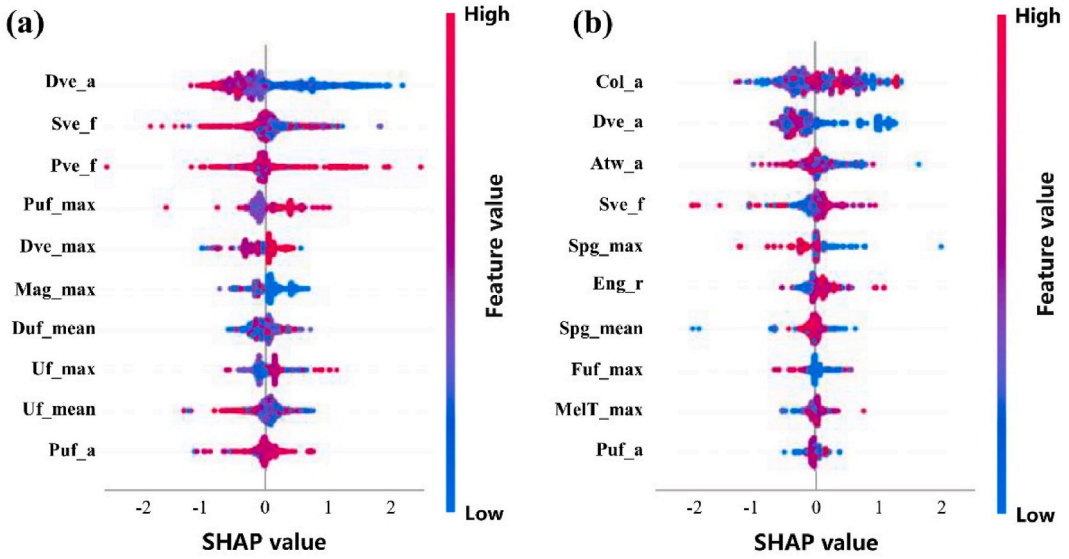

Fig. 10. SHAP summary plot for the GBR model for bandgap prediction on two test sets. (a) DB_{Mb} and (b) DB_{Pb}.

value; adjusting the percentage of s orbital valence electrons can help achieve the desired bandgap for MOs. For perovskite oxides, the average number of d orbital valence electrons, average atomic weight (Atw_a), and the maximum value of space group (Spg_max) are all negatively correlated with the bandgap value, while the electronegativity range is positively correlated. Moreover, both excessively high and low percentages of s orbital valence electrons can lead to a bandgap of 0 or negative values. From this, we draw the following conclusions and material design recommendations:

(1) **Reliability of Conclusions**: Both the SHAP analysis for 2D MOs and perovskite oxides show that the higher the average number of d orbital valence electrons, the narrower the material's bandgap. This conclusion is consistent with the known impact of d orbital valence electrons on the material's bandgap;
(2) **Factors Affecting the Bandgap**: The primary factors affecting the bandgap are related to electronic properties, such as the number of valence electrons in the d and s orbitals, and the number of unfilled p, d, and f orbitals. These factors have little relation to the elemental composition of the compound. The bandgap is closely related to electronic transition behaviors, and the SHAP analysis shows that the bandgap interacts with electrons, with the arrangement of valence electrons in different orbitals influencing the formation and width of the bandgap. The width of the bandgap in turn affects electronic transition behaviors. By combining ML with model interpretation methods, we can infer relevant physical properties from the data, further proving the adaptability of ML methods to material science;
(3) **Material Design Perspective**: We can closely integrate ML results with material design to more efficiently design ideal materials that meet specific needs. For example, to design a metal oxide with a bandgap greater than 0, one can select atomic combinations where the percentage of s orbital valence electrons is relatively low. For perovskite oxides, one can choose atomic combinations with a larger difference in electronegativity to achieve a wider bandgap. In summary, based on the feature-output relationships provided by SHAP analysis, we can design ideal oxide materials.

3.3. *Predicting formation energies of 2D MOs*

After thoroughly understanding the bandgap prediction of 2D MOs, we explore the prediction of their formation energy. Table 6 presents the performance evaluation metrics of five ensemble algorithms in predicting the formation energy of 2D MOs using datasets DB_{Mf} and DB_{Pf}.

Based on the evaluation metrics in Table 6, we can further analyze the performance of different ensemble algorithms in formation energy prediction. Table 6 provides a comprehensive comparison of the performance of various ensemble algorithms in predicting the formation energy of 2D MOs and perovskite oxides. The analysis shows that among all the evaluated algorithms, the GBR model performs the best across all evaluation metrics, whether in the 2D metal oxide or perovskite oxide datasets. Following closely behind is the XGBoost model, with a very small performance gap between the two, indicating that both algorithms effectively handle formation energy prediction tasks. Furthermore, we observe that the LightGBM algorithm exhibits high accuracy in formation energy prediction, especially when the data volume is the same, outperforming regression predictions for bandgaps. This observation implies that, under specific conditions, LightGBM may offer better predictive accuracy. Fig. 11 provides the SHAP value correlation heatmap and SHAP summary plot for the GBR model's formation energy prediction, which revealing the importance and contribution of each feature in the model's predictions.

Figs. 11(a) and 10(b) show that the model's f(x) values are more concentrated on the contribution of a few features. In Fig. 11(a), the lowest values are almost entirely attributed to the electronegativity range, while in Fig. 11(b), the higher values also mainly depend on the electronegativity range. This indicates that the formation energy is closely related to the material's electronegativity. For Fig. 11(c) and (d), the formation energy is influenced not only by valence electrons but also by electronegativity, melting point, elemental content, and atomic radius (the latter ranking 11th in both models). This suggests that the formation energy of materials is influenced by multiple factors, while the bandgap is more concentrated on electronic properties. The electronegativity range is negatively correlated with the formation energy of both materials, and a lower average group number is more likely to result in a formation energy less than 0. Lastly, the formation energy of MOs is negatively correlated with the oxygen content, while the average number of d-orbital valence electrons is positively correlated. For perovskite oxides, the formation energy is negatively influenced by the minimum total number of valence electrons and the average number of s-orbital valence electrons.

Based on the above analysis, we can draw the following conclusions about material formation energy:

Table 6
Evaluation metrics of five integrated algorithms for formation energy prediction on two test sets.

Algorithm	Assess							
	DB_{Mf}				DB_{Pf}			
	MAE	MSE	RMSE	R^2	MAE	MSE	RMSE	R^2
XGBoost	0.05	0.03	0.18	0.97	0.21	0.12	0.34	0.88
RF	0.12	0.06	0.25	0.94	0.30	0.16	0.40	0.83
GBR	0.05	0.03	0.18	0.97	0.21	0.11	0.34	0.88
AdaBoost	0.40	0.27	0.52	0.73	0.50	0.37	0.61	0.62
LightGBM	0.21	0.11	0.33	0.89	0.34	0.19	0.44	0.80

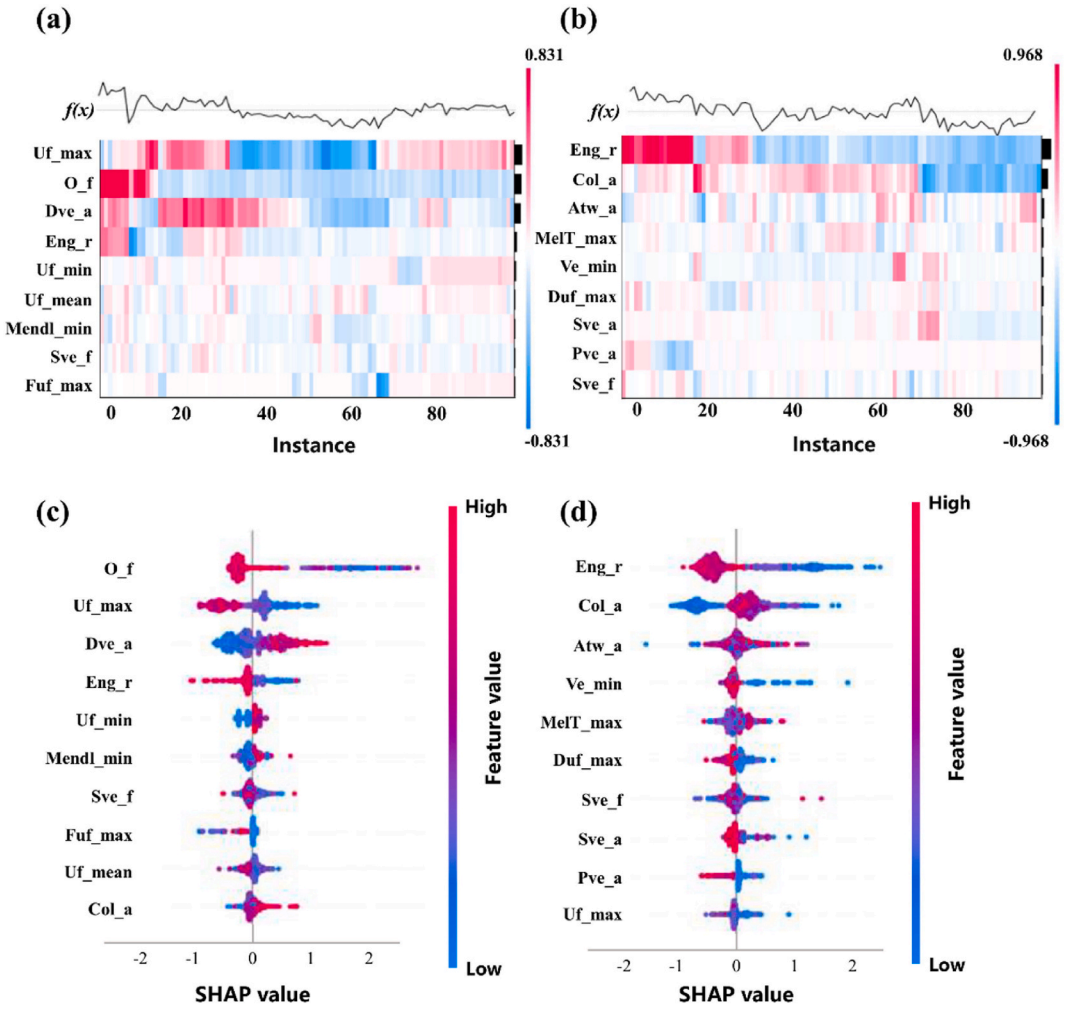

Fig. 11. GBR model for formation energy prediction. (a) Heat map of DB_{Mf} feature correlation with SHAP values, (b) heat map of DB_{Pf} feature correlation with SHAP values, (c) summarized graph of GBR model SHAP for DB_{Mf}, and (d) Summarized graph of GBR model SHAP for DB_{Pf}.

(1) **Electronegativity and Formation Energy**: Both from model training and SHAP analysis, electronegativity has a significant impact on formation energy. The greater the electronegativity range of a compound, the lower its formation energy. This association prompts further thought: electronegativity influences the formation of chemical bonds—larger differences tend to form ionic bonds, while smaller differences tend to form covalent bonds. Therefore, under normal conditions, compounds with ionic bonds are likely more stable than those with covalent bonds. However, this is just one influencing factor;

(2) **Material Design and Formation Energy**: In material research, we aim to identify compounds that are easy to form and highly stable. According to this study, the stability of MOs can be enhanced by increasing the oxygen content. This trend is pronounced in the 2D oxide dataset due to variable oxygen stoichiometry, whereas for ABO_3 perovskites, oxygen content is constant and does not affect the feature ranking. Similarly, for designing more stable perovskite oxides, choosing atoms with high valence electron counts and s-orbital electron counts can contribute to greater stability.

3.4. Classifying bandgap types of perovskite oxides

By predicting the bandgap and formation energy of 2D MOs and perovskite oxides, we have successfully demonstrated the high reliability and accuracy of ensemble algorithms in regression tasks

involving large material samples. However, the performance of ensemble algorithms in large-scale classification tasks remains uncertain. Therefore, we applied seven classification algorithms—XGBoost, RF, LR, SVC, MLPClassifier, DTC, and GBC—to classify a larger dataset and evaluate the performance of ensemble algorithms using the perovskite oxide bandgap dataset (DB_{Pd}). Table 7 presents the accuracy and AUC values of the seven algorithms in predicting direct bandgap materials, while Fig. 12 summarizes the ROC curves for the seven algorithms.

Based on Table 7, we can see that, unlike the bandgap prediction for transition MOs, the ANN performs poorly for perovskite oxides. However, the SVM stands out in this classification task, largely due to the increased dataset size and a greater number of training samples, allowing the SVC to demonstrate its strength in handling high-dimensional data. The results indicate that ensemble algorithms exhibit good predictive performance on larger datasets. Compared to the classification study of the bandgap properties of single-layer transition MOs, the performance of all ensemble algorithms has significantly improved. Therefore, for future studies involving a larger number of 2D material properties, as they are computationally efficient and consistently deliver robust performance.

To further analyze the features influencing the direct bandgap of perovskite oxides, we performed SHAP analysis on the Random Forest model. Fig. 13 presents the contribution of the RF model to the prediction results, listing the top 20 important features influencing the bandgap properties of perovskite oxides. When combined with the results of the perovskite oxide bandgap prediction, it is clear that factors such as the range of compound electronegativity, average atomic weight, d-orbital valence electron count, and the percentage of s-orbital valence electrons significantly impact the bandgap width and properties. Therefore, by carefully selecting or tuning these features, we can design perovskite oxide materials with ideal bandgap properties and widths.

To further validate the interpretability and robustness of the SHAP-based feature analysis, we conducted both multivariable regression and ablation experiments using the DB_{Mb} and DB_{Pb} datasets. First, we evaluated the model performance using only the top five SHAP-ranked features, and compared the results against models trained with all features. As summarized in Table S3, the model trained on DB_{Mb} using all features achieved an MAE of 0.11 and R^2 of 0.95, while the top-5 feature model exhibited a higher MAE of 0.19 and a lower R^2 of 0.90. This drop in performance confirms that the excluded features still contribute to predictive accuracy, although the top features account for the majority of information. For DB_{Pb}, the model performance remained almost unchanged between the full-feature and top-5 settings, indicating a more compact feature space for that dataset.

Furthermore, feature ablation experiments were conducted by removing the single most important feature (Top-1) from the input. For DB_{Mb}, removing the top feature slightly increased the MAE to 0.12 and reduced R^2 to 0.95. The relatively small performance degradation suggests that while the top-ranked feature carries significant weight, the model can partially compensate using the remaining correlated features, consistent with moderate feature collinearity observed in the correlation matrix (Fig. 4). For DB_{Pb}, the removal of the top feature had negligible effect, again supporting the robustness of the model. These experiments verify that the SHAP-derived features are not only informative but also physically meaningful and robust to feature perturbations.

Table 7
AUC values and classification accuracies of the seven algorithms after tenfold cross-validation.

Algorithmic models	AUC value	Class 1 accuracy
XGBoost	0.94	0.88
RF	0.94	0.88
LR	0.65	0.62
SVC	0.94	0.89
ANN	0.63	0.64
DTC	0.89	0.82
GBC	0.94	0.88

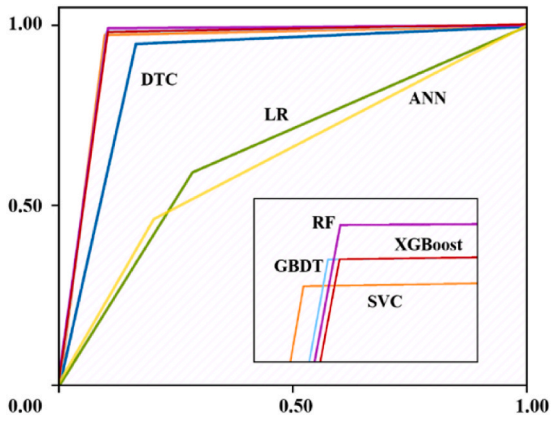

Fig. 12. ROC curves for seven classification algorithms used to classify bandgap types.

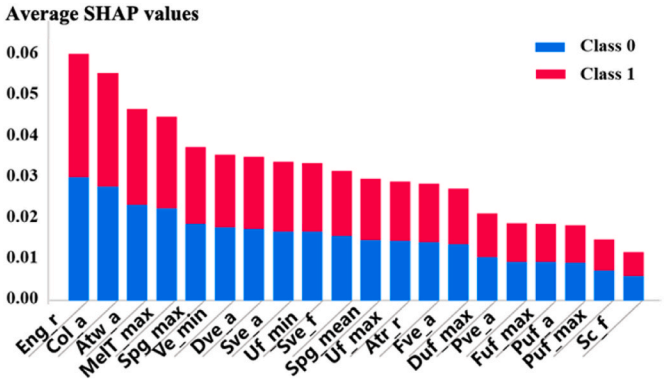

Fig. 13. SHAP summary plot illustrating feature contributions in the RF model used to classify bandgap types.

3.5. Comprehensive prediction workflow construction and validation

In the prediction of the bandgap and formation energy of 2D MOs, we generated features based on elemental characterization methods and analyzed both metal oxide and perovskite oxide datasets. After conducting feature correlation analysis and importance ranking, we removed highly correlated features that were not significant for model output, thereby establishing the final input feature matrix. Subsequently, we used various algorithms to perform regression predictions for the properties of both materials, ultimately achieving regression models with an accuracy of approximately 0.9.

After confirming the model's accuracy and stability, we applied the process to predict the bandgap and formation energy of over 20,000 unknown perovskite oxide materials. To ensure the accuracy of predictions for the unknown dataset, we used the same feature generation methods, removed features with correlations greater than 0.8, and retained the same feature values based on the previously established feature matrix. Features not present in the matrix were deleted to maintain consistency in the model's input matrix and avoid errors. The processed feature dataset was then input into the pre-trained regression prediction model, yielding predicted values for over 20,000 data points, which were saved for subsequent work.

Given that the bandgap prediction model for 2D TMOs achieves an accuracy close to 0.9, we randomly selected two perovskite oxide samples for simulation calculations to verify the prediction accuracy for

unknown data. As shown in Fig. 14, the chemical structures, simulated band structure diagrams, and performance parameters obtained from both ML and simulation calculations are displayed. The results indicate that the predicted bandgaps and formation energies are in close agreement with the simulation values, with prediction errors within the ideal range for this study, making them suitable for future research.

Based on the above research, we have developed an integrated prediction process, consolidating the findings into a unified framework aimed at establishing a reliable and accurate ML regression prediction workflow to advance material property prediction and the discovery of new materials. This process encompasses key steps such as data processing, feature engineering, model selection, and validation.

In constructing datasets for unknown materials, the most common approach is based on chemical element combinations, which can generate large datasets but often lack compound structural features. In the case of bandgap prediction for single-layer TMOs, simple ANN algorithms are more suitable for datasets with limited size and rich structural information, but they perform poorly in predicting energy properties for 2D TMOs, with an accuracy of approximately 0.6, making them unsuitable for this study. Further analysis reveals that for datasets containing more than 1,000 data points, ensemble algorithms, especially Gradient Boosting Decision Trees and XGBoost, perform excellently in both regression and classification tasks. These algorithms consistently ranked in the top three across five different prediction tasks involving materials, scales, types, and targets, with comparable accuracy and satisfactory prediction performance. Therefore, we selected GBDT and XGBoost as candidate algorithms for building more versatile and high-performance ML prediction models.

Additionally, the prediction accuracy of ML is influenced by model performance and the quality of the input matrix data. The feature engineering process for predicting the energy properties of 2D TMOs is reliable, universal, and simple to screen, requiring only the chemical formula of the compound. Therefore, it was adopted as the feature engineering module in the ML workflow. After determining the feature engineering method and algorithm model, we completed the construction of the entire workflow (as shown in Fig. 15). By using this method, a dataset containing the chemical formulas and target properties of 2D metal oxide materials can generate an applicable ML model, which is expected to provide high-precision predictions. This workflow has been validated multiple times, demonstrating high accuracy and reliability. It can serve as a reliable tool for material screening, offering effective guidance for material research and design, and is expected to provide important references for material synthesis and optimization.

4. Conclusion

In this study, we developed a high-precision, low-cost ML framework to predict the bandgap and formation energy of 2D TMOs, aiming to improve materials screening and design. By integrating data from multiple databases, seven independent datasets were constructed, and feature engineering was conducted using Pearson correlation and feature importance analysis. Among the seven models evaluated across six prediction tasks, XGBoost and GBDT consistently outperformed others. XGBoost achieved 95.4 % accuracy in classification, while both XGBoost and GBR reached R^2 values close to 0.9 in regression.

A complete ML workflow including data preprocessing, feature selection, model training, and validation, was established and applied to predict energy properties of over 20,000 previously unstudied perovskite oxides. Random sampling and DFT validation confirmed the reliability of predictions, indicating that the framework can effectively reduce reliance on costly calculations and blind experiments.

To interpret model predictions, SHAP analysis was used to identify key descriptors. For bandgap, the average number of d-orbital valence electrons showed a negative correlation, while the proportion of s-orbital electrons played a regulatory role. Other influential features included unfilled p orbitals, oxygen content (relevant only for 2D oxides), atomic mass, atomic number range, and concentrations of Co, Zr, and Ti. For formation energy, wider electronegativity ranges and higher oxygen content in 2D oxides were associated with greater stability, while d-orbital electrons showed a positive correlation.

In summary, the proposed ML framework offers accurate and interpretable predictions of key energy properties in 2D TMOs, providing a practical tool for materials discovery and a foundation for future extension to other functional systems.

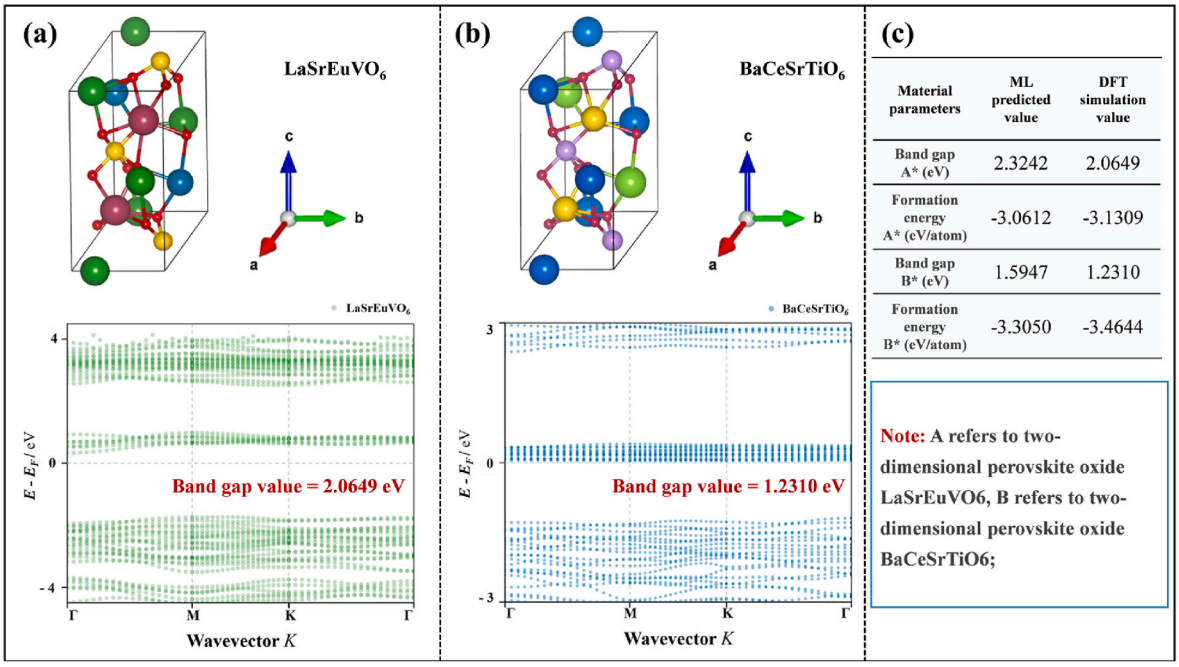

Fig. 14. ML-based prediction of bandgap for unknown perovskite oxide materials: Schematic of Results Validation. (a) Schematic representation of the chemical structure and band structure of $LaSrEuVO_6$, (b) Schematic representation of the chemical structure and band structure of $BaCeSrTiO_6$, and (c) Comparison of Predicted and Simulated Energy Properties for Two Materials.

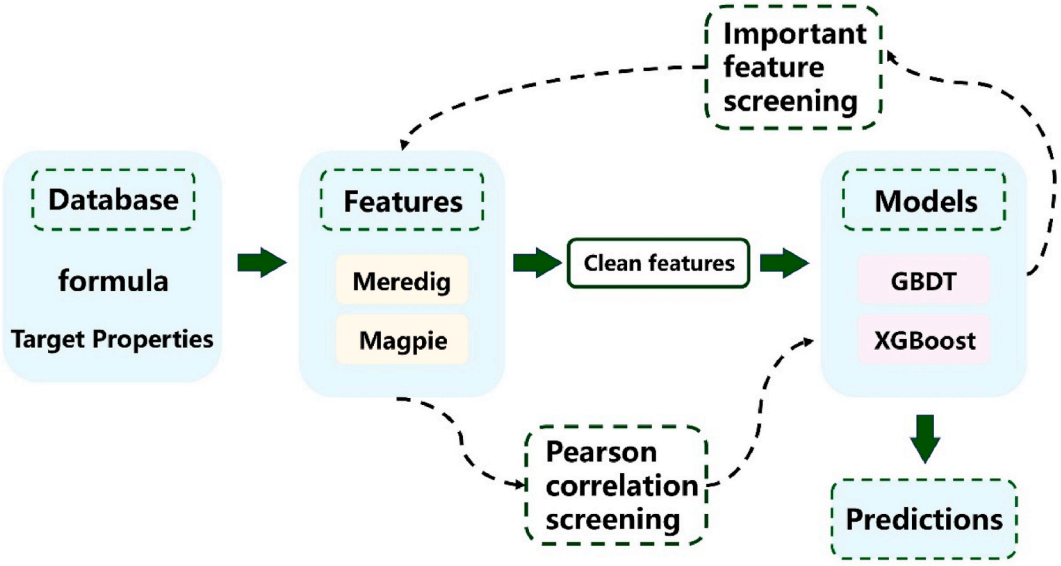

Fig. 15. Schematic diagram of the ML process for predicting the properties of 2D materials.

CRediT authorship contribution statement

Wen Yao: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation. **Wanli Jia:** Writing – review & editing, Supervision, Resources, Project administration, Investigation. **Ruofan Shen:** Visualization, Methodology, Data curation. **Jiayao Wang:** Writing – original draft, Validation, Data curation. **Lin Zhang:** Resources, Methodology, Investigation, Formal analysis. **Xinmei Wang:** Writing – review & editing, Software, Methodology, Data curation.

Statement of novelty

This study pioneers a unified machine learning framework for predicting bandgap and formation energy in diverse 2D metal oxides (MOs) and perovskites. Unlike prior works limited to narrow material classes or small datasets, we integrate multi-algorithm optimization (7 ML models), cross-scale validation (5,642 to 21,887 samples), and interpretable SHAP analysis to uncover key descriptors (e.g., d/s-orbital electrons, oxygen content). Our workflow uniquely bridges feature engineering, ensemble learning, and DFT verification, achieving state-of-the-art accuracy (95.4 %, $R^2 \approx 0.90$) while reducing computational costs. This approach advances rational design of tunable 2D oxides, addressing gaps in scalability and interpretability for high-throughput materials discovery.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grants No. 62474140), the Education Department of Shaanxi Provincial government (Grant No. 22JP058) (Grant No. 62174136) and the Ningxia Hui Autonomous Region Natural Science Foundation (Grant No. 2024AAC03158).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.physb.2025.417821.

Data availability

I shared the link to my data prior to the reference section of the paper.

References

[1] L. Tao, Y. Zhao, H. Sun, Metallenes for electrocatalytic nitrate reduction, Mater. Today Chem. 42 (2024), 102418-102418.
[2] K. Partha, C.G. Chinmayee, M.P. Lata, et al., Emerging 2D metal oxides and their applications, Mater. Today (2021) (prepublish).
[3] H.A. W, W. Hai-Chen, M.M.A. L, Two-dimensional binary metal-oxide quasicrystal approximants, 2D Mater. 8 (4) (2021).
[4] Y. Yang, X. Yin, Y. Ma, et al., Nonlinear optical properties and applications of 2D metal oxides, J. Nonlinear Opt. Phys. Mater. 33 (2) (2023).
[5] K. Kalantar-Zadeh, J.Z. Ou, T. Daeneke, et al., Two dimensional and layered transition metal oxides, Appl. Mater. Today 5 (2016) 73–89.
[6] P.K. Pradhan, N.K. Mohanty, G.K. Mishra, et al., 1 - Concepts and recent advancements in perovskite metal oxides[M]//MOHARANA S, in: T. BADAPANDA, S.K. SATPATHY, et al. (Eds.), Perovskite Metal Oxides, Elsevier, 2023, pp. 3–22.
[7] V. Acharya, A. Sharma, N.K. Chourasia, et al., Solution-processed Pb0.8Ba0.2ZrO3 as a gate dielectric for low-voltage metal-oxide thin-film transistor, Emergent Materials 3 (1) (2020) 57–62, https://doi.org/10.1007/s42247-019-00065-1.
[8] Grid..5, Clarendon Laboratory, Parks Road, Ox1 3pu, Oxford, Uk, Grid..5, Clarendon laboratory, parks road, Ox1 3pu, Oxford, Uk. Metal-halide perovskites for photovoltaic and light-emitting devices, Nat. Nanotechnol. 10 (5) (2015) 391–402.
[9] R. Waser, R. Dittmann, G. Staikov, et al., Redox-based resistive switching memories - nanoionic mechanisms, prospects, and challenges, Adv. Mater. 21 (25–26) (2009) 2632, https://doi.org/10.1002/adma.200900375.
[10] J. Suntivich, K.J. May, H.A. Gasteiger, et al., A perovskite oxide optimized for oxygen evolution catalysis from molecular orbital principles, Science 334 (6061) (2011) 1383–1385.
[11] S. Wang, X. Wu, The mechanical performance prediction of steel materials based on random forest, Frontiers in Computing and Intelligent Systems (2023).
[12] S.P. G, M.N. Mattur, N. Nagappan, et al., Prediction of nature of band gap of perovskite oxides (ABO3) using a machine learning approach, Journal of Materiomics 8 (5) (2022) 937–948, https://doi.org/10.1016/j.jmat.2022.04.006.
[13] G.M. Kanyolo, T. Masese, An idealised approach of geometry and topology to the diffusion of cations in honeycomb layered oxide frameworks, Sci. Rep. 10 (1) (2020) 13284, https://doi.org/10.1038/s41598-020-70019-9.
[14] S. Rath, G. Sudha Priyanga, N. Nagappan, et al., Discovery of direct band gap perovskites for light harvesting by using machine learning, Comput. Mater. Sci. 210 (2022), https://doi.org/10.1016/j.commatsci.2022.111476.

[15] A. Jain, K.A. Persson, G. Ceder, Research update: the materials genome initiative: data sharing and the impact of collaborative ab initio databases, APL Mater. 4 (5) (2016), 053102-053102-053114.
[16] B. Ben, W. Logan, S. Marcus, et al., A data ecosystem to support machine learning in materials science, MRS Commun. 9 (4) (2019) 1125–1133.
[17] S. Matthias, A. Martin, A. Martin, et al., FAIR data enabling new horizons for materials research, Nature 604 (7907) (2022) 635–642.
[18] F. De Angelis, The impact of machine learning in energy materials research: the case of halide perovskites, ACS Energy Lett. 8 (2) (2023) 1270–1272, https://doi.org/10.1021/acsenergylett.3c00182.
[19] Massachusetts Institute of Technology C, Ma 02139, USA, I.O.M. Research, A. S. Engineering, Singapore Innovis, et al., Accelerating materials development via automation, machine learning, and high-performance computing, Joule 2 (8) (2018) 1410–1420.
[20] A. Jain, Machine learning in materials research: developments over the last decade and challenges for the future, Curr. Opin. Solid State Mater. Sci. 33 (2024), 101189-101189.
[21] C.S. Sin, N.Y. Sheng, W.H. Qiong, et al., Advances of machine learning in materials science: ideas and techniques, Frontiers of Physics 19 (1) (2023) 13501.
[22] C. Ezeakunne, B. Lamichhane, S. Kattel, Integrating density functional theory with machine learning for enhanced band gap prediction in metal oxides, Phys. Chem. Chem. Phys. : Phys. Chem. Chem. Phys. (2025).
[23] W. Zhang, J. Guo, X. Lv, et al., Combined machine learning and high-throughput calculations predict heyd-scuseria-ernzerhof band gap of 2D materials and potential MoSi2N4 heterostructures, J. Phys. Chem. Lett. (2024) 5413–5419.
[24] H. Juan, X. Xiaomin, L. Meisheng, et al., Recent advances in perovskite oxides for non-enzymatic electrochemical sensors: a review, Anal. Chim. Acta 1251 (2023), 341007-341007.
[25] Computation Institute U O C, Chicago, Il 60637, United States, Lawrence Berkeley National Laboratory E T A, 1 Cyclotron Road, Berkeley, Ca 94720, United States, Lawrence Berkeley National Laboratory E T A, 1 Cyclotron Road, Berkeley, Ca 94720, United States, et al. Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.
[26] S.P. Ong, W.D. Richards, A. Jain, et al., Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319.
[27] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[28] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, CoRR (2016) 02754 abs/1603.
[29] L.S. M, E. Gabriel, C. Hugh, et al., From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (1) (2020) 56–67.
[30] W. Mckinney, Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61, https://doi.org/10.25080/majora-92bf1922-00a.
[31] H.C. R, M.K. Jarrod, V.D.W.S. J, et al., Array programming with NumPy, Nature 585 (7825) (2020) 357–362.
[32] A. Jain, S.P. Ong, G. Hautier, et al., Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013), 011002-011002-011011.
[33] K. Choudhary, K.F. Garrity, A.C.E. Reid, et al., The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, npj Comput. Mater. 6 (1) (2020), 173-173.
[34] L. B Z J, J. M R, B. M C, Dataset of theoretical multinary perovskite oxides[J], Sci. Data 10 (1) (2023), 244-244.
[35] Z. Lu, X. Yang, C. Jin, et al., Nonvolatile electric-optical memory controlled by conductive filaments in Ti-Doped BiFeO3, ADVANCED ELECTRONIC MATERIALS 4 (2) (2018), https://doi.org/10.1002/aelm.201700551.
[36] G.S. Priyanga, N. M M, N. N, et al., Prediction of nature of band gap of perovskite oxides (ABO3) using a machine learning approach, Journal of Materiomics 8 (5) (2022) 937–948.
[37] P. Barham, A. Donnelly, R. Isaacs, et al., Using magpie for request extraction and workload modelling [Z]. Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation - Volume 6, USENIX Association, San Francisco, CA, 2004, p. 18.
[38] Z.S.B.A. Zixin, Classification of battery compounds using structure-free mendeleev encodings, J. Cheminf. 16 (1) (2024), 47-47.
[39] H. S I. Machine Learning, Algorithms, real-world applications and research directions, SN Computer Science 2 (3) (2021), 160-160.
[40] P.L.Z. Severynen, Data Discovery Analysis on Complex Time Series Data[M], 2022.
[41] J. Hu, J. Wang, H. Chen, et al., Automated design of hybrid halide perovskite monolayers for band gap engineering, npj Comput. Mater. 10 (1) (2024), https://doi.org/10.1038/s41524-024-01323-5.
[42] B.M. Abraham, P. Sinha, P. Halder, et al., Fusing a machine learning strategy with density functional theory to hasten the discovery of 2D MXene-based catalysts for hydrogen generation, J. Mater. Chem. A 11 (15) (2023) 8091–8100, https://doi.org/10.1039/d3ta00344b.
[43] G. Ding, B. Chu, Y. Jin, et al., Comparison of orthogonal regression and least squares in measurement error modeling for prediction of material property, Adv. Mater. Res. 661 (2013) 166–170.
[44] Q.R. Cai, J. Tang, Y.J. Liu, Translated to reflect the focus on SVM (plural: support vector machines), Comput. Meas. Control 18 (11) (2010) 2478–2480+2484, https://doi.org/10.16526/j.cnki.11-4762/tp.2010.11.012.
[45] S. Pakala, D. Ahn, E. Papalexakis, Tensor completion for surrogate modeling of material property prediction, Arxiv (2025).
[46] H. Mahmudul, A. Pinar, Machine learning reinforced microstructure-sensitive prediction of material property closures, Comput. Mater. Sci. (2022) 210.
[47] D. Yabo, D. Rongzhi, C. Zhuo, et al., Computational prediction of critical temperatures of superconductors based on convolutional gradient boosting decision trees, IEEE Access 8 (2020) 57868–57878.
[48] N. Alamsyah, B. Budiman, T. Yoga, et al., XGBOOST hyperparameter optimization using randomizedsearchcv for accurate forest fire drought condition prediction, Jurnal Pilar Nusa Mandiri (2024), https://doi.org/10.33480/pilar.v20i2.5569.
[49] R. Smarak, G.S. P, N. N, et al., Discovery of direct band gap perovskites for light harvesting by using machine learning, Comput. Mater. Sci. (2022) 210.
[50] M.J.J. Douglass, Hands-on machine learning with Scikit-learn, Keras, and tensorflow, Physical and Engineering Sciences in Medicine 43 (3) (2020) 1135–1136, https://doi.org/10.1007/s13246-020-00913-z, 2nd edition.
[51] Y. Liu, B.-J. Wang, S.-G. Lv, Using multi-class AdaBoost tree for prediction frequency of auto insurance, J Appl. Finance Bank. 4 (5) (2014).
[52] W. Di-Ni, L. Lang, Z. Da, Corporate finance risk prediction based on LightGBM, Inf. Sci. 602 (2022) 259–268.
[53] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[54] R. Guidotti, A. Monreale, S. Ruggieri, et al., A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) 1–42.
[55] B. C, Pattern Recognition and Machine Learning[M], 2006.
[56] Bengio Goodfellow, Courville. Deep Learning[M], MIT press, 2016.
[57] L. Xiao, H. Liu, X. Liu, et al., Discovering key factors determining perovskite bandgap under data scarcity inspired by knowledge distillation, J. Colloid Interface Sci. 695 (2025) 137827, https://doi.org/10.1016/j.jcis.2025.137827.
[58] M.V. Jyothirmai, R. Dantuluri, P. Sinha, et al., Machine-learning-driven high-throughput screening of transition-metal atom intercalated g-C(3)N(4)/MX(2) (M = Mo, W; X = S, Se, Te) heterostructures for the hydrogen evolution reaction, ACS Appl. Mater. Interfaces 16 (10) (2024) 12437–12445, https://doi.org/10.1021/acsami.3c17389.