

Evaluation of performance of machine learning methods in mining structure-property data of halide perovskite materials

Ruoting Zhao(赵若廷), Bangyu Xing(邢邦昱), Huimin Mu(穆慧敏), Yuhao Fu(付钰豪), and Lijun Zhang(张立军)
Citation: Chin. Phys. B, 2022, 31 (**5**): 056302. DOI: 10.1088/1674-1056/ac5d2d
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Synthesis and thermoelectric properties of Bi-doped SnSe thin films

Jun Pang(庞军), Xi Zhang(张析), Limeng Shen(申笠蒙), Jiayin Xu(徐家胤), Ya Nie(聂娅), and Gang Xiang(向钢)
Chin. Phys. B, 2021, 30 (**11**): 116302. DOI: 10.1088/1674-1056/ac11da

Inverted V-shaped evolution of superconducting temperature in SrBC under pressure

Ru-Yi Zhao(赵如意), Xun-Wang Yan(闫循旺), and Miao Gao(高淼)
Chin. Phys. B, 2021, 30 (**7**): 076301. DOI: 10.1088/1674-1056/abfbcc

Two-dimensional MnN utilized as high-capacity anode for Li-ion batteries

Junping Hu(胡军平), Zhangyin Wang(王章寅), Genrui Zhang(张根瑞), Yu Liu(刘宇), Ning Liu(刘宁), Wei Li(李未), Jianwen Li(李健文), Chuying Ouyang(欧阳楚英), and Shengyuan A. Yang(杨声远)
Chin. Phys. B, 2021, 30 (**4**): 046302. DOI: 10.1088/1674-1056/abdda8

Effect of strain on electrochemical performance of Janus MoSSe monolayer anode material for Li-ion batteries: First-principles study

Guoqing Wang(王国庆), Wenjing Qin(秦文静), and Jing Shi(石晶)
Chin. Phys. B, 2021, 30 (**4**): 046301. DOI: 10.1088/1674-1056/abca24

Theoretical analysis of cross-plane lattice thermal conduction in graphite

Yun-Feng Gu(顾云风)
Chin. Phys. B, 2019, 28 (**6**): 066301. DOI: 10.1088/1674-1056/28/6/066301

DATA PAPER

Evaluation of performance of machine learning methods in mining structure–property data of halide perovskite materials

Ruoting Zhao(赵若廷)[1], Bangyu Xing(邢邦昱)[1], Huimin Mu(穆慧敏)[2], Yuhao Fu(付钰豪)[2,3], and Lijun Zhang(张立军)[1,3,†]

[1] *State Key Laboratory of Integrated Optoelectronics, Key Laboratory of Automobile Materials of MOE, Jilin Provincial International Cooperation Key Laboratory of High-Efficiency Clean Energy Materials, Electron Microscopy Center, and School of Materials Science and Engineering, Jilin University, Changchun* 130012, *China*

[2] *State Key Laboratory of Superhard Materials, College of Physics, Jilin University, Changchun* 130012, *China*

[3] *International Center of Computational Method and Software, Jilin University, Changchun* 130012, *China*

(Received 19 February 2022; revised manuscript received 12 March 2022; accepted manuscript online 14 March 2022)

With the rapid development of artificial intelligence and machine learning (ML) methods, materials science is rapidly entering the era of data-driven materials informatics. ML models serve as the most crucial component, closely bridging material structure and material properties. There is a considerable difference in the prediction performance of different ML methods for material systems. Herein, we evaluated three categories (linear, kernel, and nonlinear methods) of models, with twelve ML algorithms commonly used in the materials field. In addition, halide perovskite was chosen as an example to evaluate the fitting performance of different models. We constructed a total dataset of 540 halide perovskites and 72 features, with formation energy and bandgap as target properties. We found that different categories of ML models show similar trends for different target properties. Among them, the difference between the models is enormous for the formation energy, with the coefficient of determination (R^2) range 0.69–0.953. The fitting performance between the models is closer for bandgap, with the R^2 range 0.941–0.997. The nonlinear-ensemble model shows the best fitting performance for both the formation energy and the bandgap. It shows that the nonlinear-ensemble model, constructed by combining multiple weak learners, effectively describes the nonlinear relationship between material features and target property. In addition, the extreme gradient boosting decision tree model shows the most superior results among all the models and searches for two new descriptors that are crucial for formation energy and bandgap. Our work provides useful guidance for the selection of effective machine learning methods in the data-mining studies of specific material systems. The dataset that supported the findings of this study is available in Science Data Bank, with the link https://www.doi.org/10.11922/sciencedb.01611.

Keywords: machine learning, material informatics, first-principles calculations, halide perovskites

PACS: 63.20.dk, 71.15.Mb, 73.22.−f **DOI:** 10.1088/1674-1056/ac5d2d

1. Introduction

The rapid growth in machine learning (ML) and artificial intelligence has brought about significant development in natural science, especially in physics, materials, and biology.[1–5] The reasons for the widespread use of ML can be ascribed to two points: Firstly, A wide variety of data is currently available through experiment or simulated calculations.[6,7] Secondly, the rapid development of ML models, such as ensemble model and neural network (NN).[8,9] Although materials science is still not at a data-rich stage,[10] the rapid development of materials science databases (e.g., Material Project,[11] AFLOW[12]) and material ML methods/software (e.g., CGCNN,[13] GAP[14,15]) provides a robust foundation for achieving high-speed and large-scale materials design. Material ML method can be divided broadly into supervised learning and unsupervised learning. Supervised learning constructs mapping relationships between material descriptors (features) and target properties, including the two main aspects of interatomic potential[16] and quantitative structure–property relationship (QSPR).[17] Although they both predict the properties of the material, there are obvious differences in their essence. The focus of the interatomic potential is to construct a sufficiently accurate (close to quantum calculations) and inexpensive surrogate model for predicting the energy and force of materials. The QSPR describes a specific material system and generates problem-specific descriptors (e.g., structure–composition descriptors and easily accessible crystal/molecular properties) so that QSPR addresses the end-to-end description of thermodynamic properties. Therefore, it has a broader range of predictive properties and is better suited to search for new materials on demand. Unsupervised learning is based on the representation method of material to find the similarities and differences among different material groups.[18] The methods to distinguish the structural differences include Euclidean, Manhattan, and Tanimoto coefficients.[19]

On-demand designing of functional materials using the QSPR method has become a new hot topic in materials

†Corresponding author. E-mail: lijun_zhang@jlu.edu.cn

© 2022 Chinese Physical Society and IOP Publishing Ltd

informatics.[20] Using ML models for mining the intrinsic rules from high-throughput computational results or existing material databases has become a trend to guide theoretical/experimental design and discover new materials with optimized performance.[21–25] The advantage of this method lies in the balance between reasonable experimental requirements and a lower error rate. ML model serves as the most critical link bridging the relationship between material structure and target properties. ML models are the most critical link in the QSPR approach, which constructs linear and nonlinear relationships between material structure and target properties. Therefore, selecting the appropriate ML method is critical in constructing QSPR materials research, significantly affecting prediction accuracy and generalization ability. Generally, ML methods applied in QSPR material research are classified into the linear, kernel, and nonlinear methods. Linear methods are based on a weighted linear sum of selected subsets of material features to fit the target properties, e.g., least absolute shrinkage and selection operator (LASSO), ridge regression.[26–28] Meanwhile, these methods are also more suitable for small dataset problems, feature selection, and avoiding overfitting problems. The kernel approach transforms the original material feature set into a high-dimensional feature space and then linearly differentiates material, and the features are not directly nonlinearly processed, e.g., support vector machine (SVM). The nonlinear approach refers to the direct nonlinear processing of material features inside the method. It is used to handle when target properties do not vary smoothly in the material feature space, such as decision tree (DT), ensemble model, and neural network (NN) model.

The effectiveness of ML models for fitting material systems has still not been uniformly studied, and it remains an important problem in materials informatics. Moreover, perovskite material received considerable attention in the optoelectronic field due to its advantages, such as ultralong carrier diffusion length,[29–31] high threshold light absorption,[32,33] easily adjustable bandgap, and low-cost fabrication.[34–36] Researchers have used high-throughput computational approach and ML methods to design the novel perovskite materials that maintain superior optoelectronic properties with structural stability and non-toxicity to adapt to different optoelectronic devices.[37–45] Shuaihua Lu *et al.* constructed a gradient boosting decision tree (GBDT) ML model by the first-principles calculation of 346 HOIP structures and screened six potentially synthesizable HOIPs from 5158 structures that have not been calculated.[46] Zhenzhu Li *et al.* trained the kernel ridge regression (KRR), K-nearest neighbor (KNN), support vector machine (SVM) models by using a dataset of 354 perovskite structures. Furthermore, they found that the KRR model can significantly learn the intrinsic laws of stability of perovskite materials and agree well with the available experimental data, much better than empirical formulations such as tolerance factor and octahedral factor.[47] Adnan Ali *et al.* selected 852 types of HOIP structures and all-inorganic perovskite structures to construct an artificial neural network (ANN) model and generate a ternary phase diagram of phase stability. Based on the ternary phase diagram output from the ANN model to further guide the experiments successfully, it was found that the introduction of Cs ions at the *A*-site can increase the stability of the cubic HOIP structure.[48] Although researchers have chosen different ML methods based on the research system to build specific models,[49–53] there is no uniform research on the fitting effect of common ML models in these applications.

Therefore, in this paper, we aim to evaluate the differences in the prediction performance of different ML models for material systems. Three classes (linear model, kernel model, and nonlinear model) of twelve ML methods commonly used in the materials domain were selected, including LASSO, KRR, KNN, SVM, decision tree (DT), extremely randomized tree (ERT), AdaBoost decision tree (ABDT), random forest (RF), GBDT, light gradient boosting decision tree (LGDT), extreme gradient boosting decision tree (XGDT), and ANN model. Furthermore, a dataset of 540 structures by ion transmutation with 72-dimensional material features (elemental, structural, and combined features) was constructed using mixed halide perovskite as the study system. Formation energy and band gap were used as target properties. The results found that different classes of ML models for different target properties showed similar trends. The nonlinear-ensemble models all show excellent fitting performance and are much higher than the others. It indicates that this method, which is a combination of multiple weak learners, is more suitable for describing material systems. Meanwhile, we found a great variation between models for predicting formation energy, with R^2 ranging from 0.69 to 0.953. There are minor differences between models for predicting bandgap, with R^2 ranging from 0.941 to 0.997. It shows that the formation energy has a higher degree of nonlinearity than the bandgap. In addition, XGDT shows the most excellent results among all the models. Two descriptors for the formation energy and bandgap are searched simultaneously, CV$(\frac{\chi_B/\chi_X}{d_{B-X}})$ and mean$(\frac{\chi_B/\chi_X}{d_{B-X}})$. Our study provides a systematic discussion of model selection for specific material research systems, which may be helpful for future ML applications in materials science.

2. Computational method

2.1. First-principles calculations

The first-principles density functional theory (DFT) calculations were carried out using the plane-wave pseudopotential method implemented in Vienna *ab initio* simulation package.[54,55] The projector augmented wave pseudopotentials describe the electron–ion interactions. The generalized gradient approximation of Perdew, Burke, and Emzerhof was

used to describe the exchange–correlation function.[56,57] The kinetic cut-off energy for the plane-wave basis was set as 500 eV, and the energy convergence threshold was set as 10^{-5} eV. Monkhorst–Pack k-point meshes with the spacing of $2\pi \times 0.25$ Å^{-1} were employed for structure optimizations and self-consistent calculations. The structures (including lattice parameters and atomic positions) were fully optimized until the residual forces were converged within 0.02 eV/Å. The formation energy (E_f) and the bandgap (E_g) of the structure were calculated and regarded as predicted target properties. Since the same valence binary halides are often used as precursors for the experimental synthesis of perovskite, E_f was defined as the energy difference between the perovskite energy and the binary compounds with the same oxidation state reaction/product. The formula is as follows:

$$E_f = E_{ABX_3} - E_{AX} - E_{BX_2},$$

where E_{ABX_3} is the total energy of the perovskite, E_{AX} and E_{BX_2} are the total energies of binary metal halides (AX and BX_2 are in space group $Pm\bar{3}m$ and $Pnma$, respectively). More detailed computational methods can be referred to our previous works on halide perovskites.[41,58–63] The high-throughput first-principles calculations were performed via the Jilin Artificial-intelligence aided Material-design Integrated Package (JAMIP), which is an open-source artificial-intelligence-aided data-driven infrastructure specially designed for computational material informatics.[34,64]

2.2. Machine learning

ML calculations were implemented using the JAMIP ML module.[64] The entire process includes constructing computational material data, feature generation, feature engineering, and model construction and evaluation, as shown in Fig. 1(a). The entire material data was constructed by using cubic mixed halogen perovskite (space group $Pm\bar{3}m$, shown in Fig. 1(b)) with the chemical formula ABX_3. Specifically, we considered combinations of A = K, Rb, Cs; B = Mg, Ca, Sr, Ba, Zn, Cd, Ge, Sn, Pb; and X = F, Cl, Br, I. The data set contains 540 kinds of mixed halogen perovskite structures.

All features were uniformly preprocessed before feature input into the model to eliminate the differences in data identification with different magnitudes. The recursive feature elimination (RFE) method was used as a feature selection method. We selected feature weight as an elimination criterion for the model with feature weight. We selected four correlation coefficients for the corresponding model for the model without feature weight, including Pearson correlation (PC), Spearman correlation (SC), distance correlation (DC), and mutual information correlation (MIC). The minimum error of the corresponding model among the four correlation coefficients was taken as the final correlation coefficient. Moreover, twelve models were selected to recognize the most effective models for representing the perovskite system. All models (see Fig. 1(c)) were constructed with JAMIP by calling the corresponding ML software packages. LASSO, KRR, KNN, SVM, DT, ERT, ABDT, RF, and GBDT were called from the Scikit-learning python package, XGDT was called from the XGBoost python package, and LGDT was called from the Lightgbm python package.[65–67] The optimal hyper-parameters of all models were searched using the same standard Bayesian optimization algorithm, and ten-fold CV was used as the evaluation metric.[68] We used averaging method to search for "random_ state" hyperparameter. In Bayesian optimization search process, the result of the average 10 times 10-fold CV eliminated randomization. In the fitting performance, the result of the average 100 times eliminated the randomization, and searched for the "random_ state" with the closest average effect as the final hyperparameter.

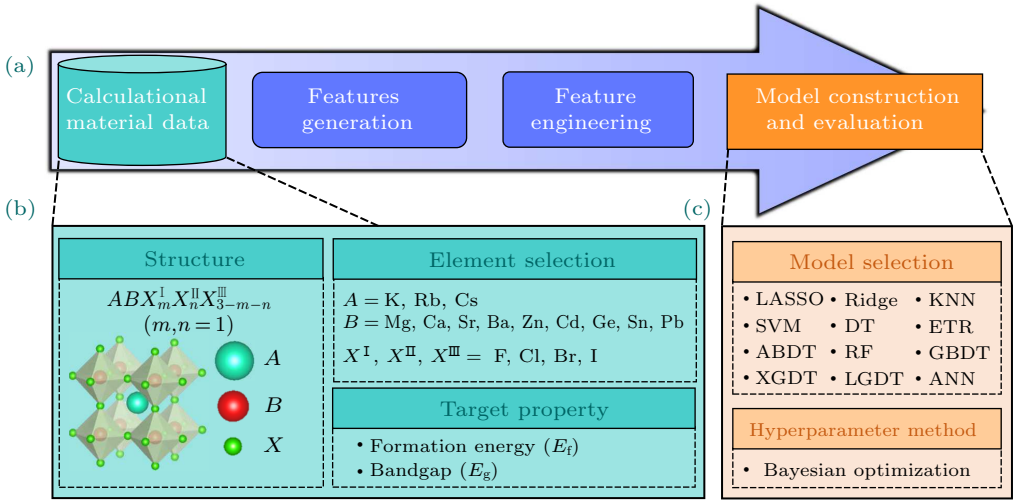


Fig. 1. Schematic diagram of the overall machine learning workflow. (a) The process consists of building a JAMIP ML calculational material data, feature generation, feature engineering, and model construction and evaluation. (b) Single perovskite schematic diagram, A-site with three alkali metal monovalent elements in A-site, nine divalent metal elements in B-site, and four halogen elements in X-site. (c) Machine learning models and hyperparameter method.

3. Results and discussion

Typically, ML studies involve selecting the most appropriate model and the most subtle input features. The initial feature set design is shown in Fig. 2(a). The elemental information of the perovskite was extracted from the Pymatgen database,[69] including atomic radius, ionic radius, ionization energy, electronic affinity energy, electronegativity, Mendeleev number, and atomic mass. The structure features of perovskite, such as (octahedral factor and tolerance factor) and combined features based on the physical/chemical intuition were constructed by the ML module of JAMIP. Moreover, the results of the DFT calculation for the formation energy and bandgap of perovskite were considered as target properties. The dataset contained 72 features and all feature information listed in the support information Table S1. All features were normalized and scaled down to between 0 and 1. Moreover, in order to fairly compare the differences in the fitting performance of perovskite among the twelve models, 80% of the total dataset was selected as the training set and 20% as the validation set. We selected the layered sampling method to ensure that the dataset was distributed evenly over the entire range area. The training set was used to train the model and search for the optimal feature set, and the validation set was used to evaluate the performance of the model.

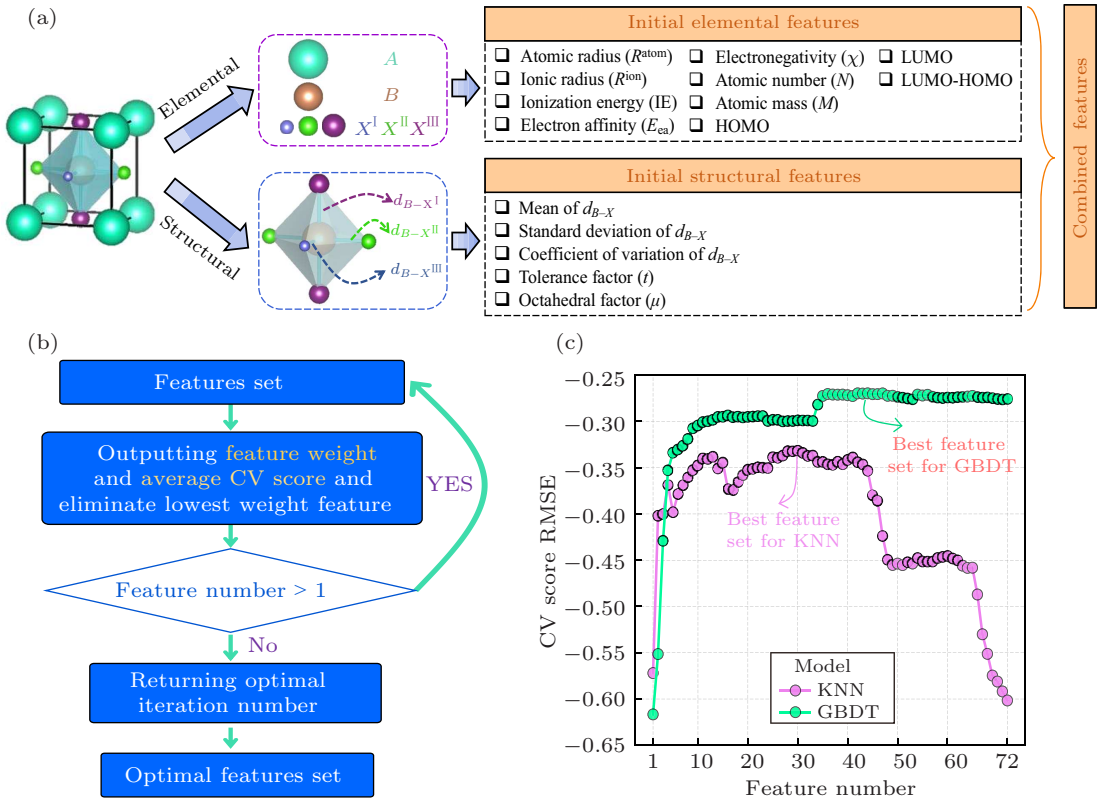


Fig. 2. Feature design and feature engineering details. (a) Perovskite feature design includes three parts: initial elemental feature, initial structural feature, and combined feature. (b) Recursive feature elimination (RFE) flow chart. (c) RFE method screening process.

Feature engineering is a crucial step for ML research, which often determines the lower limit of the fitting performance of the model, and the features determine the upper limit.[70] Feature engineering usually consists of feature design and feature selection. For system-specific ML study, feature design consists of initial elemental and structural features based on extraction from the structure and combined features based on physical–chemical intuition. Since the entire feature set is often over-complete, it can lead to a situation of the dimensional curse due to feature redundancy.[26] The RFE method was used to select the optimal feature set of each model. The whole process takes RMSE as the evaluation metric. The RFE method assesses the weight of all features, and the feature with the lowest weight is eliminated. The features subset corresponding to the lowest error is defined as the optimal feature set. We chose the ten-fold CV method for the whole process. For models that do not output feature weights information (SVM/ANN/KNN), a univariate feature selection method was selected as searching for the optimal feature subset, which evaluates the correlation weight (PC, SC, DC, MIC) between each feature and target property. The minimum error of the corresponding model among the four correlation coefficients was taken as the final correlation coefficient. The low correlation weight feature will be removed recursively in turn. The feature selection flow is shown in Fig. 2(c).

The RFE process of formation energy is shown in Fig. S1(a). All model RMSE values show a rise firstly with an increasing number of features and then tend to fluctuate after maximum value stably. After completing feature selection, the feature dimensions of all formation energy models are

shown in Table 1. The fitting performance of the nonlinear-ensemble model is better than that of other models, with the LGDT model being the most superior. LGDT is a significantly improved version of the gradient boosting decision tree, where the histogram-based decision tree algorithm and leaf-wise leaf growth strategy with depth constraints considerably increase the efficiency. The minor ten-fold CV RMSE is 0.4156 eV/formula when the feature dimension reaches 30. The XGDT model also achieves a precision close to the LGDT model (0.4167 eV/formula), but its optimal feature set has not eliminated any features. It shows that nonlinear-ensemble models (LGDT/XGDT/GBDT/ABDT) can effectively characterize the formation energy information. However, other models show a wide variation in fitting performance. The SVM is the worst fitting model with an RMSE CV of 0.7596 eV/formula. It indicates that the relationship between the formation energy and the input features is highly nonlinear. Nonlinear-ensemble models composed of multiple weak learners have better generalization capability. The RFE result of the bandgap is shown in Table 1 and Fig. S1(b). The entire RFE process is similar to the trend of the formation energy, and the nonlinear-ensemble model also performs better than the other models, with the XGDT model showing the optimal performance. The minor ten-fold CV RMSE is 0.2576 eV with feature dimension reduced to 42.

Based on the feature selection results, the second hyperparameter tuning is performed on all models with the optimal feature set and the trained models predicted for the validation set. The results of all formation energy model fitting performances are shown in Fig. 3(a) and Fig. S2. We found that the models vary widely in their evaluations, where the models with R^2 above 0.9 were only KNN/ABDT/GBDT/LGDT/XGDT. It means that only these models could learn formation energy intrinsic trends. The XGDT model performs optimally for predicting structures not involved in the training set within these models. For the structural formation energy model, the prediction results around 0 eV are usually more descriptive of model robustness. Subsequently, we accounted for the accuracy (percentage of all structures with positive/negative formation energy correctly predicted by the model), MAE, and RMSE of all models in predicting the intervals from −1 eV to 1 eV (see more details in Table S3 of support information). The XGDT model remains the top-performing model with an accuracy of 95.5455%, MAE of 0.1060 (eV/formula), and RMSE of 0.1967 (eV/formula). XGDT is an improved gradient boosting decision tree, which performs the second-order Taylor expansion for loss function compared to the GBDT model and is constructed with the weighted sum of multiple weak learner models (decision tree). Given the excellent performance of XGDT, this model can be used for all future predictions in terms of perovskite stability. All bandgap model fitting results are shown in Fig. 4(b) and Fig. S3. We found that R^2 in all models exceeded 0.9, which suggests that the models can mine the general trend of the bandgap. It also means that the intrinsic laws of perovskite formation energy are more difficult

Table 1. The feature selection procedure result of formation energy and bandgap for the twelve models.

Model	Numbers of optimal features	
	Formation energy	Bandgap
ETR	29	16
DT	41	36
SVM	35	46
LASSO	10	18
KNN	3	30
ANN	36	72
RF	30	44
Ridge	36	26
ABDT	18	46
GBDT	29	44
LGDT	30	72
XGDT	72	42

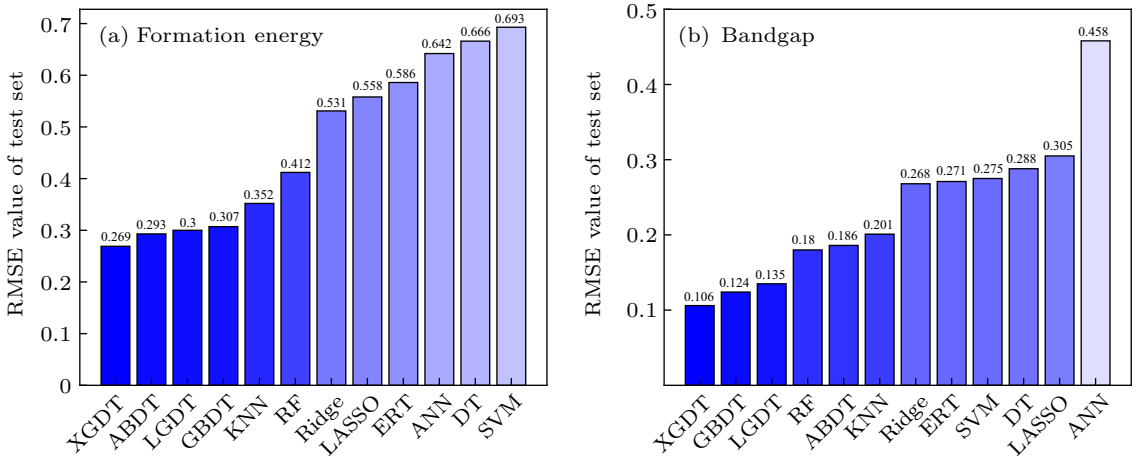


Fig. 3. The fitting error of the test set for twelve ML models. (a) Formation energy fitting results, the order of model fitting error is XGDT, ABDT, LGDT, GBDT, KNN, RF, Ridge, LASSO, ERT, ANN, DT, and SVM. (b) Bandgap fitting results, the order of model fitting error is XGDT, GBDT, LGDT, RF, ABDT, KNN, Ridge, ERT, SVM, DT, LASSO, and ANN.

to mine than the bandgap. The XGDT mode performs better than the other model with MAE of 0.144 eV, RMSE of 0.269 eV, and R^2 of 0.953. Furthermore, there is little difference in the accuracy of the predictions for the various interval ranges of the bandgap. Therefore, XGDT is one of the most promising ML models for applying to the perovskite system. In addition to the XGDT model, the LGDT and GBDT also show superior results to the other model, which belong to the nonlinear-ensemble model.

The essence of the XGDT model is a decision tree model, which can output the feature importance information to understand the relationship between structure and property. For each feature, the feature importance value is between 0 and 1, with 0 represented feature having a negligible effect on the target property and 1 represented feature having a significant effect on the target property. Moreover, the features are not independent of each other but act in concert to affect the target property. Figures 4(a) and 4(b) show the top ten feature ranks for formation energy and bandgap, respectively. For formation energy, the highest importance feature $\mathrm{CV}(\frac{\chi_B/\chi_X}{d_{B-X}})$ is defined as

$$\mathrm{CV}\left(\frac{\chi_B/\chi_X}{d_{B-X}}\right)=\frac{\sqrt{\sum_{i=1}^{6}\left(\frac{\chi_B/\chi_X}{d_{B-X}}-\frac{1}{6}\sum_{i=1}^{6}\frac{\chi_B/\chi_X}{d_{B-X}}\right)^2}}{\frac{1}{6}\sum_{i=1}^{6}\frac{\chi_B/\chi_X}{d_{B-X}}},$$

where χ_B, χ_X represent electronegativity of B, X-site, and d_{B-X} represents $[BX_6]^-$ octahedral B–X bond length. This formula indicates that the degree of variation of the B–X electronegativity ratio concerning the B–X bond length quotient plays a decisive factor in the stability of mixed halide perovskite. The second and third high important features are also combined features of elemental electronic and distance information from the B–X position, demonstrating that the relationship of the B–X position dominates the formation energy. Visual analysis of the most important feature and formation energy relationship is shown in Fig. 4(c). We found that the larger $\mathrm{CV}(\frac{\chi_B/\chi_X}{d_{B-X}})$ the structure tends to become unstable, which is consistent with the widely-accepted point that the structural stability of perovskite is roughly affected by the geometric criteria T_f and B–X bonding strength and the difference of electronegativity and distance between bond partners.[71]

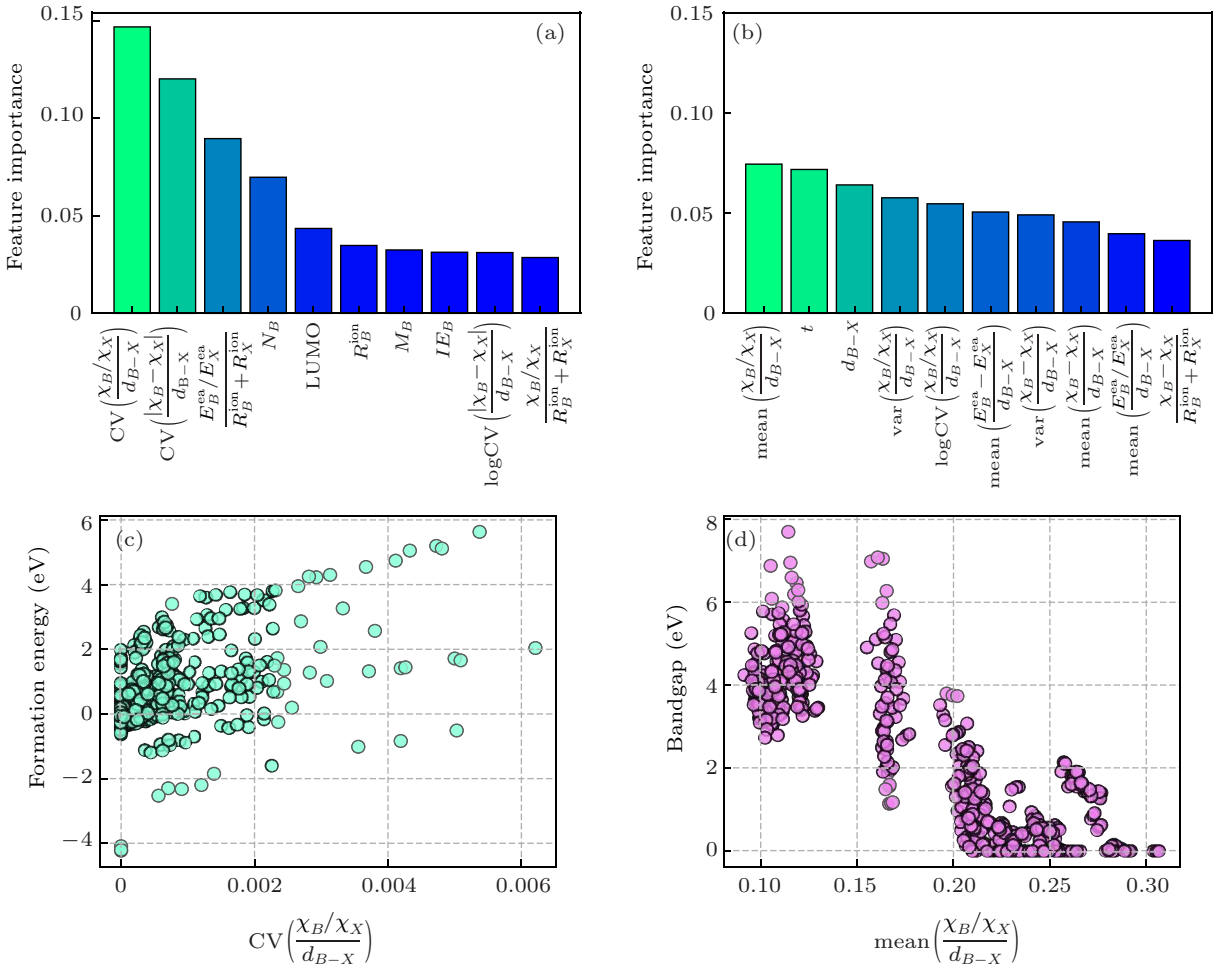


Fig. 4. The result of feature importance and two-dimensional sensitivity analysis of the XGDT model. (a) The top 10 importance features rank by the XGDT model for formation energy prediction. (b) The top 10 importance features rank by the XGDT model in bandgap prediction. (c) Relationship visualization of formation energy with $\mathrm{CV}(\frac{\chi_B/\chi_X}{d_{B-X}})$. (d) Relationship visualization of bandgap with $\mathrm{mean}(\frac{\chi_B/\chi_X}{d_{B-X}})$.

The feature rank of the bandgap is shown in Fig. 4(b). The highest importance feature mean($\frac{\chi_B/\chi_X}{d_{B-X}}$) is defined as

$$\text{mean}(\frac{\chi_B/\chi_X}{d_{B-X}}) = \frac{1}{6}\sum_{i=1}^{6}(\frac{\chi_B/\chi_X}{d_{B-X}}), \quad (1)$$

where χ_B/χ_X represent the electronegativity of B/X-site, d_{B-X} represents $[BX_6]^-$ octahedral B–X bond length. This formula shows that the mixed halide perovskite's stability is critically dependent on the mean value of the B–X electronegativity ratio and the B–X bond quotient. The second and third highest importance features are the tolerance factor and the B–X average bond length, respectively, showing that the chalcogenide band gap is affected by B/X electronegativity, B–X distance, and $[BX_6]^-$ octahedral distortion. It is mainly due to that the valence band maximum and conduction band minimum of perovskite are mainly occupied by the atomic orbitals of B and X. The atomic distance between B and X determines the degree of overlap between the s/p orbitals of B and the p orbitals of X. Visual analysis of the most important feature and bandgap relationship is shown in Fig. 4(d). We found that the bandgap gradually decreases as the mean value of the B–X electronegativity ratio to the B–X bond length quotient increases. It is mainly due to the gradual increase in octahedral distortion as the highest importance feature decreases, which results in less overlap between the s/p orbitals of B and the p orbitals of X, leading to an increase in the bandgap when the B-site electronegativity is fixed.[72] The characterization of the perovskite stability and bandgap shows that the B–X electronegativity and the degree of $[BX_6]^-$ octahedral distortion play a crucial role in the structural properties of perovskite. In summary, adjusting the appropriate halogen percentage to modulate the bandgap range within a stable structure can design perovskite materials for specific optoelectronic requirements.

4. Conclusion

In this work, we systematically analyze the differences in the fitting performance of different classes of machine learning methods for material systems. We constructed 12 machine learning models in three classes (linear, kernel, and nonlinear models). Meanwhile, halide perovskite was chosen as the research system, and 540 structures were constructed by ion transmutation. The nonlinear-ensemble model shows the best fitting performance. However, the nonlinear-single model (neural network, decision tree, support vector machine, etc.) exhibits inferior fitting performance. It shows that the nonlinear-ensemble model is more suitable for the machine learning research of material systems. The XGDT model acts as the best model for predicting the unknown structure, and root mean square error, mean absolute error, and coefficient of determination are 0.106 eV/formula, 0.079 eV/formula, and 0.997, respectively. Further analysis of the structure–property relationships from the XGDT model reveals two primary descriptors that influence the energy of chalcocite formation and bandgap, CV($\frac{\chi_B/\chi_X}{d_{B-X}}$) and mean($\frac{\chi_B/\chi_X}{d_{B-X}}$). The perovskite formation energy is mainly influenced by the B–X electronegativity and the degree of variation in $[BX_6]^-$ octahedral bond lengths, and the bandgap is mainly influenced by B–X electronegativity and $[BX_6]^-$ octahedral average bond lengths.

The superior results of the nonlinear-ensemble model benefit from employing multiple ensemble approach to optimize the higher variance and bias problem of the single model. The common ensemble approach includes boosting,[73] bagging,[74] and stacking.[75] In this study, we constructed XGDT, LGDT, GBDT, and ABDT models using the boosting method and RF model using the bagging method. All nonlinear-ensemble models taking boosting methods are better than bagging methods in formation energy and bandgap. This is mainly because the boosting method adopts updating the sample weights of every single model according to the prediction error (loss function). Then, the high bias samples are continuously trained to focus on optimizing the model and reducing the loss function. The advantage of the XGDT model for achieving the best performance is mainly reflected in Newton's method, in which the extreme value of the loss function is solved, the second-order Taylor expansion of the loss function is carried out, and the regularization term is added to avoid overfitting. This treatment is more sensitive to fitting errors.[66] Meanwhile, we found that the ANN model that performs well in the ML fitting potential does not achieve excellent results. It is because ANN algorithm often requires large data sets to support them. Otherwise, they can cause great overfitting.[76] Therefore, the nonlinear-ensemble ML models are suitable for research on small sample material datasets. In addition, the XGDT model not only achieves the best fitting performance, but also captures potential relationships between halide perovskite material descriptors and target properties (formation energy and bandgap). This approach helps to understand the properties of halide perovskites and permits more efficient work on finding descriptors. Thus, the XGDT may be a potential ML model for applying in the field of halide perovskites and even other types of material systems in the future.

Data availability

The data that support the findings of this study are openly available in Science Data Bank at https://www.doi.org/10.11922/sciencedb.01611.

Acknowledgments

The project supported by the National Natural Science Foundation of China (Grants Nos. 62125402 and 92061113). Calculations were performed in part at the high-performance computing center of Jilin University.

References

[1] Gasteiger J and Zupan J 1993 *Angewandte Chemie International Edition in English* **32** 503
[2] Mater A C and Coote M L 2019 *J. Chem. Inf. Model.* **59** 2545
[3] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 *Rev. Mod. Phys.* **91** 045002
[4] Wei J, Chu X, Sun X Y, Xu K, Deng H X, Chen J, Wei Z and Lei M 2019 *InfoMat* **1** 338
[5] Chen A, Zhang X and Zhou Z 2020 *InfoMat* **2** 553
[6] Lyu R, Moore C E, Liu T, Yu Y and Wu Y 2021 *J. Am. Chem. Soc.* **143** 12766
[7] Jablonka K M, Ongari D, Moosavi S M and Smit B 2020 *Chem. Rev.* **120** 8066
[8] Behler J 2021 *Chem. Rev.* **121** 10037
[9] Chen C, Zuo Y, Ye W, Li X, Deng Z and Ong S P 2020 *Adv. Energy Mater.* **10** 1903242
[10] Batra R, Song L and Ramprasad R 2021 *Nat. Rev. Mater.* **6** 655
[11] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Materials* **1** 011002
[12] Calderon C E, Plata J J, Toher C, Oses C, Levy O, Fornari M, Natan A, Mehl M J, Hart G, Buongiorno Nardelli M and Curtarolo S 2015 *Computational Materials Science* **108** 233
[13] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[14] Bartók A P and Csányi G 2015 *International Journal of Quantum Chemistry* **115** 1051
[15] Bartók A P, Kermode J, Bernstein N and Csányi G 2018 *Phys. Rev. X* **8** 041048
[16] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K R 2021 *Chem. Rev.* **121** 10142
[17] Musil F, Grisafi A, Bartók A P, Ortner C, Csányi G and Ceriotti M 2021 *Chem. Rev.* **121** 9759
[18] Kireeva N and Pervov V S 2017 *Phys. Chem. Chem. Phys.* **19** 20904
[19] Bajusz D, Rácz A and Héberger K 2015 *Journal of Cheminformatics* **7** 20
[20] Tao Q, Xu P, Li M and Lu W 2021 *npj Computational Materials* **7** 1
[21] Pollice R, dos Passos Gomes G, Aldeghi M, Hickman R J, Krenn M, Lavigne C, Lindner-D'Addario M, Nigam A, Ser C T, Yao Z and Aspuru-Guzik A 2021 *Acc. Chem. Res.* **54** 849
[22] Gao C, Min X, Fang M, Tao T, Zheng X, Liu Y, Wu X and Huang Z 2022 *Advanced Functional Materials* **32** 2108044
[23] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 *npj Comput. Mater.* **5** 1
[24] Liu Z, Na G, Tian F, Yu L, Li J and Zhang L 2020 *InfoMat* **2** 879
[25] Luo S, Li T, Wang X, Faizan M and Zhang L 2021 *WIREs Computational Molecular Science* **11** e1489
[26] Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli L M and Ouyang R 2018 *Phys. Rev. Materials* **2** 083802
[27] Santosa F and Symes W W 1986 *SIAM J. Sci. Stat. Comput.* **7** 1307
[28] Tibshirani R 1996 *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267
[29] Shi D, Adinolfi V, Comin R, Yuan M, Alarousu E, Buin A, Chen Y, Hoogland S, Rothenberger A, Katsiev K, Losovyj Y, Zhang X, Dowben P A, Mohammed O F, Sargent E H and Bakr O M 2015 *Science* **347** 519
[30] Dong Q, Fang Y, Shao Y, Mulligan P, Qiu J, Cao L and Huang J 2015 *Science* **347** 967
[31] Stranks S D, Eperon G E, Grancini G, Menelaou C, Alcocer M J P, Leijtens T, Herz L M, Petrozza A and Snaith H J 2013 *Science* **342** 341
[32] Tsai H, Nie W, Blancon J C, Stoumpos C C, Asadpour R, Harutyunyan B, Neukirch A J, Verduzco R, Crochet J J, Tretiak S, Pedesseau L, Even J, Alam M A, Gupta G, Lou J, Ajayan P M, Bedzyk M J, Kanatzidis M G and Mohite A D 2016 *Nature* **536** 312
[33] Yin W J, Shi T and Yan Y 2014 *Adv. Mater.* **26** 4653
[34] Zhao X G, Yang J H, Fu Y, Yang D, Xu Q, Yu L, Wei S H and Zhang L 2017 *J. Am. Chem. Soc.* **139** 2630
[35] Zhang C X, Shen T, Guo D, Tang L M, Yang K and Deng H X 2020 *InfoMat* **2** 1034
[36] Zhao Y, Li C and Shen L 2019 *InfoMat* **1** 164
[37] Li T, Luo S, Wang X and Zhang L 2021 *Adv. Mater.* **33** 2008574
[38] Xu Q, Yang D, Lv J, Sun Y Y and Zhang L 2018 *Small Methods* **2** 1700316
[39] Wang X, Li T, Xing B, Faizan M, Biswas K and Zhang L 2021 *J. Phys. Chem. Lett.* **12** 10532
[40] Liu Z, Zhao X, Zunger A and Zhang L 2019 *Advanced Electronic Materials* **5** 1900234
[41] Yang D, Lv J, Zhao X, Xu Q, Fu Y, Zhan Y, Zunger A and Zhang L 2017 *Chem. Mater.* **29** 524
[42] Zhao X G, Yang D, Ren J C, Sun Y, Xiao Z and Zhang L 2018 *Joule* **2** 1662
[43] Zhao D, Li T, Xu Q, Wang X and Zhang L 2019 *Chin. Opt.* **12** 964
[44] Yang J X, Zhang P, Wang J P and Wei S H 2020 *Chin. Phys. B* **29** 108401
[45] Jiang X W and Yin W J 2020 *Chin. Phys. B* **29** 028803
[46] Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q and Wang J 2018 *Nat. Commun.* **9** 3405
[47] Li Z, Xu Q, Sun Q, Hou Z and Yin W J 2019 *Advanced Functional Materials* **29** 1807280
[48] Ali A, Park H, Mall R, Aïssa B, Sanvito S, Bensmail H, Belaidi A and El-Mellouhi F 2020 *Chem. Mater.* **32** 2998
[49] Talapatra A, Uberuaga B P, Stanek C R and Pilania G 2021 *Chem. Mater.* **33** 845
[50] Weng B, Song Z, Zhu R, Yan Q, Sun Q, Grice C G, Yan Y and Yin W J 2020 *Nat. Commun.* **11** 1
[51] Pimachev A K and Neogi S 2021 *npj Comput. Mater.* **7** 1
[52] Ye W, Chen C, Wang Z, Chu I H and Ong S P 2018 *Nat. Commun.* **9** 1
[53] Ma X Y, Lyu H Y, Hao K R, Zhao Y M, Qian X, Yan Q B and Su G 2021 *Science Bulletin* **66** 233
[54] Kresse G and Furthmüller J 1996 *Phys. Rev. B* **54** 11169
[55] Kresse G and Furthmüller J 1996 *Computational Materials Science* **6** 15
[56] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[57] Grimme S 2006 *Journal of Computational Chemistry* **27** 1787
[58] Yang D, Fu Y, Sun Y, Li Y, Wang K, Xiao Z, Biswas K and Zhang L 2021 *Phys. Rev. Mater.* **5** 054603
[59] Tian F, Feng W, Xing B, He X, Saidi W A and Zhang L 2021 *Advanced Energy and Sustainability Research* **2** 2100087
[60] Xu Q, Stroppa A, Lv J, Zhao X, Yang D, Biswas K and Zhang L 2019 *Phys. Rev. Mater.* **3** 125401
[61] Zhao G, Xie J, Zhou K, Xing B, Wang X, Tian F, He X and Zhang L 2022 *Chin. Phys. B* **31** 037104
[62] Wang X, Fu Y, Na G, Li H and Zhang L 2019 *Acta Phys. Sin.* **68** 157101 (in Chinese)
[63] Li Y, Na G, Luo S, He X and Zhang L 2020 *Acta Phys. Chim. Sin.* **37** 2007015
[64] Zhao X G, Zhou K, Xing B, Zhao R, Luo S, Li T, Sun Y, Na G, Xie J, Yang X, Wang X, Wang X, He X, Lv J, Fu Y and Zhang L 2021 *Science Bulletin* **66** 1973
[65] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay é 2011 *Journal of Machine Learning Research* **12** 2825
[66] Chen T and Guestrin C 2016 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794
[67] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T Y 2017 *Advances in Neural Information Processing Systems* (Vol. 30)
[68] Bergstra J, Yamins D and Cox D 2013 *Proceedings of the 30th International Conference on Machine Learning* pp. 115–23
[69] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 *Computational Materials Science* **68** 314
[70] De Breuck P P, Hautier G and Rignanese G M 2021 *npj Comput Mater* **7** 83
[71] Cai X, Zhang Y, Shi Z, Chen Y, Xia Y, Yu A, Xu Y, Xie F, Shao H, Zhu H, Fu D, Zhan Y and Zhang H 2022 *Advanced Science* **9** 2103648
[72] Cai X, Zhang Y, Shi Z, Chen Y, Xia Y, Yu A, Xu Y, Xie F, Shao H, Zhu H, Fu D, Zhan Y and Zhang H *Advanced Science* **n/a** 2103648
[73] Gao W and Zhou Z H 2013 *Artificial Intelligence* **203** 1
[74] Breiman L 1996 *Mach Learn* **24** 123
[75] Rokach L 2010 *Artif. Intell. Rev.* **33** 1
[76] Hecht-nielsen R 1992 *Neural Networks for Perception* (ed H Wechsler) pp. 65–93