Zhang *et al*. *Microstructures* **2025**, *5*, 2025029
DOI: 10.20517/microstructures.2024.77

Microstructures

Research Article

Open Access

MELRSNet for accelerating the exploration of novel ultrawide bandgap semiconductors

Zhesi Zhang[1], Hongzhou Song[2], Yinghui Ji[1], Yan Cui[1], Xiang Li[3], Zili Zhang[4], Ziming Cai[5], Jie Zhang[6], Yunyi Wu[7], Huanxin Li[8], Bingcheng Luo[1]

[1]College of Science, China Agricultural University, Beijing 100083, China
[2]Institute of Applied Physics and Computational Mathematics, Beijing 100094, China.
[3]School of Energy and Power Engineering, Beihang University, Beijing 100191, China.
[4]School of Science, China University of Geosciences (Beijing), Beijing 100083, China.
[5]School of Materials Science and Physics, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China
[6]Research Center for Metamaterials, Wuzhen Laboratory, Jiaxing 314500, Zhejiang, China
[7]Research Center for Comprehensive Energy Technology, CTG Science and Technology Research Institute, Beijing 100038, China.
[8]Department of Chemistry, Physical & Theoretical Chemistry Laboratory, University of Oxford, Oxford OX1 3QZ, United Kingdom.

Correspondence to: Prof. Bingcheng Luo, College of Science, China Agricultural University, No.17 Qinghua East Road, Haidian District, Beijing 100083, China. E-mail: luobc21@cau.edu.cn; Prof. Hongzhou Song, Institute of Applied Physics and Computational Mathematics, No. 2 Fenghao East Road, Haidian District, Beijing 100094, China. E-mail: song_hongzhou@iapcm.ac.cn

How to cite this article: Zhang, Z.; Song, H.; Ji, Y.; Cui, Y.; Li, X.; Zhang, Z.; Cai, Z.; Zhang, J.; Wu, Y.; Li, H.; Luo, B. MELRSNet for accelerating the exploration of novel ultrawide bandgap semiconductors. *Microstructures* **2025**, *5*, 2025029. https://dx.doi.org/10.20517/microstructures.2024.77

Received: 29 Aug 2024 **First Decision:** 30 Oct 2024 **Revised:** 08 Nov 2024 **Accepted:** 18 Nov 2024 **Published:** 18 Mar 2025

Academic Editor: Liangzhi Kou **Copy Editor:** Guomiao Wang **Production Editor:** Guomiao Wang

Abstract

Ultrawide bandgap (UWBG) semiconductors, with bandgaps exceeding 3.4 eV of gallium nitride, offer the potential to overcome the limitations of conventional semiconductors and drive innovations in electronics and photovoltaics. However, discovering such materials remains a huge challenge due to the prohibitive cost of trial-and-error-based experiments and the complexity of cutting-edge quantum mechanical approaches. Here, we develop the Multistage Ensemble Learning Rapid Screening Network (MELRSNet), a data-driven hierarchical machine learning framework integrated with high-throughput first-principles calculations, designed for swift identification of UWBG semiconductors. Trained on the Materials Project dataset, MELRSNet utilizes elemental and structural features to classify, regress, and validate potential candidates. Its efficacy is underscored by the accurate prediction of bandgaps in UWBG oxides and the revelation of metric-bandgap relationships, aligning closely with first-principles calculations. Furthermore, MELRSNet's reliability is bolstered through the identification of eight novel ternary oxide compounds, derived from monoclinic hafnium oxide crystals, exhibiting high stability, desirable band gaps, and strong ultraviolet light absorption, marking them promising candidates for lab synthesis and subsequent applications. MELRSNet not only streamlines the discovery of UWBG semiconductors but also paves the way for high-throughput computational screening of other functional materials.

Keywords: Ultrawide bandgap semiconductor, machine learning, density functional theory, stacked generalization, LightGBM

INTRODUCTION

The emergence of semiconductors has greatly advanced modern power electronics[1], optoelectronics[2], photodetectors[3], and other applications[4,5]. Wide bandgap (WBG) semiconductors including indium gallium nitride (InGaN), silicon carbide (SiC) and gallium nitride (GaN) have undergone rapid development in the past three decades[5-8]. Recently, ultrawide bandgap (UWBG) semiconductors with bandgaps wider than 3.4 eV of GaN have attracted much attention due to their superior performance limits[1], such as higher Baliga figure of merit compared to conventional WBG counterparts[9]. The extraordinary properties of UWBG semiconductors, such as high breakdown voltage, high optical transparency, wide optical bandgap, and high thermal conductivity, make them promising candidates in the fields of high-power electronics[10], solar-blind deep-ultraviolet photodetectors[11] and extreme-environment electronics[9]. In recent years, a wide variety of UWBG semiconductors have been reported, including group III nitrides (e.g., BN, $Al_xGa_{1-x}N$, and $In_xGa_{1-x}N$[12-14]), oxides (e.g. β-Ga_2O_3[10], $Mg_xZn_{1-x}O$[15] and $ZnGa_2O_4$[16]), chalcogenides (e.g., GeS_2[17], Ga_2S_3[18,19] and $GaPS_4$[20]), and diamond[21]. Although the family of UWBG semiconductors has been continuously enriched, there are still thousands of potential substances waiting for further exploration. Therefore, designing and exploring novel UWBG semiconductors with high stability and excellent performance is important to expand the corresponding material library.

Hafnium dioxide-based materials have recently been investigated for their high stability, UWBG, and ferroelectricity, demonstrating potential applications in sensors, actuators, and memories[22-24]. Given this, it is particularly critical to explore HfO_2-based semiconductors. Using a multistage first-principles computational workflow, Garrity *et al.*[25] reported three unexplored ternary UWBG oxides for high-power electronics, namely $In_2Ge_2O_7$ thortveitite and pyrochlore, Mg_2GeO_4 spinel and $InBO_3$ calcite, and called for further exploration of unexplored ternary oxides UWBG materials. However, trial-and-error experiments are resource-, equipment- and time-consuming due to the wide range of possible compositions and structures, necessitating the use of computational simulations to screen materials. State-of-the-art computational approaches, in particular density function theory (DFT), have greatly accelerated the process of designing and predicting new materials, allowing high-throughput screening and prediction of bandgaps based on various forms of exchange-correlation functional, the most traditional of which is the generalized gradient approximation (GGA) of Perdew-Burke-Ernzerhof (PBE)[26]. Gorai *et al.*[27] calculated bandgaps of tetrahedrally bonded structures through GGA-PBE functional to find promising substitutions with better back-contact properties than the common ZnTe. However, the bandgap is inherently underestimated by about 40% due to the delocalization errors and derivative discontinuities[28]. By introducing a nonlocal Hartree-Fock exchange potential, the Heyd-Scuseria-Ernzerhof (HSE) hybrid functional achieves better accuracy in predicting bandgaps[29]. However, HSE functional requires excessive computation and time and therefore cannot be applied in large numbers for high-throughput screening of candidates with targeted properties such as bandgaps[30-32].

One of the most important identifiers of UWBG is its width greater than 3.4 eV in experiments. Increasingly, computer simulations are generating large amounts of data, promoting the construction of standard material science databases, including Materials Project (MP)[33], Open Quantum Materials Database (OQMD)[34] and Automatic Flow of Materials Discovery Library (AFLOW)[35], which enable the application of machine learning (ML) methods. ML can be used as an initial screening as it achieves a high degree of fit in a relatively short time[36]. ML methods such as kernel ridge regression (KRR), support vector machine (SVM), gradient boosting decision tree (GBDT), eXtreme gradient boosting (XGBoost), random forest (RF), and deep neural networks (DPNN) have been applied to tailor bandgaps, predict complete band structures and design materials with desired properties[37], in particular perovskites[38-41], inorganic materials[42-45], two-dimensional materials[46-48], phosphors[49] and metal-organic frameworks (MOFs)[31]. Wang *et al.*[50] reviewed the application of ML in predicting bandgaps and other properties of perovskites. Zhuo *et al.*[45] proposed SVM models to distinguish metals from non-metals and predicted bandgaps of non-metals based solely on elemental compositions. Shen *et al.*[51] screened four UWBG double perovskites by first-principles study combined with interpretable ML models. However, the accuracy of ML is highly dependent on the range of data in the training dataset and the prediction sample set. This limitation makes existing methods insufficient to accelerate the exploration of UWBG oxide semiconductors.

In this work, we developed the Multistage Ensemble Learning Rapid Screening Network (MELRSNet), a hierarchical framework based on ML and high-throughput first-principles calculations, for the rapid screening of promising UWBG semiconductors on the fly. After solving the data imbalance problem by the Synthetic Minority Oversampling Technique and Edited Nearest Neighbor (SMOTE-ENN) algorithm, we established a classification model combining five ML algorithms through a stacking strategy, to discriminate compounds as UWBG or non-UWBG. Then, we employed the LightGBM model along with Shapley Additive exPlanations (SHAP) analysis to quantitively predict bandgaps of UWBG and explore relationships between materials properties and their bandgaps. Based on the MELRSNet framework, 97 potential UWBG semiconductors were preliminarily screened from 835 candidates by combining element substitution and valence balance strategies. Meanwhile, SHAP analysis was used to map the relationship between input features reflecting material characteristics and the resulting electronic property (i.e., bandgap). Then, verified by DFT calculations, eight ternary oxides were eventually predicted as promising UWBG materials with high stability, UWBG, and outstanding optical properties.

MATERIALS AND METHODS

SMOTE-ENN

The SMOTE-ENN[52] is an algorithm dealing with data imbalance in the categorical dataset. The imbalance of data (i.e., the scale of data volume varies greatly among different categories) will make the model tend to predict the sample as the majority class, leading to biased predictions. SMOTE-ENN combines over- and under-sampling through over-sampling by SMOTE and removing redundant data by ENN. Specifically, SMOTE[53] interpolates new points by first selecting a minority class instance *a* at random and finding one of its k nearest minority class neighbors *b* at random. The synthetic instance *c* is generated according to:

$$c = a + \text{rand}(0,1) * (b - a) \tag{1}$$

where rand (0,1) is a random number from zero to one.

However, SMOTE may generate noisy samples if inducing new instances between marginal outliers and inliers. ENN is then introduced as a cleaning method to remove unnecessary and noisy samples. SMOTE-ENN was realized here through the scikit-learn package[54] in Python.

Optimization of hyperparameters

One of the most widely used hyperparameter tuning schemes is grid search. Here, we applied grid search of all ML classification models and the KRR regression model with five-fold cross-validation. It is an exhaustive search algorithm, which will try all possible combinations of parameters within the given hyperparameter range, and choose the hyperparameters set with the best model performance. Although accurate, grid search often causes a significant amount of time consumption which is not proportional to model improvement. Therefore, we adopted Bayesian optimization in the rest of the ML regression models. Bayesian optimization is dedicated to revealing the black box relationship between hyperparameters and model performance by two main functions: surrogate and acquisition. Specifically, it will start with some randomly selected samples and utilize them to compute the surrogate function. Then comes an iteration in a loop: the acquisition function is introduced to choose the next evaluation sample point, and the surrogate function is re-evaluated under the new sample points. The loop will terminate until the variance between two adjacent loops falls below the set threshold. We carried out Bayesian optimization with five-fold cross-validation through the Scikit-Optimize (skopt) package, where the surrogate function, the scoring criteria, and the number of parameters settings that are sampled were set to be Gaussian process, R^2, and 30, respectively. Bayesian optimization was proved to have similar accuracy to the grid search method with significantly reduced time cost since it tested only a limited number of hyperparameter sets (30 in our example).

Stacking

Stacked generalization (stacking)[55] is an ensemble learning technique to combine multiple models via a meta-learner. The individual classification/regression models are trained based on the complete training set, and the meta-learner is fitted based on the outputs (meta-features) of the individual classification/regression models (base-learners) in the ensemble. Good base learners should be heterogeneous strong learners. That is to say, each learner should have outstanding accuracy, and algorithms should be as distinguished as possible among different base learners. In this work, all ML models were built through an open-source scikit-learn package[54]. The base learners of the classification model were SVC, Adaboost, XGBoost, LightGBM, and RF. Hyperparameters of base learners were optimized through grid search with five-fold cross-validation. The ridge regression was used as the meta-learner of the classification stacking model.

Feature selection procedure

Recursive Feature Elimination (RFE)[56] is a wrapper-type feature selection strategy. Feature selection can allow ML algorithms to run more efficiently and avoid over-fitting or being misled by irrelevant features. By fitting the given ML algorithm, RFE ranks features by importance and discards the specified number of least important features. The procedure continues re-fitting the model and removing features until the desired number remains. Recursive feature elimination with cross-validation (RFECV) is configured similarly to the RFE, performing cross-validation evaluation of a different number of features. The selection of ML algorithms used in the core RFE highly influences the results and effect. Here, we performed RFECV in the scikit-learn package[54], choosing XGBoost, LightGBM, RF, and Adaboost as the base algorithms. In the classification model, the model accuracy kept improving as the number of features increased, indicating that there was no need to delete any feature when dealing with classification problems. XGBoost turned out to perform the best regarding the mean test squared error in the regression problem. Eventually, 43 features were screened out with the consideration of both model accuracy and training efficiency.

LightGBM

LightGBM[57] is one of the most effective implementations of the Gradient Boosting Decision Tree (GBDT) algorithm, which is especially competitive when handling large samples and multi-dimensional data. It significantly improves the model accuracy, enhances the computation speed, and reduces memory usage by introducing two novel strategies: Gradient-based one-sided sampling (GOSS) and Exclusive Feature Bundling (EFB). Through GOSS, a large amount of data instances with small gradients are out of consideration during the process of estimating the information gain, which modifies the traditional GBDT that needs scanning all the data instances. EFB reduces the dimension of features by binding mutually exclusive features. In addition, the GBDT algorithm based on histogram and leaf-wise growing technique with max depth constraints is introduced to further improve the model performance. In this work, we realized LightGBM and other ML methods through the scikit-learn package[54].

SHAP analysis

The SHAP framework was used in the classification and regression model to interpret relationships between predictions and features. Complex models such as ensemble learning or deep learning models reveal characteristics of a black box, making it difficult to understand why the model makes such outputs. SHAP builds an explainable addictive feature attribution model, where all features are viewed as important values. The contribution of each feature is then evaluated by calculating its marginal contribution when added to the model. SHAP method provides a universal solution to explain the model output and guide the inverse design of features, which is in good consistency with human intuition. Here, we applied SHAP analysis by the shap package in Python.

Density functional theory

Candidate materials were predicted using density functional theory (DFT) based on first-principles calculations through the Cambridge Serial Total Energy (CASTEP)[58,59] and Vienna ab initio simulation package (VASP) code[60,61]. The GGA with PBE[26] exchange-correlation functional and norm-conserving pseudopotential were used to optimize the geometric structure and perform the self-consistent field calculations. The Broyden–Fletcher–Goldfarb–Shannon (BFGS) minimizer[62] was employed to conduct unit cell optimization. The kinetic energy cutoff was set to be 830 eV, and the Brillouin zone was sampled with the Monkhorst-Pack mesh grid[63] of 2 × 4 × 3 k points. The convergence criteria were set as 5×10^{-6} eV/atom for the self-consistent iteration loop, 0.001 eV/Å for the residual forces on all atoms, 0.02 GPa for the force tolerance, and 5×10^{-4} Å for the displacement tolerance. Projector augmented wave (PAW) pseudopotential[60,61] in the VASP code was also used for structural relaxation and calculation of band structures, where the kinetic energy cutoff was set to be 800 eV and the first Brillouin zone was sampled by a Monkhorst-Pack mesh grid which was generated automatically. The optimization will not be converged until total energy in the self-consistent field iteration falls below 1×10^{-8} eV/atom, and the residual forces on each atom are less than 0.001 eV/Å. HSE06 hybrid functional[64,65] was used to obtain accurate band structures. Phonon dispersion calculations[66] were employed to evaluate the lattice dynamic stability by density functional perturbation theory (DFPT)[67] with the linear response method in the CASTEP code. The X-ray diffraction (XRD) patterns were theoretically simulated through VESTA[68]. Ultraviolet-visible (UV-Vis) spectrums were calculated based on time-dependent DFT[69] through CP2K package[70] and Multiwfn[71]. PBE functional was adopted with the cutting of energy 350 eV. The convergence criterion of self-consistent field iteration was set to 1×10^{-6} eV. The new charge density was mixed with the old charge density in a ratio of 40%, and the number of excited states was set to 50.

RESULTS AND DISCUSSION

Model architecture

The workflow of MELRSNet is systematically summarized in Figure 1 and contains four main procedures: data preparation, ML classification model, ML regression model, and model validation through the discovery of novel UWBG ternary oxides with monoclinic XYO_4 structure.

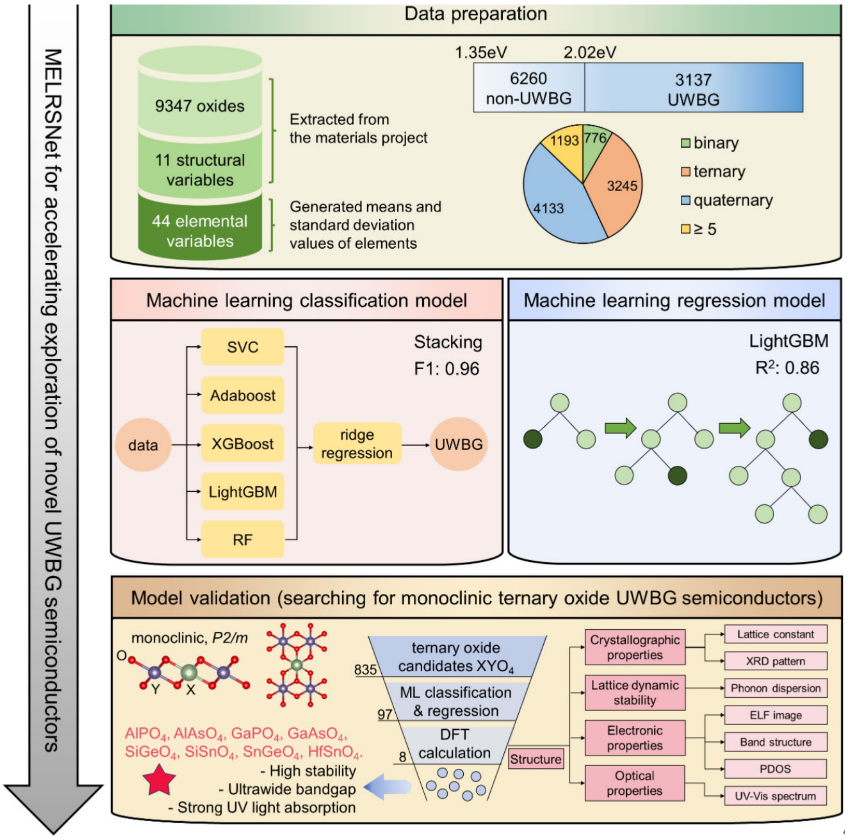
MELRSNet for accelerating exploration of novel UWBG semiconductors
Data preparation
9347 oxides
11 structural variables
44 elemental variables
Extracted from the materials project
Generated means and standard deviation values of elements
1.35eV
2.02eV
6260 non-UWBG
3137 UWBG
1193
776
3245
4133
binary
ternary
quaternary
≥ 5
Machine learning classification model
data
SVC
Adaboost
XGBoost
LightGBM
RF
ridge regression
UWBG
Stacking
F1: 0.96
Machine learning regression model
LightGBM
R^2: 0.86
Model validation (searching for monoclinic ternary oxide UWBG semiconductors)
monoclinic, *P2/m*
O
Y
X
835
ternary oxide candidates XYO_4
97
ML classification & regression
8
DFT calculation
Structure
Crystallographic properties
Lattice dynamic stability
Electronic properties
Optical properties
Lattice constant
XRD pattern
Phonon dispersion
ELF image
Band structure
PDOS
UV-Vis spectrum
$AlPO_4$, $AlAsO_4$, $GaPO_4$, $GaAsO_4$, $SiGeO_4$, $SiSnO_4$, $SnGeO_4$, $HfSnO_4$.
- High stability
- Ultrawide bandgap
- Strong UV light absorption

Figure 1. The flowchart of the MELRSNet framework. MELRSNet: Multistage Ensemble Learning Rapid Screening Network.

Data preparation

A total of 9347 oxides containing at least two elements with a GGA-PBE bandgap greater than 1.35 eV[9] and 11 corresponding structural variables were extracted as a training set from the MP[33], one of the most comprehensive open-source materials datasets. Since the band structure in the MP database is calculated through GGA-PBE and GGA + U functional[72], which exhibits an inherent underestimation of bandgaps, a moderate adjustment was made to better reflect the experimental bandgap. Specifically, the experimental bandgap for the lower limit of the WBG material is 2.3 eV[9], while the calculated PBE bandgap is 1.35 eV, which is an underestimation of 41.1%[73]. Likewise, a PBE bandgap of 2.02 eV was defined as the lower limit for the screened materials in the dataset, based on the experimental bandgap of the UWBG material (3.4 eV)[74]. Deviation correction of 41.1% is also consistent with the internal test results of the MP database, where the experimental gaps are approximately 1.6-fold of the computed gaps[72]. Based on previous research[45], the 44 elemental variables are obtained according to the mean and standard deviation of 22 elemental properties such as atomic number, ionic radius, and Pauling electronegativity. Of these, 3137 materials are labeled as UWBG due to their PBE bandgap greater than 2.02 eV, while the remaining 6210 materials are labeled as non-UWBG. Therefore, an initial classification dataset consisting of 9347 samples is created with one binary dependent variable (0 for non-UWBG, 1 for UWBG) and 55 features (11 structural and 44 elemental). These 3137 UWBG samples are extracted as the regression dataset. All features and their explanations are listed in Supplementary Table 1. Among the training database, there are 776 binary oxides, 3245 ternary oxides, 4133 quaternary oxides, and 1193 items with elements greater than or equal to five, reflecting the comprehensive characterization of all WBG and UWBG oxides in our dataset. The elemental and numeric distributions of our training dataset are exhibited in Figure 2A and Figure 2B.

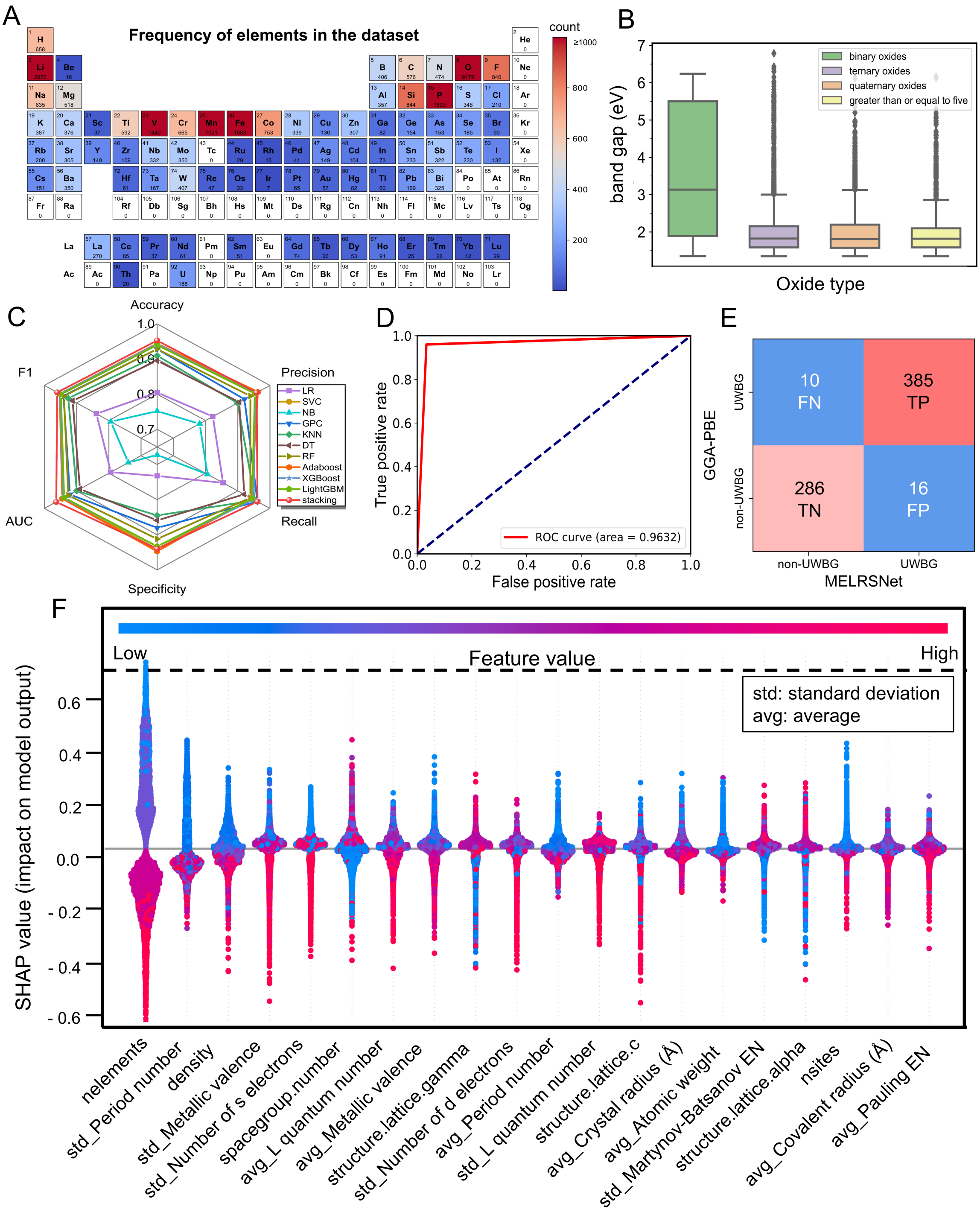

Figure 2. Distribution of training dataset and performances of the classification model. A, The occurance frequencies of different elements in the training dataset. B, The bandgap distribution of different types of oxides. C, The radar chart of eleven ML models regarding six evaluation indicators. D, The ROC curve of the stacking classification model. E, The confusion matrix of the stacking classification model. F, SHAP analysis for the feature importance of the stacking classification model. ML: machine learning; SHAP: Shapley Additive exPlanations. ROC: receiver operating characteristic

Classification model to discriminate UWBG materials

Firstly, we categorized the WBG and UWBG materials to determine whether the compound might be UWBG materials. To eliminate the issue of data imbalance in the dataset, the SMOTE-ENN was applied [52]. This algorithm balances the class distributions by oversampling the minority class through interpolation and then removing redundant samples through an ENN strategy. The classification precision on the test dataset of XGBoost algorithm has improved from 83.51% to 94.57% after utilizing the SMOTE-ENN strategy, which justifies its effectiveness. The quantities of non-UWBG and UWBG samples were reduced to 2962 and 4009, respectively. Standardization was then performed to ensure that the mean of each variable value is zero and the variance is one. To achieve better performance of the model, it is essential to select features that not only perfectly reflect the trend of the dependent variable, but also avoid feature autocorrelation, the curse of dimensionality, and the waste of computational resources[40]. RFECV[56] was employed for feature engineering. Four different ML algorithms were used as the ranking criterion, all of which indicate that there is no redundant variable in the classification problem, as shown in Supplementary Figure 1A. The test set size was fixed to 0.1 through a selection procedure as shown in Supplementary Figure 1B.

Different ML algorithms significantly affect the performance of the model. Here, ten state-of-the-art data-driven ML classification models were firstly employed for training and classification, which included logistic regression (LR), support vector classification (SVC), Naïve Bayes (NB), Gaussian process classification (GPC), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), Adaboost, XGBoost, and LightGBM. The hyperparameters of each model were optimized through grid search with 5-fold cross-validation. The hyperparameter details are given in Supplementary Table 2. Five indexes were derived to evaluate the performance of the model, where Accuracy = (TP + TN)/(TP + TN + FP + FN), Precision = TP/(TP + FP), Recall = TP/(TP + FN), Specificity = TN/(TN + FP), and F1 = 2 × (Precision × Recall)/(Precision + Recall). TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. Besides, the receiver operating characteristic (ROC) curve plots the trend between the true positive rate (recall) and the false positive rate (1-specificity). The area under the curve of ROC (AUC) reflects the model's ability to make precise classifications. Recall reflects the proportion of UWBG samples successfully classified as UWBG, while precision shows the proportion of samples classified as UWBG that are indeed real UWBG materials. These two values are of most interest because we are more concerned with UWBG materials than non-UWBG materials. Therefore, the F1 score, which combines both recall and precision, was considered as the main evaluation indicator. Among the ten models, SVC, Adaboost, XGBoost, LightGBM, and RF performed the best on the test set, with F1 scores of 0.9525, 0.9523, 0.9515, 0.9489, and 0.9395, respectively. NB achieved the worst performance, with an F1 score on the test set of only 0.7938. The five most effective models performed satisfactorily on the test set, but all showed slight overfitting as the performance was 5% better on the training set than on the testing set. Therefore, a stacked generalization (stacking)[55] model with SVC, Adaboost, XGBoost, LightGBM, and RF as base learners and ridge regression as meta learner was constructed to further enhance the classification efficiency, take advantage of different strong learners, and make the classification results more robust. Details of the stacking model are given in the Materials and Methods. The stacking model proved to be in better agreement with the dataset than other classification models, thus being selected as the mapping model from the input structural and elemental features to the binary UWBG classification problem. Performances of all state-of-the-art classification models on the testing set are shown in Figure 2C, Table 1 and Supplementary Table 3. As the ROC curve and the confusion matrix of the stacking model on the testing set shown in Figure 2D and Figure 2E, only 26 of 697 compounds are misclassified, indicating the high efficiency of differentiating UWBG and non-UWBG by our classification model.

Table 1. Comparisons of the performance of stacking classification model and LightGBM regression model with other state-of-the-art machine learning models

F1 score of classification model										
LR	SVC	NB	GPC	KNN	DT	RF	Adaboost	XGBoost	LightGBM	stacking
0.84	0.95	0.79	0.94	0.93	0.91	0.94	0.95	0.95	0.95	0.96
R^2 score of regression model										
LinR	GPR	KRR	DT	RF	Adaboost	XGBoost	LightGBM			
0.63	0.76	0.82	0.79	0.83	0.71	0.86	0.86			

LR: logistic regression, SVC: support vector classification, NB: Naïve Bayes, GPC: Gaussian process classification, KNN: k-nearest neighbor, DT: decision tree, RF: random forest. LinR: linear regression. GPR: Gaussian process regression, KRR: kernel ridge regression. The models used in our framework are marked in red.

Although ensemble learning methods such as stacking generally demonstrate superior performance on the tasks, the "black box" character makes it challenging to obtain the basis for their predictions. In order to enhance the explainability of the ensemble stacking model, we employed a training dataset to perform the SHAP analysis. As shown in Figure 2F, the summary plot summarizes the entire distribution of SHAP value for the top 20 most important features. The y-axis denotes the features, which are ordered from left to right by the influence on the output. The dot colors from blue to red show the feature values from lowest to highest. Therefore, the high feature value in the region with a positive SHAP value and the low value with a negative SHAP value indicate the positive relationship between the feature and output. Here, the number of unique elements in the compound (nelements) ranked first in classifying the UWBG materials. Compounds with fewer elemental constituents are more likely to exhibit UWBG, possibly due to their potential to have simpler crystal structures, higher symmetry, and reduced defect levels, which collectively stabilize the electronic band structure and widen the bandgap. Large standard deviation of period numbers (std_Period number) for elements and small spacegroup number reflect the large differences among elements and the low symmetry system, providing higher possibility to the instability such as lattice mismatch and energy level splitting, and thus cause the reduction in bandgap[75]. The large density and large sites per cell (nsites) generally increased the interactions of electrons among elements, leading to the narrowing of forbidden band. Larger electronegativity difference tends to from ionic bonds with stronger electron bonding of nucleus, making the electrons harder to move, resulting in wider bandgaps[76].

Regression model for bandgap prediction

We then built a regression model to further predict the bandgaps of potential UWBG materials. Similarly, we selected 43 optimal features through RFECV, as shown in Supplementary Figure 2A. Four different ML algorithms are used as the ranking criterion, where XGBoost performed the best with the lowest mean squared error in five folds. The mean absolute error (MAE), the root-mean-squared error (RMSE), the coefficient of determination (R^2), and the Pearson correlation coefficient (rho) were used as evaluation criteria, and the optimal test set ratio was set as 0.1 based on XGBoost algorithm [see Supplementary Figure 2B for details]. Each feature was normalized and transformed to the range [0,1] respectively.

Herein, we applied eight different ML regression algorithms, including linear regression, Gaussian process regression (GPR), KRR, DT, RF, Adaboost, XGBoost, and LightGBM. The hyperparameters of KRR were optimized through grid search with five-fold cross-validation. Considering the huge time cost of grid search with complex hyperparameters, the hyperparameters of the later five algorithms were evaluated using five-fold cross-validation under Bayesian optimization. The computational details are given in Supplementary Table 4 and Supplementary Figure 2C. Four indexes including MAE, RMSE, R^2, and rho were chosen to evaluate the fitting efficiency of models. Small MAE and RMSE values indicate small model errors, while large R^2 and rho values indicate a closer fit to the model, all reflecting the excellent performance of the ML model. Figure 3A, Table 1 and Supplementary Table 5 summarize the performances of nine ML models on the testing set, where XGBoost and LightGBM performed better than other regression models, with R^2 values on the test set of 0.8584 and 0.8588, respectively. The DT-based ensemble learning regression models, including RF, XGBoost and LightGBM, all had R^2 values greater than 0.8, achieving higher accuracy than the single DT model, whose R^2 value on the test set was 0.7899. Compared with XGBoost, LightGBM adopts several strategies to reduce the complexity of training, including Gradient-based One-Side Sampling (GOSS) and EFB for reducing the amount of data and features, leaf-wise split, and histogram-based method for accelerating the training process. Therefore, due to the demand for smaller memory requirements and faster computing speed, we selected LightGBM as the final regression model.

The LightGBM model achieved outstanding consistency with the training set, where MAE, RMSE, R^2, and rho are 0.1207, 0.1833, 0.9713, and 0.9858, respectively. Figure 3B presents its corresponding performances on the test set. Most samples fit well by our LightGBM regression model, but there do exist outliers. According to the red circle in Figure 3B, the two worst-fitting samples are N_8O_4 and $Cs_4Al_4H_{96}S_8O_{80}$, which may be caused by the small sample size of binary and pentagonal oxides in the training dataset. Although the feature engineering and regularization have been applied, there is a slight over-fitting of the final model, which is most likely due to the insufficient numbers of dataset samples and the right-skewed distribution of samples on the bandgaps, which can be seen in Figure 3C. To get further insight on the statistical fitting consistency, the hypothesis tests for correlation and differences were calculated and listed in Supplementary Table 6. The correlation test shows a significant linear relationship between the test results of LightGBM and DFT, while the differences test indicates the prediction consistency on bandgaps, both of which prove the excellent fitting performances of our LightGBM regression model. Since the dataset comprises DFT calculation data, the underestimation of the experimental bandgap still exists. However, as for predicting UWBG materials, LightGBM has been proven to outperform the accuracy of most DFT methods with much less time cost.

To enhance the interpretability of the regression model and gain insight into the relationships between variables and regression results, the SHAP method was used to analyze the importance of features of the LightGBM regression model. From Figure 3D, 18 out of 20 most influential features are elemental features, indicating the elemental composition of the material may have a greater impact on its bandgap than its structural features. This inference aligns with the fact that polymorphs of Ga_2O_3 (α-, β-, ε-, κ-, γ-) exhibit the UWBG mutually, from 3.62 eV of κ-Ga_2O_3 to 5.3 eV of α- Ga_2O_3[77]. Consistent with previous research findings related to perovskites, factors such as electronegativities, atomic radii, and Mendeleev number significantly affect the bandgap of materials[51,78,79]. Among the top five influential features, "std_Atomic weight", "std_Atomic radius (Å)", "avg_Number of d electrons", and "std_Atomic number" were negatively correlated with the band gap value. Large "std_Atomic weight" and "std_Atomic number", symbols of large differences of atoms in a compound, may introduce instability such as lattice mismatch and an expansion of electronic energy level distribution to the lattice structure, thus leading to the lower band gap by the formation of defect state. Generally, the lattice constants scale up as the "std_Atomic radius (Å)" becomes larger, which will bring electron band overlap, thereby reducing the bandgap. For semiconductor materials, the *d*-orbitals are usually the valence electron orbital with higher electron density. As the number of *d*-electrons increases, the position of the valence band maximum (VBM) moves up, resulting in a decrease in the bandgap width. In particular, the spacegroup number of materials, which reflects the periodicity and symmetry of atomic arrangement in crystals, shows a positive relationship with the bandgap. The larger the spacegroup number is, the higher the symmetry of the materials, the smaller the splitting of electronic energy bands, and the greater the energy difference between the valence and conduction band, leading to a

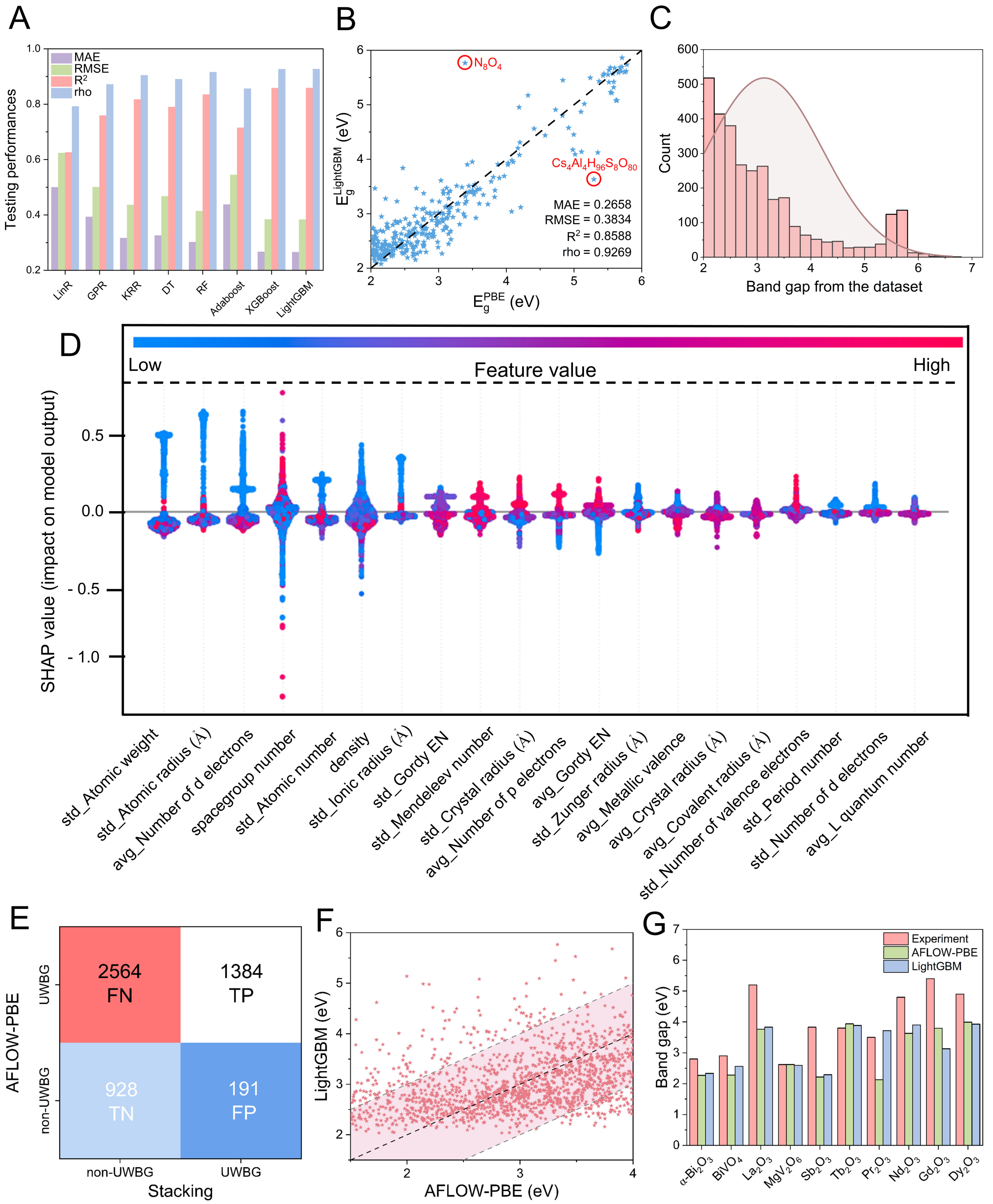

Figure 3. Performances of the regression model and validation procedure on the AFLOW dataset. A, Model performances of eight regression models based on the test set. B, Comparison between the predicted band gap by LightGBM and the calculated PBE band gap from the test dataset. The red circles display the two worst-fitting samples. C, Distribution of band gap in the regression dataset, which shows the right-skewed characteristic. D, SHAP analysis for the feature importance of the LightGBM regression model. E, Confusion matrix of stacking classification model based on the AFLOW dataset. F, Comparison between the predicted band gap and the calculated PBE band gap from the AFLOW database. G, Comparison among the bandgaps of several representative oxides obtained by experiments, PBE-functional calculation, and the LightGBM prediction. AFLOW: Automatic Flow of Materials Discovery Library; PBE: Perdew-Burke-Ernzerhof; SHAP: Shapley Additive exPlanations.

larger bandgap of semiconductor materials. The other structural feature, density, negatively affects the band gap mainly because increasing density generally results in decreased atomic spacing and enhanced interactions among atoms, potentially causing shifts in the energy levels of the conduction band and valence band and even phase transitions to a more compact structure, thereby leading to a reduced bandgap. The SHAP analysis provides a comprehensive explanation on the influence from features to the output, and is of great importance to guide the design of materials with desired properties based on the important features.

Model validation based on the performances on the AFLOW dataset

To study the general applicability of our proposed model in exploring UWBG oxide semiconductors, we evaluated its performance on another dataset called Automatic FLOW for Materials Discovery (AFLOW)[80]. AFLOW provides a globally available database for over 3 million compounds and their various calculated properties. Specifically, we filtered entries from the AFLOW database that contain oxygen, have two to eight types of elements, possess PBE-calculated bandgaps ranging from 1.35 eV to 4 eV, and originate from structures in the Inorganic Crystal Structure Database (ICSD)[81]. The ICSD is a standard database for completely identified inorganic crystal structures with high quality. To validate the performance on the new data, we removed entries already included in the training dataset, specifically those with the same chemical formula and space group number as the samples in the training set. For the entries with the same chemical formula and same space group number, we viewed them as the duplicate system and only remained the first entry. After obtaining their structural features and calculating elemental features, 5067 items were eventually filtered as the input dataset.

Then, the trained stacking classification model was employed to categorize these entries into non-UWBG and UWBG. To evaluate the efficiency of classification result, we manually labeled structures with a PBE calculated bandgap extracted from AFLOW dataset larger than 2.02 eV as UWBG and others as non-UWBG. The confusion matrix was then built as shown in Figure 3E and the performance statistics were calculated. The FN, TP, TN, and FP values are 2564, 1384, 928, and 191, correspondingly. The large FN count indicates that a large number of UWBG compounds in the AFLOW dataset were classified as non-UWBG by our stacking model. This misclassification results in poor accuracy and recall value. However, the precision and specificity achieved as high as 0.88 and 0.83. This discrepancy suggests that the model is highly effective in identifying true positives when it predicts a positive class, thus achieving high precision. However, the low recall indicates that the model fails to identify a substantial number of actual positive instances, resulting in a high number of false negatives. This situation may arise due to the inherent class imbalance (the number of UWBG instances is 3.5-fold of non-UWBG instances) in the filtered AFLOW dataset. Meanwhile, the manual classification and the simple transformation from the PBE bandgap to the experimental bandgap may introduce imprecisions to the read-world classification. Our classification model shows its strictness; that is, it would rather misclassify some UWBGs as non-UWBGs than predict non-UWBGs as UWBGs. However, this "strict" classification criterion is valuable due to our primary focus: we aim to ensure that as many of the predicted UWBG materials as possible are truly UWBG in reality. This is particularly important because these materials will undergo time-consuming DFT calculations later, and the false positive structures will cause unnecessary computational costs. Fortunately, our model performs well in this regard, with only 191 samples incorrectly classified as UWBG.

A total of 1575 structures classified as UWBG were then put in the LightGBM regression model. Figure 3F compares the calculated bandgap by PBE functional from the AFLOW dataset and the predicted bandgap by the LightGBM regression model. The black dashed line represents the regression result equal to the bandgap from the original dataset, and the two grey dashed lines indicate a deviation of ±1 eV between the regression result and the dataset bandgap. It can be seen that most of the dots (1317 out of 1575) fall into the pink region, which means the discrepancy between predicted and calculated bandgaps is less than 1 eV. In order to gain a deeper insight into the outliers that fall out of the pink region, we calculated the distribution of outliers' proportion based on the number and type of elements. As shown in Supplementary Figure 3A, as the number of elements increased, the proportion of outliers increased, mainly due to the insufficient training samples with a large number of elements. Similarly, Supplementary Figure 3B shows that the elements with large outlier proportions such as the rare earth elements are mostly those seldomly appeared in the training dataset, as plotted before in Figure 2A. This finding indicates that the model tends to make mistakes if it did not learn sufficient information during the training process, which again emphasizes the importance of the quality of training data. To compare the predicted bandgap with experiments, we manually extracted ten oxides with simple compositions and experimentally measured bandgaps, as shown in Figure 3G and Supplementary Table 7. Training from the PBE calculated data, the model cannot overcome the limitation of underestimating the bandgap without any transfer technique. However, one of the advantages of our ML-based model is its minimal time and computational cost, which will significantly accelerate the preliminary screening process for materials.

Model validation for the discovery of monoclinic ternary oxide UWBG

After confirming the usability on the new dataset, MELRSNet was validated for the applicability for the discovery of novel UWBG oxide semiconductors. Derived from the HfO_2 structure with their structural inputs and corresponding generated elemental parameters, we built 835 ternary oxide structures with composition XYO_4, monoclinic crystal system, and space group *P*2/*m*. We chose different elements at the X and Y sites and ensured that the sum of the valences of X and Y is 8 to satisfy the chemical equation equilibrium. Specifically, we selected 11 monovalent and six heptavalent elements, 33 divalent and six hexavalent elements, 40 trivalent and 12 pentavalent elements, and 14 tetravalent elements, as seen in Figure 4A. Out of 835 compounds, 97 materials were classified as candidate UWBG materials by our stacking classification model (see the full list in Table S8). Then, the MELRSNet has been used to predict the bandgaps of selected formulas, with the predicted bandgaps ranging from 2.15 eV to 5.60 eV. After completing the preliminary screening process by the ML model, we performed first-principles calculations based on the DFT of the selected UWBG candidates to assess the crystallographic properties, lattice dynamic stability, electronic properties, and optical properties, and thus getting physical insights in their potentials for UWBG applications. Finally, nine compounds were calculated to be stable by phonon dispersion curves, namely $AlPO_4$, $AlAsO_4$, $GaPO_4$, $GaAsO_4$, $SiGeO_4$, $SiSnO_4$, $SnGeO_4$, $HfSnO_4$, and $HfBiO_4$. Except for $HfBiO_4$, whose PBE and HSE band structure shows metallic character, all other eight materials have desired bandgaps and are verified as UWBG ternary oxide materials by DFT calculations.

The optimized structures of selected nine materials are drawn in Figure 4B. All materials belong to the space group *P*2/*m* and have a monoclinic crystal structure. The lattice angle β shows different degrees of lattice tilt, where $SiSnO_4$ has the highest degree (β = 59.76°). The c-axis of $AlPO_4$ and $GaPO_4$ is elongated to meet the convergence criteria during geometric optimization, becoming two-dimensional monolayer materials. The parameters of selected materials are summarized in Figure 4C, Figure 4D, with detailed values in Table 2. To assess the lattice dynamic stability of these materials, the phonon dispersion curve was calculated by the DFPT methods along the high-symmetry lines in the Brillouin zone, as shown in Figure 4E, Figure 4F, and Supplementary Figure 4. The frequencies of $HfSnO_4$ show minor imaginary frequency. The frequencies of the other eight materials are positive, indicating no imaginary phonon modes of the seven materials. XRD patterns are simulated theoretically and displayed in Figure 4G. Except for $AlPO_4$ and $GaPO_4$, whose peak intensities appeared at 2θ = 4.82° and 3.40°, respectively, the other seven compounds showed peak intensities at around 2θ = 15.50°. The electron localization function (ELF)[82] images are shown in Figure 4H, Figure 4I, and Supplementary Figure 5. ELF, with values ranging from 0 to 1, demonstrates the degree of electron localization. The blue contours (ELF = 0.00) around Al atoms show a deficiency of electrons, while the green counterparts (ELF = 0.50) of P, Hf, and Bi atoms and O-p orbital display the delocalization

Table 2. Lattice parameters, bond lengths, and band gap calculated by PBE functional and HSE06 functional of selected nine stable materials

Materials	*a*	*b*	*c*	*β*	*α* = *γ*	X-O1	X-O2	Y-O1	Y-O2	E_g^{PBE}	E_g^{HSE}
$AlPO_4$	2.66	4.64	18.31	88.56	90.00	1.89	1.80	1.65	1.76	5.14	6.93
$AlAsO_4$	2.78	4.81	5.57	88.62	90.00	1.91	1.84	1.78	1.86	3.48	4.35
$GaPO_4$	2.76	4.98	25.99	86.54	90.00	2.07	1.95	1.66	1.80	3.89	5.71
$GaAsO_4$	2.88	5.09	5.94	79.20	90.00	2.09	1.98	1.80	1.90	2.95	4.34
$SiGeO_4$	2.80	4.87	5.66	89.75	90.00	1.93	1.91	1.82	1.84	4.41	6.38
$SiSnO_4$	2.87	4.96	6.12	59.76	90.00	1.96	1.93	1.80	1.87	4.76	5.83
$SnGeO_4$	2.97	5.13	5.64	89.77	90.00	1.98	1.96	1.93	1.95	3.67	5.56
$HfSnO_4$	3.09	5.38	5.71	86.91	90.00	2.07	2.04	1.98	1.99	3.77	5.20
$HfBiO_4$	3.24	5.71	5.77	86.95	90.00	2.07	2.07	2.18	2.15	0.00	0.00

E_g^{PBE} and E_g^{HSE} are in the unit of eV, and other figures are in the unit of Å. PBE: Perdew-Burke-Ernzerhof; HSE: Heyd-Scuseria-Ernzerhof.

features, which imply the formation of covalent bonds between these atoms. The red contours (ELF = 1.00) of O atoms represent the aggregation of electrons, implying the charge transfer from X and Y atoms to the O atoms, which have larger electronegativity.

Band structures of nine novel materials were calculated by GGA functional under norm-conserving pseudopotentials. It should be noticed that although electronic band structure is convincing, there exists a general underestimation of bandgaps by the traditional GGA calculation. Therefore, the HSE hybrid functional (HSE06) with projector-augmented wave (PAW) pseudopotential has been introduced to improve the accuracy of calculated bandgaps (see Materials and Methods for details). We measured the computation time required to calculate the bandgap for each material using different methods, with the results visualized in Figure 4J. LightGBM predicted the bandgap of each material in less than 0.001 s, while calculations using PBE and HSE06 functionals took around 100 s to near 10000 s, depending on the material's structure. Moreover, the time advantage will become more apparent in more complex systems, as the prediction time of LightGBM almost remains unaffected by the complexity of the structure and elemental composition. Therefore, as we previously stated, an accuracy and resources-cost trade-off is achieved by the LightGBM regression model, which provides an opportunity for pre-screening potential UWBG semiconductor materials.

The bandgaps calculated by PBE and HSE06 functional are displayed in Figure 5A, where the PBE-calculated VBM of eight materials has been aligned to 0 eV. It can be seen from Table 2, Figure 5B, Figure 5C, Supplementary Figure 6, and Supplementary Figure 7 that all the HSE06 bandgaps are wider than 3.4eV. To get further insights into the electronic mechanisms, we investigated the partial density of states (PDOS) of selected UWBG materials, as shown in Figure 5D, Figure 5E, and Supplementary Figure 8. The p orbitals of oxygen atoms contribute most to the VBM for all selected materials, while the conduction band minimum (CBM) mainly comprises O-p and Y-s orbitals. d orbitals of Hf atoms in $HfSnO_4$ also show a large contribution near CBM. Y-p, X-s, and X-p orbitals generally affect higher conduction bands at over 6 eV. These UWBGs provide promising applications in high-power devices with enhanced electronic properties including a larger on/off ratio and threshold voltage[83,84].

Optical properties were then calculated based on TD-DFT methods to explore the application prospects of these materials in optical devices. Figure 5F shows the calculated UV-Vis spectrums of eight stable UWBG semiconductors. Eight novel UWBG materials have higher molar absorption coefficients and absorption peaks in the UV region. Except for $HfSnO_4$, other materials show double peaks with similar absorption

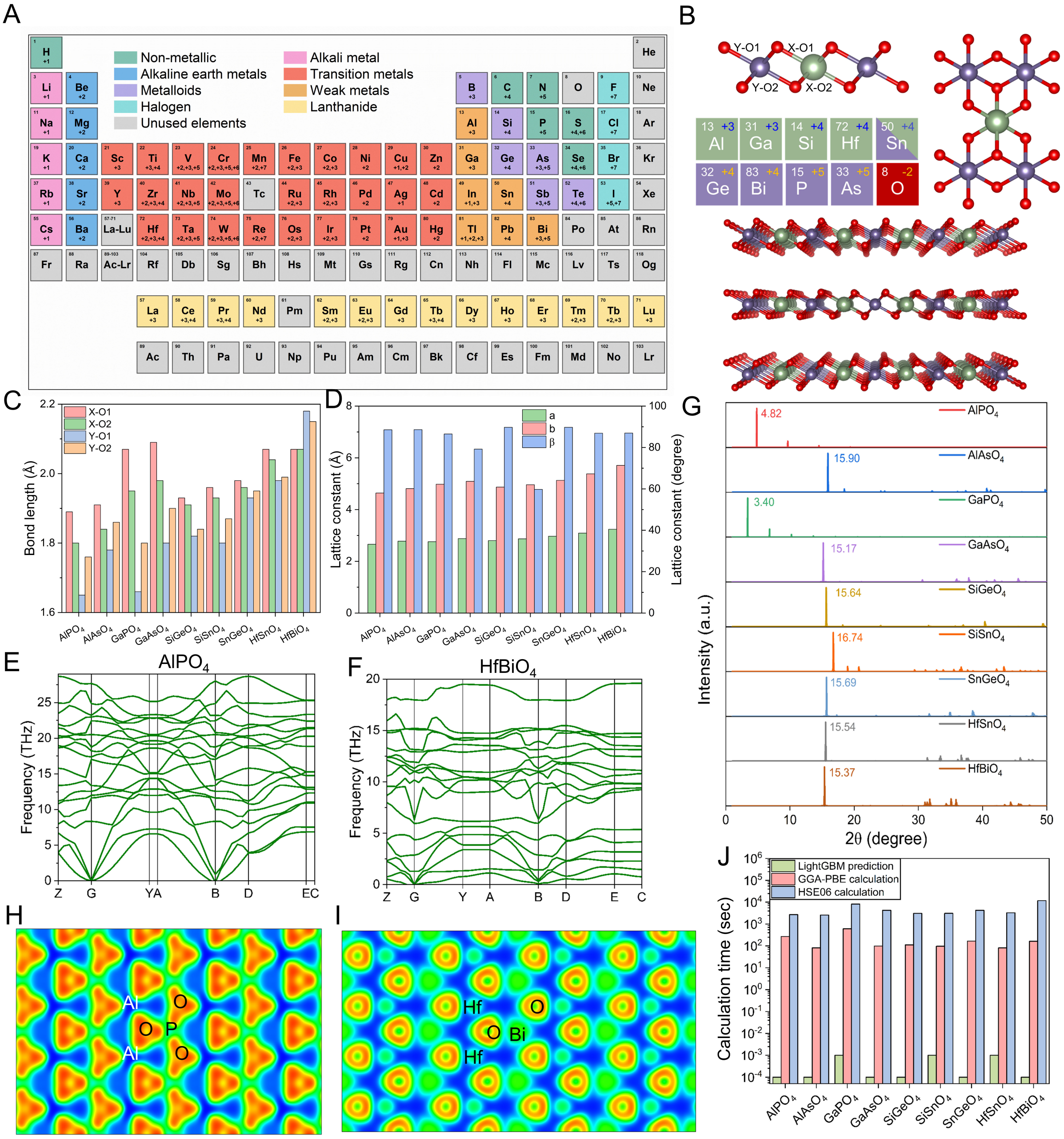

Figure 4. Selection and structures of validation ternary oxide materials. A, Selection of X and Y elements in XYO_4 compounds, where the sum of the valences of X and Y is controlled to be eight. Unused elements are labeled gray. B, Optimized structures of nine selected XYO_4 compounds by MELRSNet. From left to right are their side views and top views. The perspective view is at the bottom. Corresponding atoms with their atom number and valence number are labeled in the middle left corner. Four types of bonds are labeled on the side view. C**,** Optimized bond lengths of selected nine stable materials. Four columns represent four types of bonds in XYO_4 crystals that are explained in Figure 4B. D, Lattice constants of selected XYO_4 crystals. E**,** Phonon dispersion curves of $AlPO_4$. F**,** Phonon dispersion curves of $HfBiO_4$. G, Theoretical XRD patterns of nine selected materials. ELF image of H, $AlPO_4$, and I, $SnGeO_4$ at the planar view. J, Comparison of calculation time based on 32 CPU cores for the PBE functional, HSE06 functional, and LightGBM predictions. MELRSNet: Multistage Ensemble Learning Rapid Screening Network; ELF: electron localization function; CPU: central processing unit; PBE: Perdew-Burke-Ernzerhof; HSE: Heyd-Scuseria-Ernzerhof.

coefficients at over 190 nm and other peaks at shorter wavelengths. As for $HfSnO_4$, a single absorption peak is observed at around 300 nm. This provides the potential applications of these materials in deep ultraviolet

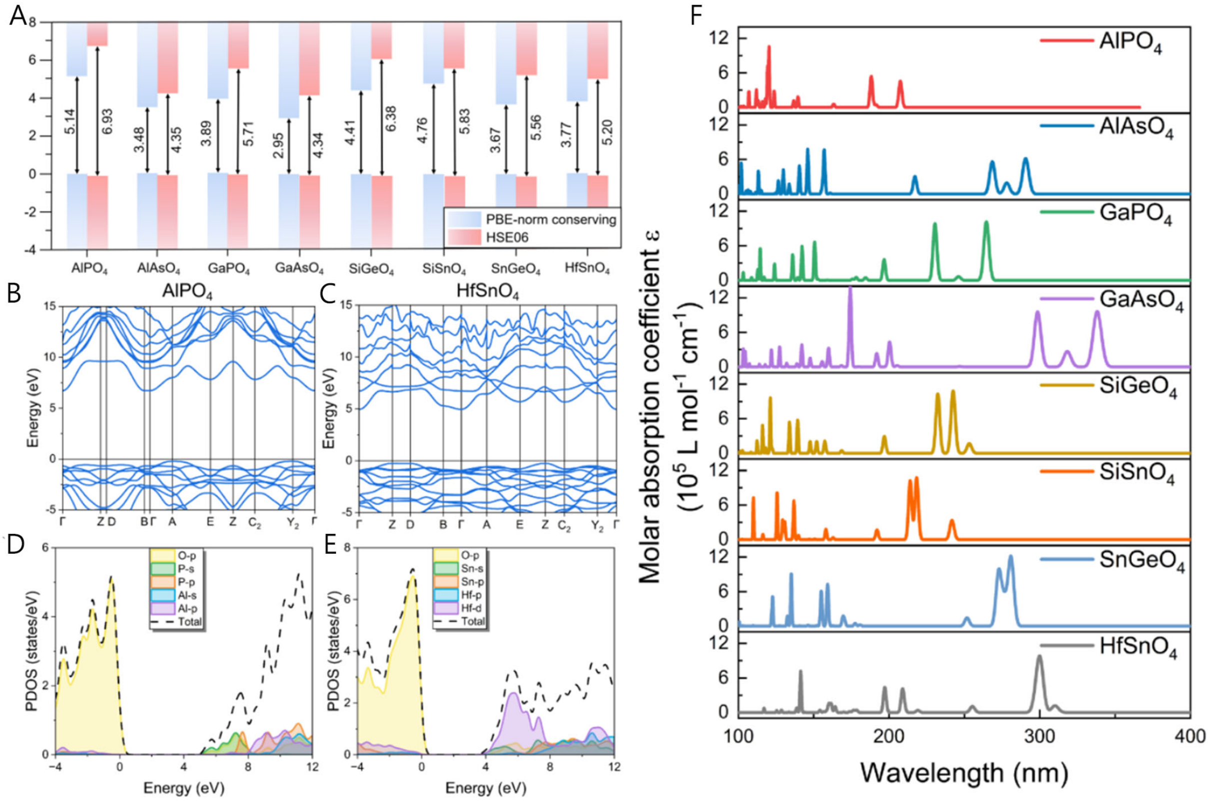

Figure 5. Electronic and optical properties of selected UWBG semiconductors. A, Band edges of eight selected ultrawide bandgap semiconductors. Corresponding bandgaps are labeled. HSE06 calculated band structures of B, $AlPO_4$ and C, $HfSnO_4$. PDOS of D, $AlPO_4$ and E, $HfSnO_4$. F, UV-Vis spectrums of selected ultrawide bandgap semiconductors. UWBG: Ultrawide bandgap; HSE: Heyd-Scuseria-Ernzerhof; PDOS: partial density of states.

(DUV) devices, such as photodetectors, light emitters, and lithography machines [85,86].

CONCLUSIONS

In summary, we developed the MELRSNet, a framework that integrates ML classification models, regression models, and DFT calculations, leveraging structural and elemental inputs for the discovery of novel UWBG oxides. The MELRSNet demonstrates remarkable accuracy in distinguishing UWBG from non-UWBG compounds, achieving an F1 score of 0.96 on the testing set through its stacking classification model. Furthermore, the regression model performed a satisfying prediction of the bandgaps of UWBG materials. SHAP analysis enhances the explainability of “black-box” ensemble learning model, providing a more comprehensive understanding between the features and the output. DFT calculations were employed to validate and refine the optimized structure, stability, electronic, and optical properties of selected materials. The applicability of MELRSNet was first validated through the study on the AFLOW dataset with no overlapped entry with the training dataset. The potential of MELRSNet for the discovery of novel UWBG oxides has been validated by applying the model to ternary oxide candidates derived from the monoclinic HfO_2 structure with the formula XYO_4 and space group *P*2/*m*, where we successfully identified eight novel ternary oxides with UWBG. These materials exhibit desirable bandgaps, high stability, and promising UV light absorption. MELRSNet offers an efficient balance between accuracy and resource consumption. In contrast to relying solely on DFT calculations, MELRSNet algorithms enable rapid prediction of material properties, serving as a preliminary screening tool for novel materials. This approach significantly reduces the computational demands associated with state-of-the-art first-principles methods. MELRSNet overcomes a significant hurdle faced by traditional trial-and-error methods, presenting vast potential in materials design and discovery.

Moreover, MELRSNet addresses the common issue of data imbalance through the application of SMOTE-ENN, enhancing classification accuracy. The use of model blending (stacking) combines the strengths of various base models, significantly boosting classification performance. The LightGBM regression model achieves comparable accuracy to PBE functional bandgap calculations but with much greater speed. Importantly, our dataset focuses on UWBG samples, providing a novel avenue for expanding the UWBG database. There exist several limitations of the current framework, such as the limited quality of training data due to the lack of experimental dataset, insufficient consideration of the properties needed for a material to be UWBG semiconductor. However, with the development of benchmark datasets and state-of-the-art ML models, many strategies could be employed to further enhance the performance of our MELRSNet framework. For example, the model ability is anticipated to improve if replacing the dataset with experimental or more precise computational data, employing strategies such as transfer learning to refine GGA-PBE data, and incorporating additional properties such as the stability and carrier mobility and other important factors for UWBG semiconductors in the ML evaluation steps to comprehensively characterize material structure and properties. Although specific ML models and datasets would change according to the different application demands, our workflow provides a universal solution in the field of materials discovery and properties prediction.

DECLARATIONS

Authors' contributions

Conceptualization: Luo, B.; Song, H.
Software:Zhang, Z. (Zhesi Zhang); Ji, Y.; Zhang, Z. (Zili Zhang)
Investigation: Li, X.; Zhang, Z. (Zili Zhang)
Resources: Wu, Y.; Li, H.
Data Curation: Cai, Z.; Zhang, J.
Writing - Original Draft: Zhang, Z. (Zhesi Zhang)
Supervision:Luo, B.; Song, H.
Writing - Review & Editing:: Luo, B.; Song, H.; Ji, Y.; Cui, Y.
Visualization: Zhang, Z. (Zhesi Zhang); Cui, Y.

Availability of data and materials

The raw data supporting the findings of this study are available within this Article and its Supplementary Materials. Further data is available from the corresponding authors upon reasonable request.

Financial support and sponsorship

This research was supported by the National Natural Science Foundation of China (Grant No. 12135019 and No. 52202154), the 2115 Talent Development Program of China Agricultural University and the High-performance Computing Platform of China Agricultural University. Bingcheng Luo acknowledges support from the National Supercomputer Center in Tianjin, and the optimization calculations were performed on the Tianhe new generation supercomputer.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Zhou, F.; Gong, H.; Xiao, M.; et al. An avalanche-and-surge robust ultrawide-bandgap heterojunction for power electronics. *Nat. Commun.* **2023**, *14*, 4459. DOI PubMed PMC
2. Kim, H.; Uddin, S. Z.; Lien, D. H.; et al. Actively variable-spectrum optoelectronics with black phosphorus. *Nature* **2021**, *596*, 232-7. DOI
3. Luo, L.; Huang, Y.; Cheng, K.; et al. MXene-GaN van der Waals metal-semiconductor junctions for high performance multiple quantum well photodetectors. *Light. Sci. Appl.* **2021**, *10*, 177. DOI PubMed PMC
4. Makita, T.; Kumagai, S.; Kumamoto, A.; et al. High-performance, semiconducting membrane composed of ultrathin, single-crystal organic semiconductors. *Proc. Natl. Acad. Sci. USA.* **2020**, *117*, 80-5. DOI PubMed PMC
5. Nguyen, T. K.; Barton, M.; Ashok, A.; et al. Wide bandgap semiconductor nanomembranes as a long-term biointerface for flexible, implanted neuromodulator. *Proc. Natl. Acad. Sci. USA.* **2022**, *119*, e2203287119. DOI PubMed PMC
6. Park, S. H.; Yuan, G.; Chen, D.; et al. Wide bandgap III-nitride nanomembranes for optoelectronic applications. *Nano. Lett.* **2014**, *14*, 4293-8. DOI
7. Shi, J.; Zhang, J.; Yang, L.; Qu, M.; Qi, D. C.; Zhang, K. H. L. Wide bandgap oxide semiconductors: from materials physics to optoelectronic devices. *Adv. Mater.* **2021**, *33*, e2006230. DOI
8. Wang, Y.; Xu, W.; Fu, L.; et al. Realization of robust and ambient-stable room-temperature ferromagnetism in wide bandgap semiconductor 2D carbon nitride sheets. *ACS. Appl. Mater. Interfaces.* **2023**, *15*, 54797-807. DOI
9. Tsao, J. Y.; Chowdhury, S.; Hollis, M. A.; et al. Ultrawide-bandgap semiconductors: research opportunities and challenges. *Adv. Elect. Materials.* **2018**, *4*, 1600501. DOI
10. Zhang, J.; Dong, P.; Dang, K.; et al. Ultra-wide bandgap semiconductor Ga_2O_3 power diodes. *Nat. Commun.* **2022**, *13*, 3900. DOI PubMed PMC
11. Xie, C.; Lu, X.; Tong, X.; et al. Recent progress in solar-blind deep-ultraviolet photodetectors based on inorganic ultrawide bandgap semiconductors. *Adv. Funct. Materials.* **2019**, *29*, 1806006. DOI
12. Chen, Y.; Liu, K.; Liu, J.; et al. Growth of 2D GaN single crystals on liquid metals. *J. Am. Chem. Soc.* **2018**, *140*, 16392-5. DOI
13. Razeghi, M. Short-wavelength solar-blind detectors-status, prospects, and markets. *Proc. IEEE.* **2002**, *90*, 1006-14. DOI
14. Wang, J.; Xie, N.; Xu, F.; et al. Group-III nitride heteroepitaxial films approaching bulk-class quality. *Nat. Mater.* **2023**, *22*, 853-9. DOI
15. Ohtomo, A.; Kawasaki, M.; Koida, T.; et al. Mg x $Zn_{1-x}O$ as a II–VI widegap semiconductor alloy. *Appl. Phys. Lett.* **1998**, *72*, 2466-8. DOI
16. Han, D.; Liu, K.; Yang, J.; et al. Performance enhancement of a p-Si/n-$ZnGa_2O_4$ heterojunction solar-blind UV photodetector through interface engineering. *J. Mater. Chem. C.* **2021**, *9*, 10013-9. DOI
17. Yang, Y.; Liu, S.; Wang, X.; et al. Polarization-sensitive ultraviolet photodetection of anisotropic 2D GeS_2. *Adv. Funct. Materials.* **2019**, *29*, 1900411. DOI
18. Zheng, Y.; Tang, X.; Wang, W.; Jin, L.; Li, G. Large-size ultrathin α-Ga_2S_3 nanosheets toward high-performance photodetection. *Adv. Funct. Materials.* **2021**, *31*, 2008307. DOI
19. Zhou, N.; Gan, L.; Yang, R.; et al. Nonlayered two-dimensional defective semiconductor γ-Ga_2S_3 toward broadband photodetection. *ACS. Nano.* **2019**, *13*, 6297-307. DOI PubMed
20. Yan, Y.; Yang, J.; Du, J.; et al. Cross-substitution promoted ultrawide bandgap up to 4.5 eV in a 2D semiconductor: gallium thiophosphate. *Adv. Mater.* **2021**, *33*, e2008761. DOI PubMed
21. Maldovan, M.; Thomas, E. L. Diamond-structured photonic crystals. *Nat. Mater.* **2004**, *3*, 593-600. DOI PubMed
22. Li, L.; Wu, M. Binary compound bilayer and multilayer with vertical polarizations: two-dimensional ferroelectrics, multiferroics, and nanogenerators. *ACS. Nano.* **2017**, *11*, 6382-8. DOI
23. Liu, X.; Geng, X.; Liu, H.; et al. Recent progress and applications of HfO_2-based ferroelectric memory. *Tsinghua. Sci. Technol.* **2023**, *28*, 221-9. DOI
24. Schroeder, U.; Park, M. H.; Mikolajick, T.; Hwang, C. S. The fundamentals and applications of ferroelectric HfO_2. *Nat. Rev. Mater.* **2022**, *7*, 653-69. DOI
25. Garrity, E. M.; Lee, C.; Gorai, P.; Tellekamp, M. B.; Zakutayev, A.; Stevanović, V. Computational identification of ternary wide-band-gap oxides for high-power electronics. *PRX. Energy.* **2022**, 1. DOI

26. Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865-8. DOI PubMed
27. Gorai, P.; Krasikov, D.; Grover, S.; Xiong, G.; Metzger, W. K.; Stevanović, V. A search for new back contacts for CdTe solar cells. *Sci. Adv.* **2023**, *9*, eade3761. DOI PubMed PMC
28. Perdew, J. P.; Levy, M. Physical content of the exact kohn-sham orbital energies: band gaps and derivative discontinuities. *Phys. Rev. Lett.* **1983**, *51*, 1884-7. DOI
29. Heyd, J.; Peralta, J. E.; Scuseria, G. E.; Martin, R. L. Energy band gaps and lattice parameters evaluated with the Heyd-Scuseria-Ernzerhof screened hybrid functional. *J. Chem. Phys.* **2005**, *123*, 174101. DOI PubMed
30. Masood, H.; Sirojan, T.; Toe, C. Y.; et al. Enhancing prediction accuracy of physical band gaps in semiconductor materials. *Cell. Rep. Phys. Sci.* **2023**, *4*, 101555. DOI
31. Rosen, A. S.; Iyer, S. M.; Ray, D.; et al. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578-97. DOI
32. Wang, X.; Huang, Y.; Xie, X.; et al. Bayesian-optimization-assisted discovery of stereoselective aluminum complexes for ring-opening polymerization of racemic lactide. *Nat. Commun.* **2023**, *14*, 3647. DOI PubMed PMC
33. Jain, A.; Ong, S. P.; Hautier, G.; et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL. Materials.* **2013**, *1*, 011002. DOI
34. Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM.* **2013**, *65*, 1501-9. DOI
35. Curtarolo, S.; Setyawan, W.; Wang, S.; et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227-35. DOI
36. Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M. Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **2020**, *11*, 4428. DOI PubMed PMC
37. He, L.; Li, Y.; Torrent, D.; Zhuang, X.; Rabczuk, T.; Jin, Y. Machine learning assisted intelligent design of meta structures: a review. *Microstructures* **2023**, *3*, 2023037. DOI
38. Li, C.; Hao, H.; Xu, B.; et al. A progressive learning method for predicting the band gap of ABO $_3$ perovskites using an instrumental variable. *J. Mater. Chem. C.* **2020**, *8*, 3127-36. DOI
39. Lu, S.; Zhou, Q.; Ma, L.; Guo, Y.; Wang, J. Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning. *Small. Methods.* **2019**, *3*, 1900360. DOI
40. Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 3405. DOI PubMed PMC
41. Park, H.; Mall, R.; Ali, A.; Sanvito, S.; Bensmail, H.; El-mellouhi, F. Importance of structural deformation features in the prediction of hybrid perovskite bandgaps. *Comput. Mater. Sci.* **2020**, *184*, 109858. DOI
42. Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj. Comput. Mater.* **2020**, *6*, 352. DOI
43. Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B.* **2016**, 93. DOI
44. Shi, Z.; Tsymbalov, E.; Dao, M.; Suresh, S.; Shapeev, A.; Li, J. Deep elastic strain engineering of bandgap through machine learning. *Proc. Natl. Acad. Sci. USA.* **2019**, *116*, 4117-22. DOI PubMed PMC
45. Zhuo, Y.; Mansouri, T. A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **2018**, *9*, 1668-73. DOI PubMed
46. Dau, M. T.; Al, K. M.; Michon, A.; Reserbat-Plantey, A.; Vézian, S.; Boucaud, P. Descriptor engineering in machine learning regression of electronic structure properties for 2D materials. *Sci. Rep.* **2023**, *13*, 5426. DOI PubMed PMC
47. Fung, V.; Zhang, J.; Hu, G.; Ganesh, P.; Sumpter, B. G. Inverse design of two-dimensional materials with invertible neural networks. *npj. Comput. Mater.* **2021**, *7*, 670. DOI
48. Lu, S.; Zhou, Q.; Guo, Y.; Zhang, Y.; Wu, Y.; Wang, J. Coupling a crystal graph multilayer descriptor to active learning for rapid discovery of 2D ferromagnetic semiconductors/half-metals/metals. *Adv. Mater.* **2020**, *32*, e2002658. DOI PubMed
49. Zhuo, Y.; Mansouri, T. A.; Oliynyk, A. O.; Duke, A. C.; Brgoch, J. Identifying an efficient, thermally robust inorganic phosphor host via machine learning. *Nat. Commun.* **2018**, *9*, 4377. DOI PubMed PMC
50. Wang, Z.; Yang, M.; Xie, X.; et al. Applications of machine learning in perovskite materials. *Adv. Compos. Hybrid. Mater.* **2022**, *5*, 2700-20. DOI
51. Shen, H.; Wu, J.; Chen, Z.; et al. First-principles study combined with interpretable machine-learning models of bayesian optimization for the design of ultrawide bandgap double perovskites. *J. Phys. Chem. C.* **2023**, *127*, 21410-22. DOI
52. Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD. Explor. Newsl.* **2004**, *6*, 20-9. DOI
53. Chawla, N. V. ; Bowyer KW.; Hall LO.;Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321-57. DOI
54. Pedregosa, F. ; Varoquaux G.; Gramfort A, et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **2011**, *12*, 2825-30. DOI
55. Wolpert, D. H. Stacked generalization. *Neural. Networks.* **1992**, *5*, 241-59. DOI

56. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.46*, 389-422. DOI
57. Ke, G. ; Meng Q.; Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc.; 2017, p. 3149-57. DOI
58. Clark, S. J.; Segall, M. D.; Pickard, C. J.; et al. First principles methods using CASTEP. *Z. Kristallogr. Cryst. Mater.* **2005**, *220*, 567-70. DOI
59. Payne, M. C.; Teter, M. P.; Allan, D. C.; Arias, T. A.; Joannopoulos, J. D. Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* **1992**, *64*, 1045-97. DOI
60. Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B. Condens. Matter.* **1996**, *54*, 11169-86. DOI PubMed
61. Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B.* **1999**, *59*, 1758-75. DOI
62. Pfrommer, B. G.; Côté, M.; Louie, S. G.; Cohen, M. L. Relaxation of crystals with the quasi-newton method. *J. Comput. Phys.* **1997**, *131*, 233-40. DOI
63. Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B.* **1976**, *13*, 5188-92. DOI
64. Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207-15. DOI
65. Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, *125*, 224106. DOI PubMed
66. Refson, K.; Tulip, P. R.; Clark, S. J. Variational density-functional perturbation theory for dielectrics and lattice dynamics. *Phys. Rev. B.* **2006**, 73. DOI
67. Baroni, S.; de, G. S.; Dal, C. A.; Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **2001**, *73*, 515-62. DOI
68. Momma, K.; Izumi, F. *VESTA 3* for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **2011**, *44*, 1272-6. DOI
69. Leeuwen R. Mapping from densities to potentials in time-dependent density-functional theory. *Phys. Rev. Lett.* **1999**, *82*, 3863-6. DOI
70. Kühne, T. D.; Iannuzzi, M.; Del, B. M.; et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103. DOI
71. Lu, T.; Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012**, *33*, 580-92. DOI
72. Jain, A.; Hautier, G.; Moore, C. J.; et al. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295-310. DOI
73. Borlido, P.; Aull, T.; Huran, A. W.; Tran, F.; Marques, M. A. L.; Botti, S. Large-scale benchmark of exchange-correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory. Comput.* **2019**, *15*, 5069-79. DOI PubMed PMC
74. Yang, J.; Liu, K.; Chen, X.; Shen, D. Recent advances in optoelectronic and microelectronic devices based on ultrawide-bandgap semiconductors. *Prog. Quantum.* **2022**, *83*, 100397. DOI
75. Yin, Y.; Wang, A.; Sun, Z.; Xin, C.; Jin, G. Machine learning regression model for predicting the band gap of multi-elements nonlinear optical crystals. *Comput. Mater. Sci.* **2024**, *242*, 113109. DOI
76. Chen, X.; Lu, S.; Chen, Q.; Zhou, Q.; Wang, J. From bulk effective mass to 2D carrier mobility accurate prediction via adversarial transfer learning. *Nat. Commun.* **2024**, *15*, 5391. DOI PubMed PMC
77. Biswas, M.; Nishinaka, H. Thermodynamically metastable α-, ε- (or κ-), and γ-Ga_2O_3: From material growth to device applications. *APL. Materials.* **2022**, *10*, 060701. DOI
78. Gladkikh, V.; Kim, D. Y.; Hajibabaei, A.; Jana, A.; Myung, C. W.; Kim, K. S. Machine learning for predicting the band gaps of ABX_3 perovskites from elemental properties. *J. Phys. Chem. C.* **2020**, *124*, 8905-18. DOI
79. Fowler, W. B. Influence of electronic polarization on the optical properties of insulators. *Phys. Rev.* **1966**, *151*, 657-67. DOI
80. Curtarolo, S.; Setyawan, W.; Hart, G. L.; et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218-26. DOI
81. Zagorac, D.; Müller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **2019**, *52*, 918-25. DOI PubMed PMC
82. Becke, A. D.; Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. *J. Chem. Phys.* **1990**, *92*, 5397-403. DOI
83. Su, J.; Guo, R.; Lin, Z.; et al. Unusual electronic and optical properties of two-dimensional Ga_2O_3 predicted by density functional theory. *J. Phys. Chem. C.* **2018**, *122*, 24592-9. DOI
84. Anam, B.; Gaston, N. Structural, thermal, and electronic properties of two-dimensional gallium oxide (β-Ga_2O_3) from first-principles design. *Chemphyschem* **2021**, *22*, 2362-70. DOI
85. Kubota, Y.; Watanabe, K.; Tsuda, O.; Taniguchi, T. Deep ultraviolet light-emitting hexagonal boron nitride synthesized at atmospheric pressure. *Science* **2007**, *317*, 932-4. DOI PubMed
86. Wang, Y.; Meng, J.; Tian, Y.; et al. Deep ultraviolet photodetectors based on carbon-doped two-dimensional hexagonal boron nitride. *ACS. Appl. Mater. Interfaces.* **2020**, *12*, 27361-7. DOI