Predicting band gaps of ternary oxides for phosphor hosts from machine learning

YUEYU ZHOU, JING GAO, YITING GUI, JUN WEN*, YAN WANG, QUANJIN LIU, GUISHENG JIANG, XIAOXIAO HUANG, QIANG WANG, CHENLONG WEI
School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University, Anqing 246133, China

The prediction of the properties of inorganic compounds by data-driven machine learning methods has gradually become a research hotspot in the field of materials science. In this work, the machine learning models of the least absolute shrinkage and selection operator (Lasso), kernel ridge regression (KRR), Gaussian process regression (GPR), random forests regression (RFR), support vector regression (SVR) and gradient boosting regression (GBR) were utilized to predict band gaps of ternary oxides for phosphor hosts. The results show that the GBR is a robust and feasible model with higher performance. Besides, the importance of each feature is analyzed quantificationally based on the GBR model. It indicates that two features (i.e., the average of molar heat capacity and the range of metallic valence) play a great role in affecting the predictive performance of band gaps. Besides, the Shapley additive explanation (SHAP) is used to elaborate the results from the GBR model. This work not only demonstrates the feasibility of machine learning to predict band gaps based only on the chemical composition but also contributes to the prediction of the other properties of inorganic materials.

(Received May 2, 2022; accepted December 6, 2022)

Keywords: Machine learning methods, Band gaps, Inorganic compounds, Materials science

1. Introduction

Exploring new materials with high performance for the specific application is an important subject in materials science. The research methods of materials science can be roughly divided into experimental and computational (or theoretical) methods. Experimental methods mainly rely on the accumulation of the experience of researchers. Although they are relatively intuitive, it may sometimes be faced with several problems, such as high preparation costs, long research and development cycle and low efficiency. In recent years, first-principles calculations, molecular dynamics simulation, finite element simulation and other calculation methods have achieved remarkable results in semiconductor materials [1], rare earth luminescent materials [2, 3], new energy materials [4] and so on [5, 6]. However, the demand for new materials is increasing along with the cycle of research getting shorter, so the shortcoming of high computing cost for high-throughput computing methods is becoming increasingly prominent. The traditional calculation methods may be difficult to meet the efficient screening and discovery of new materials.

Thanks to the rapid development of artificial intelligence and the improvement of hardware facilities, the application of machine learning in the study on materials has become possible. Based on the massive calculated and experimental results, the machine learning methods can quickly learn useful information and achieve the prediction of material properties (such as the structural and mechanical properties of alloys [7-11]). Ahmad et al. [12] employed the support vectormachine (SVM), random forest (RF), Adaboost and k-nearest neighbor (KNN) models to achieve the prediction of the shear strength of rockfill materials. Kauwe et al. [13] confirmed the feasibility of the prediction for the heat capacity of solid inorganics by using the linear regression (LR), support vector regression (SVR), and random forest regression (RFR). Formation energies of oxygen vacancies in metal oxide materials can also be predicted through the machine learning model [14]. Machine learning technologies has also allowed researchers to predict the glass formation ability [15, 16], Debye temperature [17] and energy-level structures [18].

As well known, inorganic phosphors have a wide range of applications in solid-state lighting and display [19, 20]. They are composed of hosts (such as oxides, nitrides and halides) and luminescent centers (such as rare-earth and transition metal ions). The screening for the optimum host is important for finding high-performance phosphors. It is noted that the band gap is an important property, corresponding to the difference of the energy between the bottom of the conduction band and the top of the valence band of semiconductors or insulators [21, 22]. It determines band structures of materials and influences their electronic structures and optical properties. In this work, six important machine learning models, namely the least absolute shrinkage and selection operator (Lasso), kernel ridge regression (KRR), Gaussian process regression (GPR), RFR, SVR and gradient boosting regression (GBR) are developed to predict band gaps of phosphor hosts. 773 potential hosts and 129 features were used as the input of the models. By comparing the output results of the models, it is found that the the GBR gives better predictive performance of band gaps of the selected materials.

Furthermore, the importance score of the features based on the GBR model is analyzed deeply and the possible underlying reasons are explained reasonably. A meticulous understanding about the band gap of materials may promote the process of screening out stable and effective phosphor hosts. This work provides useful help for the design and synthesis of phosphors.

2. Calculation methods

Machine learning was proposed by Samuel [23] in 1959, and is an inter-discipline involving the probability theory, statistics and computer science. The principle of machine learning is that the model make the accurate judge and decision through learning the past experience or data. A simple workflow of machine learning for predicting material properties is shown in Fig. 1, which contains the data collection and preprocessing, model building and evaluation. In this work, the samples with the target material property (i.e., the band gap) for specific types of compounds are collected from the open-access database. The machine learning models are trained by using featurized data, and further evaluated by the performance metrics.

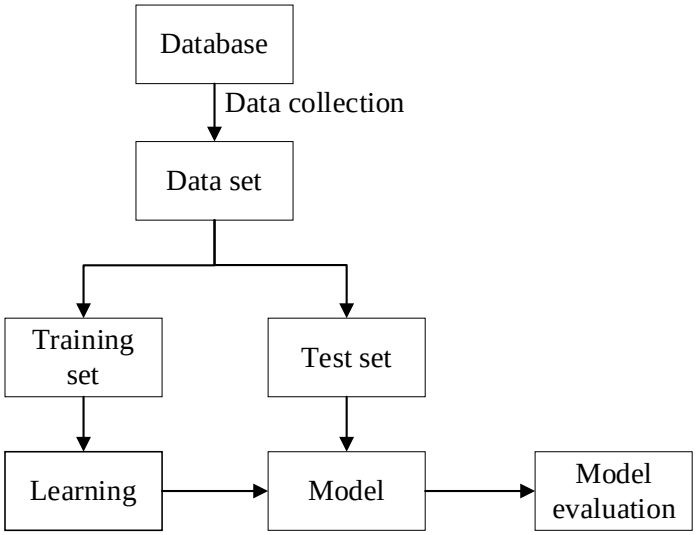

Fig. 1. The workflow of the prediction model of material properties via machine learning

2.1. Data collection and preprocessing

The aim of this work is to predict band gaps of ternary inorganic oxides for phosphor hosts. The data collection is firstly carried out in the Materials Project database [24]. It is an open-access database that contains lots of inorganic materials with relevant properties information. A total of 2032 $A_\alpha B_\beta O_\gamma$-type inorganic compounds (α, β and γ are positive integers) were collected as the raw data set, and the elements of *A* and *B* are illustrated in Fig. 2. It is noted that the oxides with relatively small band gaps (<3.0 eV) are eliminated, considering that a wide band gap is critical for the hosts of phosphors [25]. Therefore, the final data set in this work consists of the values of the band gaps for 733 oxide samples. Then, the data set above was divided randomly into two sets: the training (696 samples) and test sets (37 samples). The feature generation is a very important step that directly affects the performance of the model. 43 different variables of the elements [26] listed in Table 1 (e.g., the period, electronegativity, atomic radius, Mendeleev number, boiling point etc.) are extended to 129 features via the operations of the weighted average, the sum and the max-min, by using Python Materials Genomics (Pymatgen) library [27]. Then, the features and the target property are normalized by using the StandardScaler module in the machine learning library Scikit-learn (Sklearn) [28] to guarantee the normal distribution with 0 mean and 1 variance for each data. The selected models are trained by using different numbers of features, according to the calculated Pearson correlation coefficients between the features. It is found that the reduction of the number of features hardly improves the accuracy of the model, so all 129 features are included for the train of the model.

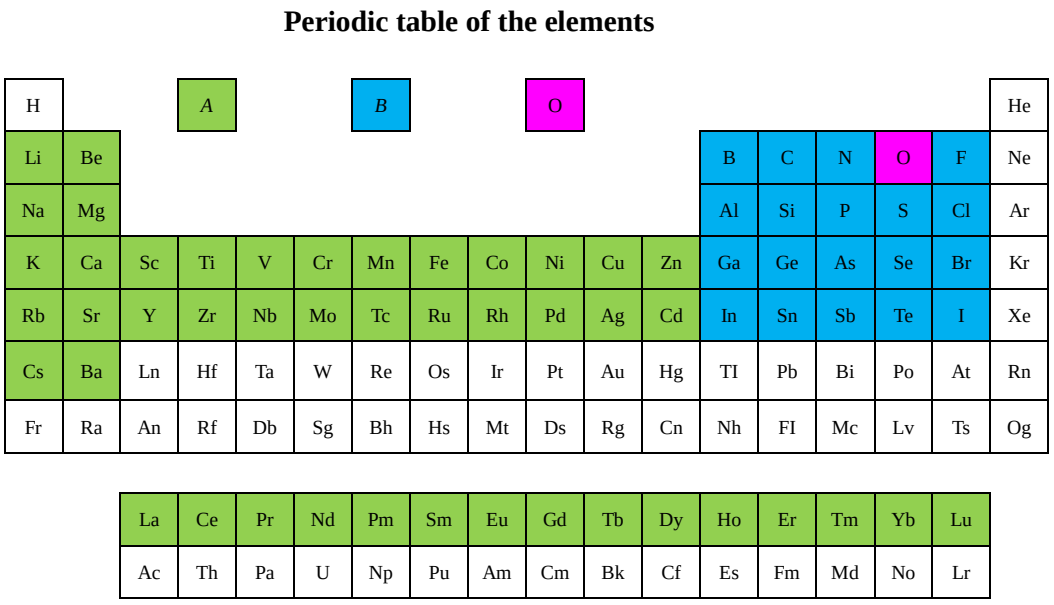

Fig. 2. The element composition of phosphor hosts ($A_\alpha B_\beta O_\gamma$) considered in this work. 2032 $A_\alpha B_\beta O_\gamma$-type inorganic compounds and their band gap values are collected from Materials Project database considering the specific rules of composition (color online)

2.2. Machine learning model

2.2.1. Lasso

The regression models (i.e., the Lasso, KRR, GPR, RFR, SVR and GBR) are utilized in this work, in consideration that the predicted outcomes (band gaps) in the supervised learning are continuous variables. The Lasso model has been proposed by Tibshirani in 1996 [29]. In this model, there is an upper bound for the sum of the absolute values of some model parameters. It is noted that a regularization process of the penalizing coefficient of regression variables is applied, in order to achieve that goal. The *L*1 regularization method was used herein, and its objective function is expressed as follows:

$$\min_{w} \frac{1}{2n_{\text{samples}}} \|wx - y\|_2^2 + \lambda \|w\|_1 \tag{1}$$

where $\|wx - y\|_2^2$ represents the sum of squared errors between the predicted and actual values. λ is a constant and $\|w\|_1$ is *L*1-norm of the coefficient vector.

2.2.2. KRR

The KRR is a non-linear prediction model combining nuclear technique and ridge regression [30]. It uses the kernel function to map the original non-linear data to a kernel space, so that the data can be linearly separable in the kernel space, and then the data can be linear ridge regression in the kernel space. Compared with Lasso, *L*2 regularization was used in KRR, and its loss function is expressed as follows:

$$\min_{w} \|wx - y\|_2^2 + \lambda \|w\|_2^2 \tag{2}$$

where $\|w\|_2^2$ is *L*2-norm of the coefficient vector.

2.2.3. GPR

Gaussian process is a supervised learning method, which is used to solve classification and regression problems. It refers to a stochastic process in which any finite random variables and their subsets obey a Gaussian distribution. The GPR is a Gaussian process to achieve the purpose of the regression [31]. The covariance function and means function are key variables, which uniquely determine the properties of Gaussian regression. The GPR is a prior-based regression analysis model based on Bayesian and statistical theories, and shows good effects in complex regression problems, such as the high dimension, small sample size and nonlinearity.

2.2.4. RFR

The RF is introduced by Breiman in the early 2000s [32]. It is an ensemble machine learning model, which is based on a large number of independent decision trees. Their output results are then summarized, in order to achieve better performance than the individual model. In general, the RF is not only used for the classification (RFC) but also the regression (RFR) purposes. In this work, the RFR is adopted to predict band gaps. The construction procedure is summarized in following:

Step (1): The training set data is divided into a series of

sub-datasets with *N* samples by the bootstrap statistical method. The *N* samples with *M* features are selected one by one (with the return) and used to train a decision tree, being regarded as the node sample.

Step (2): The samples of nodes are split in the process of training. The *m* features are selected randomly from the *M* features ($m << M$). Then, one of the m features is selected as the sample node according to a certain strategy.

Step (3): Each sample node is split according to Step (2), until that it cannot be split.

Step (4): Many decision trees are built according to Steps (1)–(3), achieving the goal to construct the RF. Final results are obtained by averaging the predicted values of all decision trees.

2.2.5. SVR

Boser et al. proposed the theory of support vector machine (SVM) [33], which can be applied to solve both the classification and regression problems. Considering that some data are not completely linearly separable, the conception of soft margin was introduced as follows:

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{Subject } y_i(wx_i + b) \geq 1 - \xi_i \quad (3)$$

where $C >> 0$ is a penalty parameter, each ξ_i is related to the distance between the object i and the respective margin hyperplane. ω and b is the normal vector and the bias of the hyperplane, respectively. The dot product of the high dimensional space is replaced by the kernel function operation of the low dimensional space, in order to deal with the curse of dimensionality for the high dimensional space. There are many kinds of kernel functions, such as linear and radial basis function (RBF). In this study, the RBF is adopted, as shown in the following equation:

$$k(x, x_i) = \exp\left(-\frac{\| x - x_i \|^2}{2\sigma^2}\right) \quad (4)$$

where x_i is the center of kernel function and σ represents the width parameter of function.

2.2.6. GBR

As one of the models closing to real distribution fitting in traditional machine learning models, the Gradient Boosting Decision Tree (GBDT) is an ensemble model used for classification and regression [34, 35]. In the GBDT, the accuracy of final regression results is improved by means of adopting an additive model (a linear combination of basis function), which can constantly reduce the residual in the training process. A weak estimator is produced and then trained, on the basis of the residual error of the prior estimator in each of the iterations. It should be noted that the GBDT is particularly good at handling dense numerical features, since the features with the largest statistical information will be picked to build the trees. The regression model of the GBDT (i.e., the GBR) was used in this work.

2.3. Performance metrics

The performance of a machine learning model on the test set is particularly important, as it directly reflects the prediction ability of the model. The coefficient of determination (R^2), the mean absolute error (MAE) and the mean squared error (MSE) were selected to evaluate the predictive capacity of the model. Their values can be computed by the equations (5)–(7):

$$R^2 = 1 - \frac{\sum_{j=1}^{m}(T_j - P_j)^2}{\sum_{j=1}^{m}(T_j - \overline{T})^2} \quad (5)$$

$$\text{MAE} = \frac{1}{m}\sum_{j=1}^{m}\left|T_j - P_j\right| \quad (6)$$

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}(T_j - P_j)^2 \quad (7)$$

where m denotes the number of samples, T_j is the real values, $\overline{T}$ is the mean value of real values and P_j is the predicted value by using the fitted model.

3. Results and discussion

The training set is used to train and optimize the six classical machine learning models (Lasso, KRR, GPR, RFR, SVR and GBR), in order to better predict band gaps of phosphor hosts. The adjustable parameters of selected models (see Table 2) are respectively optimized according to the GridSearchCV method [28]. It should be noticed that GridSearchCV method, which combines grid search and cross validation, are used for parameter adjustment in this work. The parameters of each model are given in Table 2 and the other parameters for each model are default values. It should be noted that the parameters of each model are using the GridSearchCV method to set the search space, and the optimal parameters are sought under the cross validation method. Here, taking GBR as an example, the learning curves of Random_state and Max_depth parameters are shown in Figs. 3a-3b. They show the

implementation process of the parameter tuning, and reflect the trends between parameters and model performance in a certain search space. The optimal parameters are determined for model training through the learning curve of parameters. According to the trained models, the band gap prediction results of the test set are shown in Table 3. It is easy to find that the predicted band gaps of the majority of compounds are in a good agreement with DFT-calculated ones for most of the models (such as the GPR, RFR, SVR and GBR).

The predicted and DFT-calculated band gaps are demonstrated in Figs. 4a-4d. They intuitively reveal the degree of deviation between the predicted and DFT-calculated band gaps. The gray lines are related to ideal situations, that is, the predicted band gaps are equal to the DFT-calculated ones. The red lines represent the linear fitting results. It is found that the predictive error of the Lasso and KRR models are relatively large, along with most of the points deviating from the ideal line. The GPR, RFR and SVR models perform well, and the predictive performance of the GBR is the most excellent among six models, with the smallest angle between red and gray lines. These results also indicate that in the case of the optimal model parameters, different models have various predictive capacities although on the same test set.

The performance metrics R^2, MAE and MSE are usually used to evaluate the performance of the models (see Table 2). It is easy to find that the GBR model gets the maximum R^2, which reaches up to 0.822, and its MSE (0.094) is the smallest. The difference of predicted results among GPR, RFR and SVR models is small, along with that their R^2 values are 0.73, 0.756 and 0.775, respectively. However, Lasso and KRR models perform bad and just have the accuracy of 0.522 and 0.580, respectively. There are many factors that affect the performance of the models, such as the selection of the data set and features, as well as the parameter tuning. Herein, the reason for the bad performance of Lasso and KRR models may be that they are not sensitive enough to the nonlinear data.

In order to confirm the accuracy of predicted results, the above calculations are repeated for ten times under the condition that the parameters were unchanged. The mean values of R^2, MAE and MSE of ten tests for six models are calculated, and it is found that the GBR gets the best predictive performance. The number of times that the GBR performed best, reached up to five. It is proving that the GBR is superior to the other five models in this circumstance. Although the R^2 of the GBR is not high, it has the best performance in comparison with the other five models for the same test. The reason for this may be that, the parameters are the same in the process of conducting 10 tests for a certain model.

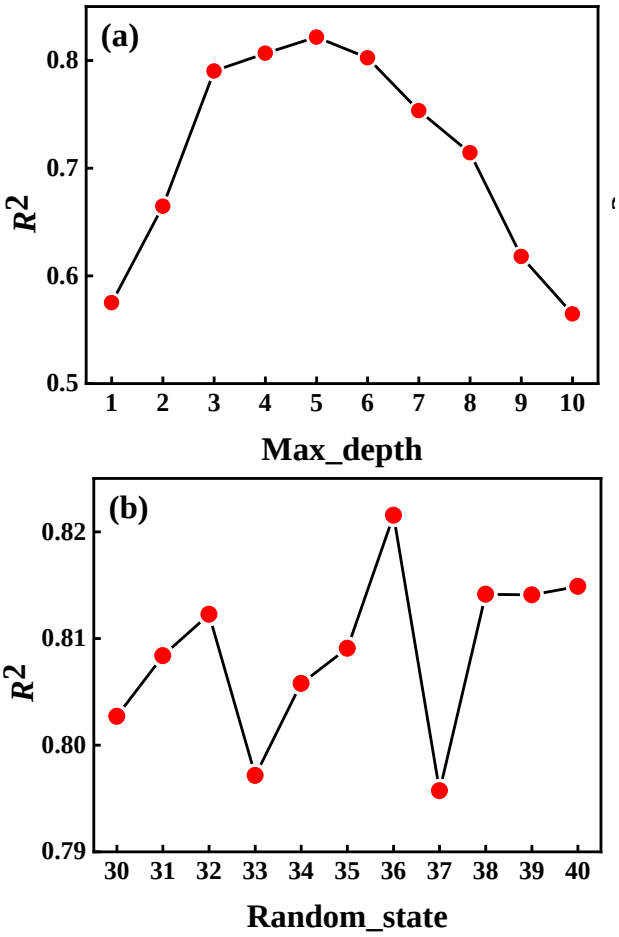

Fig. 3. The learning curves of the parameters of (a) Max_depth and (b) Random_state (color online)

Table 1. 43 different element variables used for the generation of features

Feature number	Element variables	Feature number	Element variables
1-3	First ionization potential	67-69	Period number
4-6	Allred-Rochow electronegativity	70-72	Zunger radii sums
7-9	Atomic number	73-75	Abs valence
10-12	Atomic radius	76-78	Family number
13-15	Atomic weight	79-81	Gilman number of valence electrons
16-18	Boiling point	82-84	Group number
19-21	Cohesive energy	85-87	Heat atomization
22-24	Covalent radius	88-90	Heat of fusion
25-27	Critical temperature	91-93	Heat of vaporization
28-30	Density	94-96	Ionic radius
31-33	Gordy electronegativity	97-99	Quantum number *l*
34-36	Melting point	100-102	Metallic valence
37-39	Mendeleev number	103-105	Number of valence electrons
40-42	Molar density	106-108	Number of outer shell electrons
43-45	Molar heat capacity	109-111	Polarizability
46-48	Nagle electronegativity	112-114	Specific heat
49-51	Number of unfilled d valence electrons	115-117	Thermal conductivity
52-54	Number of unfilled f valence electrons	118-120	Number of s valence electrons
55-57	Number of unfilled p valence electrons	121-123	Number of p valence electrons
58-60	Number of unfilled s valence electrons	114-126	Number of d valence electrons
61-63	Orbital radius	127-129	Number of f valence electrons
64-66	Pauling electronegativity		

Table 2. Optimized parameters and the evaluation metrics of band gap prediction employing regression models on the test set

Model	Parameters	R^2	MAE	MSE
Lasso	alpha = 0.0006, max_iter = 10000	0.522	0.422	0.252
KRR	kernel = “polynomial”, alpha = 0.68, coef0 = 3.99, degree = 3	0.580	0.385	0.222
GPR	kernel = “RBF”, alpha = 0.1, n_restarts_optimizer = 6	0.735	0.270	0.140
RFR	max_depth = 17	0.756	0.291	0.137
SVR	kernel = “rbf”, C = 5, epsilon = 0.25, gamma = 0.0188	0.775	0.264	0.119
GBR	n_estimators = 391, random_state = 36, learning_rate = 0.1, max_depth = 5	0.822	0.245	0.094

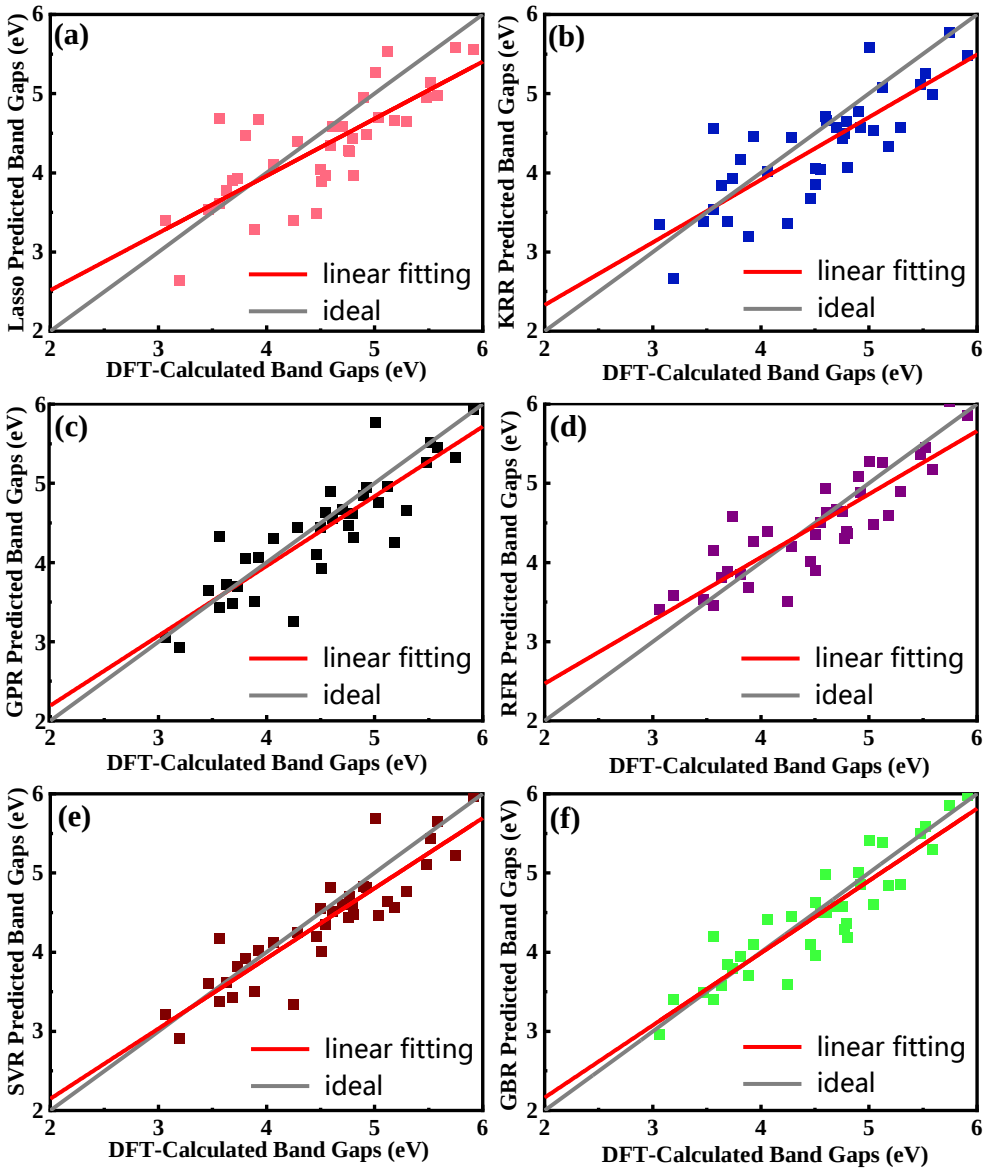

Fig. 4. Predicted band gaps of 37 compounds for the test set from (a) Lasso, (b) KRR, (c) GPR, (d) RFR, (e) SVR and (f) GBR models, in comparison with the DFT-calculated values (color online)

The GradientBoostingRegressor module from the Sklearn library was used in order to analyze the contribution of each feature to the prediction of band gaps. It should be noted that the feature importance score sum to 1.0. The higher importance score represents the deeper degree of contribution of the feature. The features with the importance score of larger than 0.75% are plotted in Fig. 5, according to the GBR model. One can find that the features "the average of molar heat capacity" and "the range of metallic valence" (with the highest two values of the feature importance) have a much more important effect on the prediction. The molar heat capacity is defined as the amount of heat needed, in order to make one mole of the substance to cause an increase of one unit in its temperature. It reflects the electron transition process from the top of the valence band to the bottom of the conduction band. Besides, the metallic valence of the elements represents the oxidation state, affecting the degree of electron excitation or relaxation. Both features indeed contribute to the prediction of band gaps. The importance score of the third- to sixth-ranked of features are in the range from 3.03% to 7.54%.

The Shapley additive explanation (SHAP) analysis [35] is further carried out, in order to explain the rationality of the importance of features derived from the GBR. According to the SHAP analysis, the feature importance scores for 37 samples of the test set from the GBR are obtained, as shown in Fig. 6.

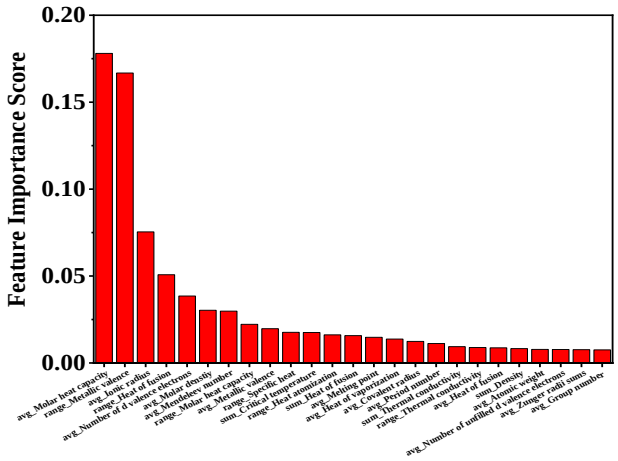

Fig. 5. The feature importance scores (representing the contribution to the prediction of band gaps) based on the GBR (color online)

It is found that the feature "the range of metallic valence" has the widest distribution, indicating the most important influence on the prediction of band gaps. In addition, the feature "the average of molar heat capacity" has the second most important influence. The top 2 important features determined from the SHAP analysis are consistent with those from the GBR.

Table 3. Band gaps (in units of eV) of the test set predicted from six machine learning models, along with the DFT-calculated values The relative errors of predicted results are also listed, in comparison with DFT-calculated values

Composition	Lasso	KRR	GPR	RFR	SVR	GBR	DFT
Al_4CdO_7	3.399(11%)	3.350(9%)	3.056(3%)	3.406(11%)	3.217(5%)	2.965(−3%)	3.064
$BaSi_2O_5$	4.440(−29%)	4.645(−3%)	4.624(5%)	4.394(−8%)	4.609(−4%)	4.364(−9%)	4.792
$Ca_8Si_5O_{18}$	3.972(−25%)	4.048(−11%)	4.632(5%)	4.503(−1%)	4.357(−4%)	4.503(−1%)	4.546
$Cs_4Si_3O_8$	3.776(−6%)	3.835(6%)	3.722(4%)	3.813(5%)	3.620(0.4%)	3.579(−1%)	3.633
$Dy_4B_6O_{15}$	4.949(−38%)	5.115(−7%)	5.260(5%)	5.362(−2%)	5.108(−7%)	5.495(0.3%)	5.476
Er_2SiO_5	4.285(−29%)	4.439(−7%)	4.466(4%)	4.642(−2%)	4.435(−7%)	4.585(−4%)	4.759
$Gd_4Ga_2O_9$	2.642(6%)	2.670(−16%)	2.934(3%)	3.589(12%)	2.905(−9%)	3.403(7%)	3.192
Ho_2SiO_5	4.585(−28%)	4.576(−3%)	4.679(5%)	4.672(−1%)	4.604(−2%)	4.585(−2%)	4.702
HoIO	3.545(−2%)	3.393(−2%)	3.644(4%)	3.539(2%)	3.603(4%)	3.496(1%)	3.465
$HoPO_4$	5.559(−43%)	5.487(−7%)	5.938(6%)	5.855(−1%)	5.974(1%)	5.979(1%)	5.912
$KClO_3$	4.977(−39%)	4.992(−11%)	5.452(5%)	5.183(−7%)	5.650(1%)	5.295(−5%)	5.582
KPO_3	4.660(−34%)	4.338(16%)	4.255(4%)	4.600(−11%)	4.571(−12%)	4.848(−6%)	5.183
$La_2P_4O_{13}$	4.482(−31%)	4.573(−7%)	4.944(5%)	4.884(−1%)	4.820(−2%)	4.860(−1%)	4.920
La_2SiO_5	4.039(−24%)	4.058(−10%)	4.447(4%)	4.357(−3%)	4.550(1%)	4.622(3%)	4.500
La_3PO_7	4.110(−16%)	4.023(−1%)	4.308(4%)	4.387(8%)	4.131(2%)	4.414(9%)	4.062
$La_4Ga_2O_9$	3.282(−13%)	3.191(−18%)	3.511(4%)	3.681(−5%)	3.502(−10%)	3.708(−5%)	3.885
$Lu_3Al_5O_{12}$	4.697(−33%)	4.531(−10%)	4.756(5%)	4.482(−11%)	4.469(−11%)	4.603(−9%)	5.036
MgB_4O_7	5.580(−41%)	5.773(1%)	5.332(5%)	6.038(5%)	5.224(−9%)	5.857(2%)	5.743
MgP_4O_{11}	5.530(−34%)	5.087(−1%)	4.958(5%)	5.271(3%)	4.638(−9%)	5.387(5%)	5.121
$MgSO_3$	5.268(−32%)	5.587(12%)	5.770(6%)	5.276(5%)	5.697(14%)	5.415(8%)	5.007
MoP_3O_9	3.404(−20%)	3.356(−21%)	3.256(3%)	3.508(−17%)	3.336(−21%)	3.592(−15%)	4.246
$NaAlO_2$	3.964(−29%)	4.067(−15%)	4.316(4%)	4.370(−9%)	4.475(−7%)	4.188(−13%)	4.805
Nd_3PO_7	4.404(−21%)	4.447(4%)	4.441(4%)	4.204(−2%)	4.257(−1%)	4.457(4%)	4.284
NdB_3O_6	5.140(−38%)	5.258(−5%)	5.516(6%)	5.458(−1%)	5.437(−1%)	5.586(1%)	5.519
$NdCl_3O_{12}$	4.952(−31%)	4.773(−3%)	4.850(5%)	5.084(4%)	4.836(−1%)	5.005(2%)	4.896

Composition	Lasso	KRR	GPR	RFR	SVR	GBR	DFT
NdFO	4.645(−36%)	4.567(−14%)	4.662(5%)	4.891(−8%)	4.771(−10%)	4.860(−8%)	5.293
PrBrO	3.485(−24%)	3.679(−18%)	4.101(4%)	4.015(−10%)	4.198(−6%)	4.101(−8%)	4.462
$RbBO_2$	4.675(−13%)	4.457(13%)	4.069(4%)	4.270(9%)	4.025(2%)	4.097(4%)	3.927
$ScAlO_3$	4.347(−26%)	4.707(2%)	4.901(5%)	4.938(8%)	4.822(5%)	4.985(9%)	4.593
$Sr_{10}Al_6O_{19}$	3.907(−8%)	3.380(−8%)	3.488(3%)	3.896(6%)	3.434(−7%)	3.843(4%)	3.688
Sr_2GeO_4	3.612(−5%)	3.538(−1%)	3.432(3%)	3.456(−3%)	3.380(−5%)	3.401(−5%)	3.566
SrN_2O_6	3.929(−9%)	3.934(5%)	3.706(4%)	4.584(23%)	3.817(2%)	3.799(2%)	3.733
Tm_2SiO_5	4.593(−26%)	4.661(1%)	4.558(5%)	4.630(1%)	4.519(−2%)	4.503(−2%)	4.607
YBrO	3.890(−25%)	3.853(−15%)	3.930(4%)	3.896(−14%)	4.006(−11%)	3.954(−12%)	4.507
ZnP_4O_{11}	4.267(−29%)	4.498(−6%)	4.624(5%)	4.302(−10%)	4.703(−1%)	4.286(−10%)	4.769
Zr_3SO_9	4.478(−11%)	4.164(9%)	4.055(4%)	3.849(1%)	3.915(3%)	3.949(4%)	3.809
ZrS_2O_8	4.686(−5%)	4.566(28%)	4.335(4%)	4.155(17%)	4.171(17%)	4.198 (18%)	3.564

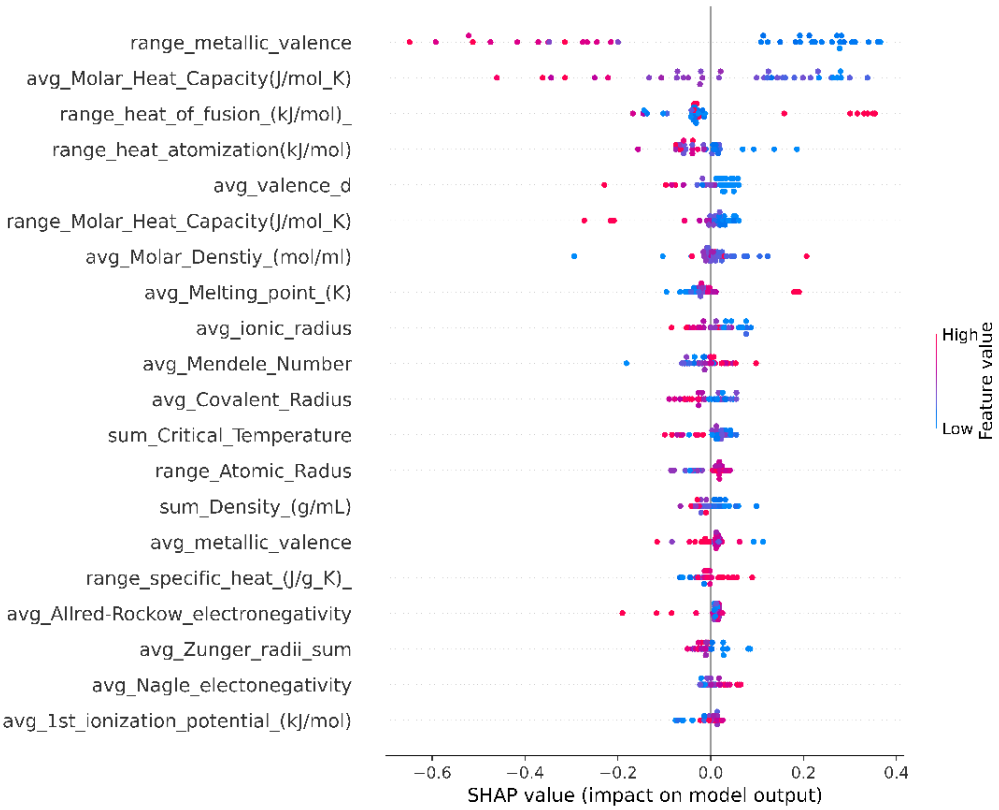

Fig. 6. Correlation between the band gaps and input features (color online)

4. Conclusions

In this work, six machine learning models (i.e., the Lasso, KRR, GPR, RFR, SVR and GBR) are adopted to predict band gaps of ternary oxides for phosphor hosts, according to the chemical composition properties of specific elements. 733 oxide compounds and their 129 features are chosen as the input of machine learning models. The predictive performances of the selected models are examined and compared. One can find that the GBR leads to the best performance, showing that it is a useful and relatively accurate machine learning model for the prediction of band gaps. The performance metrics R^2 of the GPR, RFR, SVR and GBR are in the range of 0.735-0.822, while the R^2 of the Lasso and KRR are much lower. Moreover, the results of the feature importance ranking reveal that "the average of molar heat capacity" and "the range of metallic valence" are important physical quantities affecting the performance of the band-gap prediction based on the GBR model. In addition, the conclusion is further verified by the SHAP analysis. This work shows that machine learning methods are powerful tools for predicting

band gaps of ternary oxides. Besides, the optimized models may also be used to predict other material properties for rapidly discovering new functional materials.

Acknowledgements

The authors gratefully appreciate the National Natural Science Foundation of China (Grants 11974022 and 11974315), University Natural Science Research Project of Anhui Province (Grants 2022AH030104, KJ2021A0636, KJ2021A0639 and KJ2020A0820), Project of Support Program for Excellent Young Talents in Colleges and Universities of Anhui Province (Grant gxyqZD2019046), and Postgraduate Science Research Program in Colleges and Universities of Anhui Province (Grant YJS20210514).

References

[1] F. X. Liu, P. Xu, Mater. Res. Express **8**, 045901 (2021).
[2] Z. F. Zhang, S. M. Hou, T. Wang, S. J. Liu, X. G. Yang, Q. J. Li, P. F. Shen, B. B. Liu, H. P. Gao, Y. L. Mao, J. Alloys Compd. **859**, 157882 (2021).
[3] B. Dong, B. S. Cao, Y. Y. He, Z. Liu, Z. P. Li, Z. Q. Feng, Adv. Mater. **24**, 1987 (2012).
[4] Y. Pan, E. Yu, Int. J. Energy Res. **45**, 11284 (2021).
[5] M. Sadiq, M. N. Khan, M. Arif, A. Naveed, K. Ullah, S. Afridi, Mater. Res. Express **8**, 095507 (2021).
[6] H. Kim, I. Park, D. H. Seo, S. Lee, S. W. Kim, W. J. Kwon, Y. U. Park, C. S. Kim, S. Jeon, K. Kang, J. Am. Chem. Soc. **134**, 10369 (2012).
[7] C. Wen, Y. Zhang, C. X. Wang, D. Z. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, Acta Mater. **15**, 10369 (2019).
[8] U. Bhandari, M. R. Rafi, C. Y. Zhang, S. Z. Yang, Mater. Today Commun. **1**, 101871 (2021).
[9] J. Zhang, B. Xu, Y. X. Xiong, S. H. Ma, Z. Wang, Z. G. Wu, S. J. Zhao, npj Comput. Mater. **8**, 1 (2022).
[10] U. Bhandari, C. Y. Zhang, C. Y. Zeng, S. M. Guo, A. Adhikari, S. Z. Yang, Crystals. **11**, 46 (2021).
[11] J. F. Durodola, Prog. Mater. Sci. **1**, 100797 (2022).
[12] M. Ahmad, P. Kamiński, P. Olczak, M. Alam, M. J. Iqbal, F. Ahmad, S. Sasui, B. J. Khan, Appl. Sci. **11**, 6167 (2021).
[13] S. K. Kauwe, J. Graser, A. Vazquez, T. D. Sparks, Integrating Materials and Manufacturing Innovation. **7**, 43 (2018).
[14] Z. Y. Wan, Q. D. Wang, D. C. Liu, J. H. Liang, Phys. Chem. Chem. Phys. **23**, 15675 (2021).
[15] Y. T. Sun, H. Y. Bai, M. Z. Li, W. H. Wang, J. Phys. Chem. Lett. **8**, 3434 (2017).
[16] Z. Li, Z. L. Long, S. Lei, T. Zhang, X. W. Liu, D. M. Kuang, Comput. Mater. Sci. **197**, 110656 (2021).
[17] Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duck, J. Brgoch, Nat. Commun. **9**, 1 (2018).
[18] Y. Zhuo, S. Hariyani, S. H. You, P. Dorenbos, J. Brgoch, J. Appl. Phys. **128**, 013104 (2020).
[19] C. C. Lin, R. S. Liu, J. Phys. Chem. Lett. **2**, 1268 (2011).
[20] X. Wang, H. F. Shi, H. L. Ma, W. P. Ye, L. L. Song, J. Zan, X. K. Yao, X. Y. Ou, G. H. Yang, Z. Zhao, M. Singh, C. Y. Lin, H. Wang, W. Y. Jia, Q. Wang, J. H. Zhi, C. M. Dong, X. Y. Jiang, Y. G. Tang, X. J. Xie, Y. Yang, J. P. Wang, Q. S. Chen, Y. Wang, H. H. Yang, G. Q. Zhang, Z. F. An, X. G. Liu, W. Huang, Nat. Photonics. **15**, 187 (2021).
[21] E. Yablonovitch, J. Opt. Soc. Am. B. **10**, 283 (1993).
[22] C. Kittel, P. McEuen, Introduction to solid state physics, Wiley, New York. **8**, 105 (1996).
[23] A. L. Samuel, IBM J. Res. Dev. **3**, 210 (1959).
[24] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, APL Mater. **1**, 011002 (2013).
[25] J. Brgoch, S. P. DenBaars, R. Seshadri, J. Phys. Chem. C. **117**, 17955 (2013).
[26] S. K. Kauwe, T. Welker, T. D. Sparks, Integrating Materials and Manufacturing Innovation, **9**, 213 (2020).
[27] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, Comput. Mater. Sci. **68**, 314 (2013).
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, J. Mach. Learn. Res. **1**, 2825 (2011).
[29] R. Tibshirani, J. R. Stat. Soc. **58**, 267 (1996).
[30] N. Cristianini, J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press (2000).
[31] C. Williams, C. E. Rasmussen, NeurIPS **8** (1995).
[32] L. Breiman, Machine Learning. **45**, 5 (2001).
[33] B. E. Boser, I. M. Guyon, V. N. Vapnik, In Proceedings of The Fifth Annual Workshop on Computational Learning Theory, 144 (1992).
[34] J. H. Friedman, Comput. Stat. Data Anal. **38**, 367 (2002).
[35] J. H. Friedman, Ann. Stat. **29**, 1189 (2001).
[36] W. He, B. Li, R. Liao, H. Mo, L. Tian, Knowledge-Based Systems **237**, 107778 (2022).

*Corresponding author: wenjunkd@mail.ustc.edu.cn; jwen@aqnu.edu.cn