PLOS ONE

RESEARCH ARTICLE

Bandgap prediction of two-dimensional materials using machine learning

Yu Zhang[1]*, Wenjing Xu[1], Guangjie Liu[1]*, Zhiyong Zhang[1], Jinlong Zhu[1], Meng Li[2]

1 Department of Computer Science and Technology, Changchun Normal University, Changchun, China, **2** College of Information Science and Engineering, Shenyang University of Technology, Shenyang, China

* zhangyu@ccsfu.edu.cn (YZ); 756058599@qq.com (GL)

Citation: Zhang Y, Xu W, Liu G, Zhang Z, Zhu J, Li M (2021) Bandgap prediction of two-dimensional materials using machine learning. PLoS ONE 16(8): e0255637. https://doi.org/10.1371/journal.pone.0255637

Editor: Michael Loong Peng Tan, Universiti Teknologi Malaysia, MALAYSIA

Received: April 18, 2021

Accepted: July 20, 2021

Published: August 13, 2021

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: This research was funded by the National Natural Science Foundation of China, grant number: 61604019, Joint fund of Science & Technology Department of Liaoning Province and State Key Laboratory of Robotics, China, grant number: 2020-KF-22-08, and Science Research Fund Project of Liaoning Provincial Department of education, grant number: LQGD2019015.

Abstract

The bandgap of two-dimensional (2D) materials plays an important role in their applications to various devices. For instance, the gapless nature of graphene limits the use of this material to semiconductor device applications, whereas the indirect bandgap of molybdenum disulfide is suitable for electrical and photo-device applications. Therefore, predicting the bandgap rapidly and accurately for a given 2D material structure has great scientific significance in the manufacturing of semiconductor devices. Compared to the extremely high computation cost of conventional first-principles calculations, machine learning (ML) based on statistics may be a promising alternative to predicting bandgaps. Although ML algorithms have been used to predict the properties of materials, they have rarely been used to predict the properties of 2D materials. In this study, we apply four ML algorithms to predict the bandgaps of 2D materials based on the computational 2D materials database (C2DB). Gradient boosted decision trees and random forests are more effective in predicting bandgaps of 2D materials with an R^2 >90% and root-mean-square error (RMSE) of ~0.24 eV and 0.27 eV, respectively. By contrast, support vector regression and multi-layer perceptron show that R^2 is >70% with RMSE of ~0.41 eV and 0.43 eV, respectively. Finally, when the bandgap calculated without spin-orbit coupling (SOC) is used as a feature, the RMSEs of the four ML models decrease greatly to 0.09 eV, 0.10 eV, 0.17 eV, and 0.12 eV, respectively. The R^2 of all the models is >94%. These results show that the properties of 2D materials can be rapidly obtained by ML prediction with high precision.

Introduction

The bandgap is a significant electronic property of certain materials that has been utilized for the fabrication of semiconductor devices. Moore's law has dominated the chip manufacturing industry for over 50 years. However, Moore's law is nearing its end, as reported by *Nature* [1]. One of the promising methods to prolong this era is to replace silicon, the most common material used to make chips, with other materials for next-generation electronics [2–4]. Owing to their unique nanosheet structure, large surface area, and extraordinary physical and chemical properties, atomically thin two-dimensional (2D) materials have attracted considerable attention since the successful isolation and characterization of graphene in 2004 [5].

Competing interests: The authors have declared that no competing interests exist.

Two-dimensional materials have become the most promising substitutes for silicon. It is to be noted that 2D materials with different structures have different electrical properties. For instance, intrinsic graphene is gapless, which makes it impossible to switch off [6–8]. This major shortcoming of graphene limits its range of potential device applications. Conversely, molybdenum disulfide (MoS_2), which originates from the transition metal dichalcogenide (TMDC) family, is a semiconductor with an indirect bandgap of ~1.89 eV [9, 10]. In addition, hexagonal boron nitride (h-BN) is an insulator with a wide gap of ~ 5.9 eV [11, 12]. Therefore, considering the diverse properties of different 2D materials, the priority is to study the electrical properties of 2D materials used in manufacturing semiconductor devices.

Based on this motivation, it is anticipated that many theoretical and experimental studies have been conducted on their electronic properties [13–15]. Although the results of experimental studies are more reliable, it is difficult to carry out experiments on these thin materials owing to experimental constraints. Theoretical research is an effective alternative to experimental research. Over the past decades, conventional first-principles calculations have been a powerful tool for calculating the structures and properties of materials [16–18]. For example, they have been used to study the structural and electronic properties of perfect, doped, and defective 2D materials [19–21], the interaction between the substrate and 2D materials [22, 23], and the interaction between the contacts and 2D materials [24, 25]. Although the results calculated by the first-principles theory are usually consistent with experimental results, this method is computationally expensive and time consuming [26, 27].

Recent progress in machine learning (ML), a data-driven technique, has been effectively used for material research [28–37]. For instance, ML has been used in guiding chemical synthesis, assisting material characterization, and designing new materials [38]. In particular, ML methods are considered suitable for predicting a series of material properties. Cherukara et al. created the first atomic-level model using ML to accurately predict the thermal properties of stanene [39]. Dieb et al. employed ML to determine the most stable structures of boron-doped graphene [40]. Wan et al. developed a convolutional neural network (CNN) model to predict the thermal conductivity of porous graphene [41]. Dong et al. developed deep learning algorithms to predict the bandgaps of hybrids of graphene and h-BN with arbitrary supercell configurations [42]. Baboukani et al. presented an ML method for predicting nanoscale friction in 2D materials [43]. Moreover, ML interatomic potentials have shown outstanding efficiency in predicting novel materials [44, 45], lattice dynamics [46], estimating the thermal conductivity [47, 48], and exploring the phononic properties of 2D materials [49].

In this study, we employ random forest (RF), support vector regression (SVR), gradient boosted decision tree (GBDT), and multi-layer perceptron (MLP) models to predict the bandgaps of 2D materials based on the open C2DB database. First, eight elemental features were selected to train our ML models. The GBDT and RF models were found to be superior to the SVR and MLP with a prediction accuracy >90% and prediction RMSE of 0.24 eV, and 0.27 eV, respectively. In addition, when the bandgap without SOC is added into the elemental feature space, the prediction RMSE of GBDT, RF, SVR, and MLP decreases to 0.09 eV, 0.10 eV, 0.17 eV, and 0.12 eV, respectively. Furthermore, the prediction accuracy R^2 of the four ML models is 98%, 98%, 95%, and 97%, respectively.

Materials and methods

Machine learning

Machine learning in this study was performed in Python 3.6 code with Scikit-learn frameworks for SVR, MLP, RF, and GBDT. All hyperparameters were optimized, and model performance was evaluated by grid search based on the averaged RMSE of the validation set.

In existing studies, most data used for the prediction of 2D materials were calculated based on the first-principles theory. Consequently, the accuracy of the prediction is easily challenged by the root cause. The data used here are from the C2DB, which is a common 2D material database [50]. The database contains approximately 4000 different 2D monolayer crystal structures. The dataset was split into training and test datasets. The training and validation datasets were 2817 and 313, respectively. For the SVR method, the prediction value for the bandgap is

$$\hat{f}_{SVR}(x) = \sum_{i=1}^{N}(\hat{a} - \alpha^*)K(x_i, x) + b \tag{1}$$

where K is the kernel function used to measure the difference between the training data x_i and prediction data x.

The kernel function is the radial basis function (RBF) denoted by

$$K(x_i, x) = \exp(-\frac{\|x_i - x\|^2}{2\sigma^2})$$

with the width $\sigma > 0$.

The b can be obtained by solving the following Lagrange function

$$T(\omega, b, \zeta_i, \zeta_i^*) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{N}(\zeta_i + \zeta_i{}^*) \tag{2}$$

$$\text{s.t. } y_i - \hat{f}_{SVR}(x_i) \leq \varepsilon + \zeta_i$$

$$\hat{f}_{SVR}(x_i) - y_i \leq \varepsilon + \zeta_i^*$$
$$\zeta_i, \zeta_i^* \geq 0, i = 1, \cdots, N$$

where C is the regularization parameter.

For the MLP model, the prediction value for the bandgap is

$$\hat{f}_{MLP}(x) = \sum_{i=1}^{N} W_o^{ji}\phi(W_l^{ji}x_i + b_l) + b_o \tag{3}$$

where $\varphi(z) = z$ is the activation function, W_l^{ji} and $W_o{}^{ji}$ represents the weight of the jth neuron in the l_{th} and output layers connected to the ith neuron in the $(l\text{-}1)_{\text{th}}$ and $(o\text{-}1)_{\text{th}}$ layers, respectively; b_1,b_0 are the biases of the hidden and output layers, respectively.

The weights $W = (W_1, W_2, \cdots W_m)^{\text{T}}$ are repeatedly updated to minimize the loss function

$$L(W_1 \cdots W_m) = \frac{1}{2}\|\hat{f}_{MLP}(x_i) - y_i\|_2^2 + \frac{\alpha}{2}\|W\|_2^2 \tag{4}$$

where $a > 0$ is a non-negative tuning parameter that controls the magnitude of the penalty $\frac{\alpha}{2}\|W\|_2^2$.

The weight is updated by

$$W^{i+1} = W^i - \eta\frac{\partial L^i}{\partial W} \tag{5}$$

where η is the learning rate.

For the GBDT and RF models, the algorithmic details are provided in Figs 1 and 2.

Algorithm Gradient Boosting Decision Tree

1: Initialize $F_{pb0}(\mathrm{x})$ by $F_{pb0}(\mathrm{x}) = \arg\min \sum_{i=1}^{N} L(y_i, \theta)$, where L is loss function.

2: Calculate the negative gradient of the M_{th} tree' s loss function, $\mathrm{g}_{m,i} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x)}]_{F(x)=F_{m-1}(x)}$, where M is the number of the regression tree, $m = 1,2,\cdots M$; N is the number of the testing set, $i = 1,2,\cdots N$.

3: Obtain the M_{th} regression tree by fitting the data $(\mathrm{x}_i, g_{m,i})$ based on CART regression tree.

4: Calculate the best fitting value $\theta^*_{m,j} = \arg\min \sum_{\mathrm{x}_i \in R_{m,j}} L(\mathrm{y}_i, F_{\mathrm{m}-1}(x_i) + \theta)$, where $R_{m,j}$ is the set of leaf nodes, $j = 1,2,\cdots J_m$, J_m is the number of leaf nodes.

5: Update $F_{\mathrm{m}}(x)$, $F_{\mathrm{m}}(x) = F_{\mathrm{m}\text{-}1}(x) + \sum_{j=1}^{J_{\mathrm{m}}} \theta_{m,j} I(x \in R_{\mathrm{m,j}})$.

6: The predicted bandgap $F_{pb}(\mathrm{x})$ is obtained by $F_{pb}(\mathrm{x}) = F_{pb0}(\mathrm{x}) + \sum_{m=1}^{M} \sum_{j=1}^{J_{\mathrm{m}}} \theta_{m,j} I(x \in R_{m,j})$.

Fig 1. Gradient boosting decision tree algorithm.

https://doi.org/10.1371/journal.pone.0255637.g001

Algorithm Random Forest

1: Input training data set $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_i, y_i)\}$

2: Construct the t_{th} tree' s training set D_t by randomly choosing n training samples from the training set D with return, where $n < i$.

3: Solve $(j^*, s^*) = \min_{\mathrm{j,s}}[\min_{c_1} \sum_{x_i \in D_{t1}(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in D_{t2}(j,s)} y_i - c_2]$ to find minimum segmentation variable j and segmentation point s, where c_1 is the mean value of the dataset D_{t1} and c_2 is the mean value of the dataset D_{t2}.

4: Use (j^*, s^*) to divide D_t and construct the t_{th} CART tree by $f_t(x) = \sum_{i=1}^{m} \hat{c}_{ti} I(x \in D_{ti}), i = 1,2$ where $D_{t1}(j^*, s^*) = \{x \mid x^{(j^*)} \le s^*\}, D_{t2}(j^*, s^*) = \{x \mid x^{(j^*)} > s^*\}$, and $\hat{c}_{tm} = \frac{1}{N_m} \sum_{x_i \in D_{tm}(j,s)} y_i, x \in D_{tm}, m = 1,2$.

5: Goto step 1 and 2 until the stop condition is met.

6: The predicted bandgap is obtained by $f_{pb}(x) = \sum_{k=1}^{K} \sum_{i=1}^{m} \hat{c}_{ti} I(x \in D_{ti}) / K, k = 1,2 \cdots K$ where K is the number of the CART trees constructed.

Fig 2. Random forest algorithm.

https://doi.org/10.1371/journal.pone.0255637.g002

Criteria of evaluation

We used the mean absolute error (MAE), RMSE, and explained variance (R^2) to evaluate the prediction accuracy of each model on the test set.

$$MAE = \frac{1}{k}\sum_{i=1}^{k}|y_{tb}^{i} - \hat{y}_{pb}^{i}| \tag{6}$$

$$RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_{tb}^{i} - \hat{y}_{pb}^{i})^2} \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{k}(y_{tb}^{i} - \hat{y}_{pb}^{i})^2}{\sum_{i=1}^{k}(y_{tb}^{i} - \bar{y}_{tb}^{i})^2} \tag{8}$$

In formulas (6)–(8),y_{tb}^{i} is the true bandgap value randomly selected from the test set, $\bar{y}_{tb}^{i}$ is the average value of y_{tb}^{i},$\hat{y}_{pb}^{i}$ is the predicted value of the corresponding regression model, and i = 1,2,. . .,k, where k = 313.

Spearman method

The spearman correlation coefficient is calculated by [51]

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{9}$$

where X_i is the value of the i_{th} feature x in the training set, $\bar{x}$ is the average value of all the features, y_i is the value of the i_{th} feature y, $\bar{y}$ is its average value, and i = 1, 2, . . ., N, where N is the total number of samples in the training set.

Results and discussions

Feature selection is a key step in ML algorithms. A good sample dataset can improve the performance of ML algorithms. The most common elemental information, including Dosef, Hform, Natoms, Mass, Cellarea, Energy, Smax, Fmax, and Volume, was first chosen for feature correlation analysis. The elemental features selected are listed in Table 1. The Pearson linear

Table 1. Definitions of the features selected [50].

Feature Name	Definitions
Dosef	density of states at the Fermi energy
Hform	heat of formation
Natoms	number of atoms
Mass	sum of atomic masses in unit cell
Energy	total energy
Fmax	maximum force
Smax	maximum stress on unit cell
Volume	volume of unit cell
Gap_nosoc	gap w/o soc

https://doi.org/10.1371/journal.pone.0255637.t001

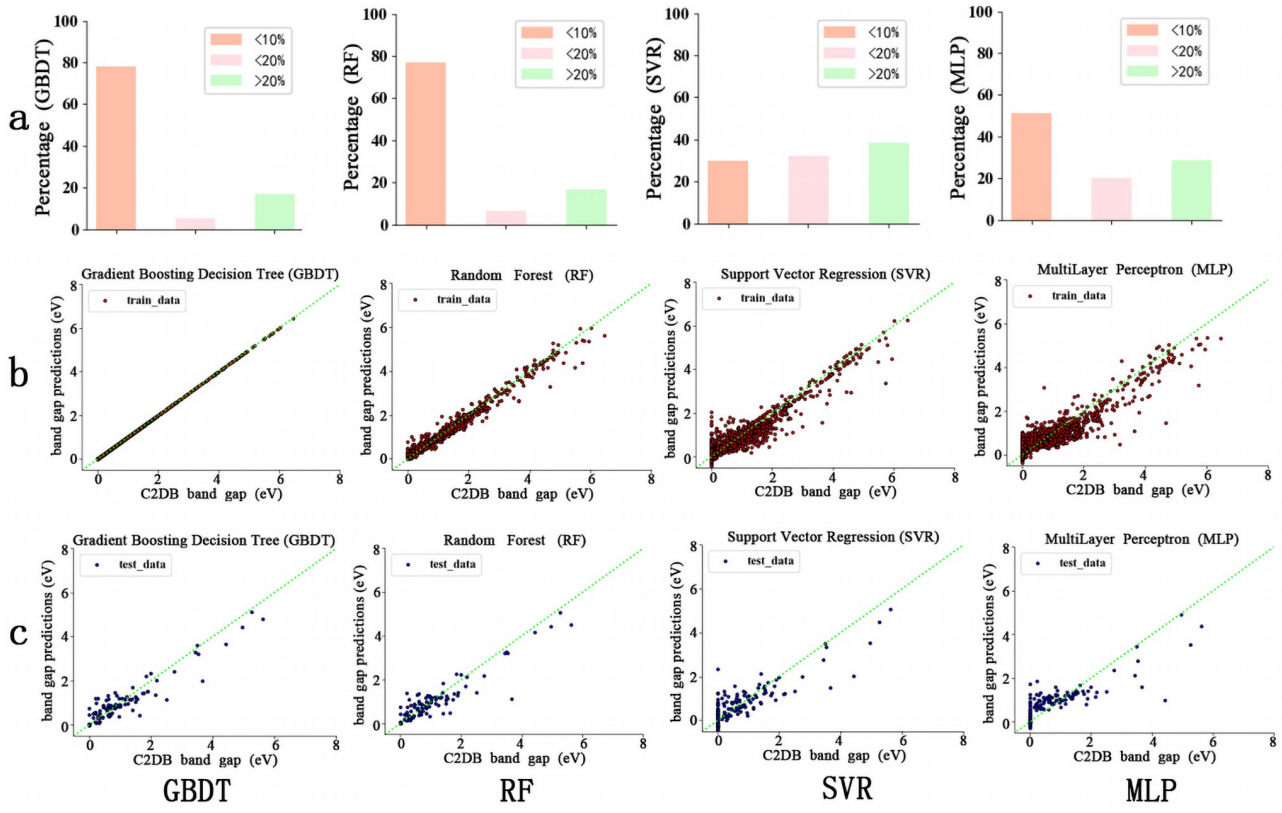

Fig 3. Prediction performance of four ML algorithms (GBDT, RF, SVR and MLP) without the feature Gap_nosoc. (a) Error levels of ML predicted bandgaps (b) Fitness between predicted bandgaps and true bandgaps on the training sets (c) Fitness between predicted bandgaps and true bandgaps on the test sets.

https://doi.org/10.1371/journal.pone.0255637.g003

correlation coefficient map can analyze and identify features with high correlation, and eliminate the multiple collinearities between features, which leads to the distortion or inaccuracy of model estimation; this is particularly evident in linear models, such as SVR. Therefore, among many features with a strong correlation (correlation value greater than 0.8), only one effective feature is reserved. However, the Pearson correlation coefficient is not stable and is affected by the outlier values. Thus, scatter plots are also used. Scatter plots of features not only show outlier values but also show a high feature correlation. The scatter plots of the feature correlation and Pearson linear correlation coefficient map are provided in S1 and S2 Figs.

Four types of ML models: SVR, MLP, RF, and GBDT are used. The performance comparison of the four ML models with the 8-dimensional feature space (without Gap_nosoc) is shown in Fig 3. The prediction accuracy is characterized by the absolute error of the predicted bandgaps (E_{ML}) to the C2DB bandgaps (E_{C2DB}), which is calculated as $|E_{ML}-E_{C2DB}|$. As shown in Fig 3(a), RF and GBDT can predict the bandgaps within 10% absolute error for approximately 80% of the cases, whereas MLP shows approximately 50% of the cases. By contrast, the prediction results from SVR deviate much more from the C2DB benchmarks, showing >20% error for approximately 40% of the cases. Fig 3(b) and 3(c) show the fitness between the predicted bandgaps and true bandgaps on the training and test sets, respectively. The RF and GBDT exhibit a strong direct linear correlation between the predicted ML values and the C2DB values, whereas the SVM and MLP show weaker correlations. The predicted bandgaps of the four models for new materials with bandgap values between 4 and 6 showed relatively

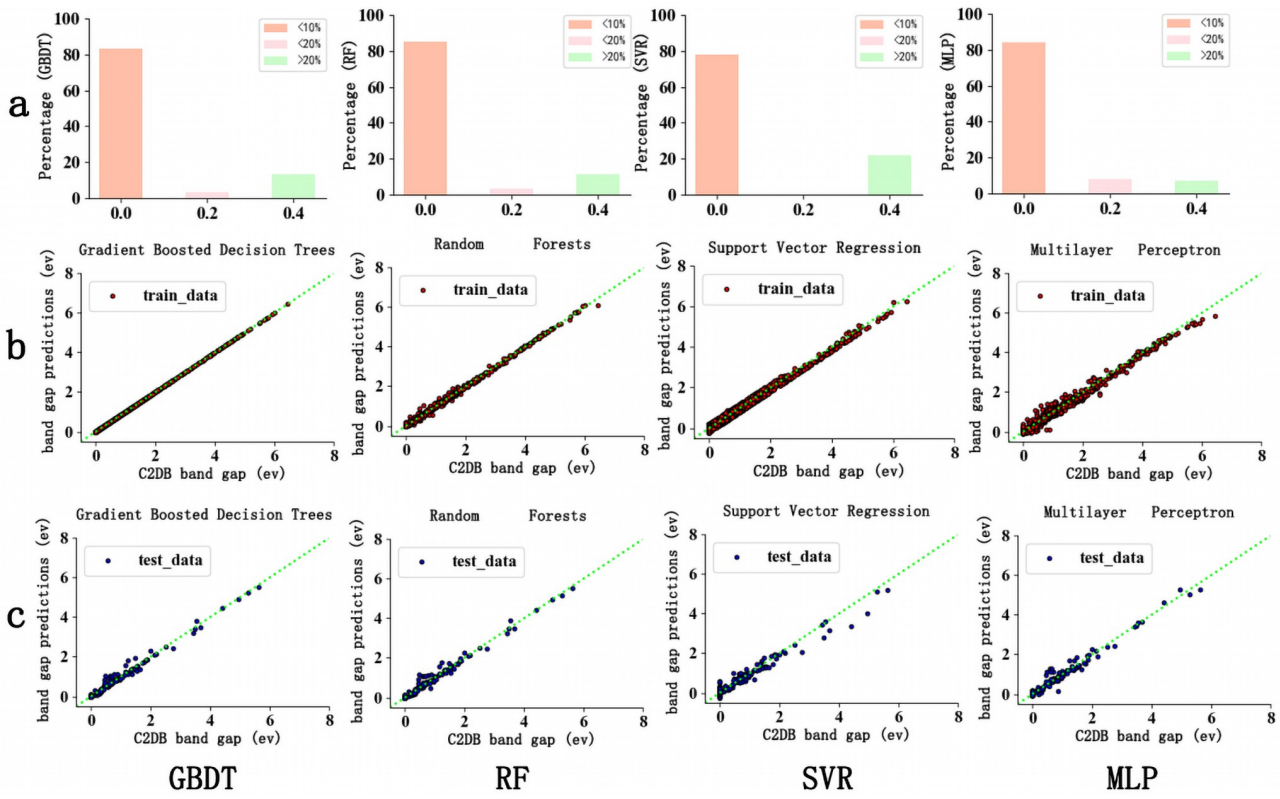

Fig 4. Prediction performance of four ML algorithms (GBDT, RF, SVR and MLP) with the feature Gap_nosoc. (a) Error levels of ML predicted bandgaps (b) Fitness between predicted bandgaps and true bandgaps on the training sets (c) Fitness between predicted bandgaps and true bandgaps on the test sets.

https://doi.org/10.1371/journal.pone.0255637.g004

deviations, especially for SVR and MLP. This is because the amount of data with bandgap values between 4 and 6 in the training set is small. Consequently), the model cannot obtain a better prediction ability for new materials with bandgap values between 4 and 6 in the learning process.

When the bandgap calculated without SOC was used as a feature, the performance of all four ML models was greatly enhanced, as shown in Fig 4. Fig 4(a) shows that the GBDT, RF, and MLP can predict the bandgap within 10% absolute error for >80% of the cases, and SVR shows >75% of the cases. The four ML models exhibit a very strong direct linear correlation between the predicted ML values and the C2DB values for both the training and test sets, as shown in Fig 4(b) and 4(c). The predicted bandgap values between 4 and 6 improved significantly. Only SVR showed relatively slight deviations.

Other indicators of the MAE, RMSE, and R^2 were used to evaluate the regression performance of the ML models. Normally, the MAE and RMSE are much smaller and closer to 0, and R^2 is greater and closer to 1, indicating better performance and higher prediction accuracy of the model. The MAE, RMSE, and R^2 of the four ML models based on an 8-dimensional feature space are provided in Table 2. The GBDT shows the lowest RMSE of ~0.24 eV for the predicted bandgaps. The RMSE values slightly increase to 0.27 eV for the RF. By contrast, the prediction accuracy of the SVR and MLP algorithms is lower, showing higher RMSEs of 0.41 eV and 0.47 eV, respectively. R^2 is an indicator of the correlation between the prediction and real values and is considered one of the most important metrics for evaluating the accuracy of the prediction models. Table 2 illustrates the bandgaps predicted by the GBDT and RF with

Table 2. Statistics of predicted bandgaps by SVR, GBDT, RF and MLP algorithms based on an 8-dimensional feature space.

Models	Parameters	MAE	RMSE	R^2
SVR	C = 50, epsilon = 0.2 gamma = 50,kernel = rbf	0.10	0.41	0.75
GBDT	n_estimators = 21000, max_depth = 21, min_samples_split = 5, max_features = 0.8, learning_rate = 0.001	0.12	0.24	0.92
RF	n_estimators = 15000, max_depth = 20, min_samples_split = 5,max_features = 0.8, min_samples_leaf = 3	0.10	0.27	0.90
MLP	solver = adam, hidden_layer_sizes = (262,140,139,180), activation = tanh,alpha = 1e-8, tol = 1e-6, max_iter = 5000, learning_rate_init = 0.01	0.24	0.43	0.73

https://doi.org/10.1371/journal.pone.0255637.t002

Table 3. Statistics of predicted bandgaps by SVR, GBDT, RF and MLP algorithms based on a 9-dimensional feature space.

Models	Parameters	MAE	RMSE	R^2
SVR	C = 50, epsilon = 0.2 gamma = 50,kernel = rbf	0.11	0.17	0.95
GBDT	n_estimators = 21000, max_depth = 21, min_samples_split = 5, max_features = 0.8, learning_rate = 0.001	0.03	0.09	0.98
RF	n_estimators = 15000, max_depth = 20, min_samples_split = 5,max_features = 0.8, min_samples_leaf = 3	0.03	0.10	0.98
MLP	solver = adam, hidden_layer_sizes = (262,140,139,180), activation = tanh,alpha = 1e-8, tol = 1e-6, max_iter = 5000, learning_rate_init = 0.01	0.06	0.12	0.97

https://doi.org/10.1371/journal.pone.0255637.t003

~92% and 90% relevance to the values from the C2DB database. Both SVR and MLP had >70% relevance. Both RF and GBDT are known to be more effective in predicting the bandgaps of 2D materials than other models. Among them, GBDT yields the best prediction results. This is because of the fusion of the decision tree and the gradient descent algorithm. Thus, the model has the advantages of high prediction accuracy, ability to deal with nonlinear data, flexibility to deal with various data types, stronger fitting ability, and better generalization ability for unknown new datasets. Table 3 shows the MAE, RMSE, and R^2 of the four ML models based on a 9-dimensional feature space. The prediction precision was greatly improved when the bandgap calculated without SOC was added to the feature space. The bandgaps predicted by the four models had >94% relevance to the values from the C2DB database. The RMSE values of all the models fall drastically to 0.17 eV, 0.09 eV, 0.10 eV, and 0.12 eV, respectively. In summary, these results show that the RF and GBDT are superior to the SVR and MLP without highly relevant features in predicting the bandgaps of 2D materials. Provided that highly relevant features are chosen, ML models are less important and show high prediction accuracy.

The correlation index between the features used for model training and the target was calculated by the Spearman method, which was used to judge the importance of features to the target. In addition, the RF and GBDT models have their feature importance scoring functions. Feature importance is shown in Fig 5. For the 8-dimensional feature space, as shown in Fig 5(a), the Spearman correlation coefficient indicates that the features of "Dosef," "Hform," "Volume," "Smax," and "Energy" have relatively high scores. For the RF and GBDT, the features of "Dosef," "Hform," and "Energy" have relatively high scores. For the 9-dimensional feature space, as shown in Fig 5(b), both "Dosef" and "Gap_nosoc" have the highest scores based on the Spearman correlation coefficient. For the RF and GBDT, the features of "Dosef," "Hform," and "Gap_nosoc" have relatively high scores. These results reveal that the features of "Dosef," "Hform," and "Gap_nosoc" have a high influence on the performance of the training model.

Conclusion

Recently, researchers have performed ML to investigate the properties of 2D materials. Common databases of 2D materials are few and less developed. Therefore, most published studies used their calculated data to train ML models. The prediction accuracy of these data can be

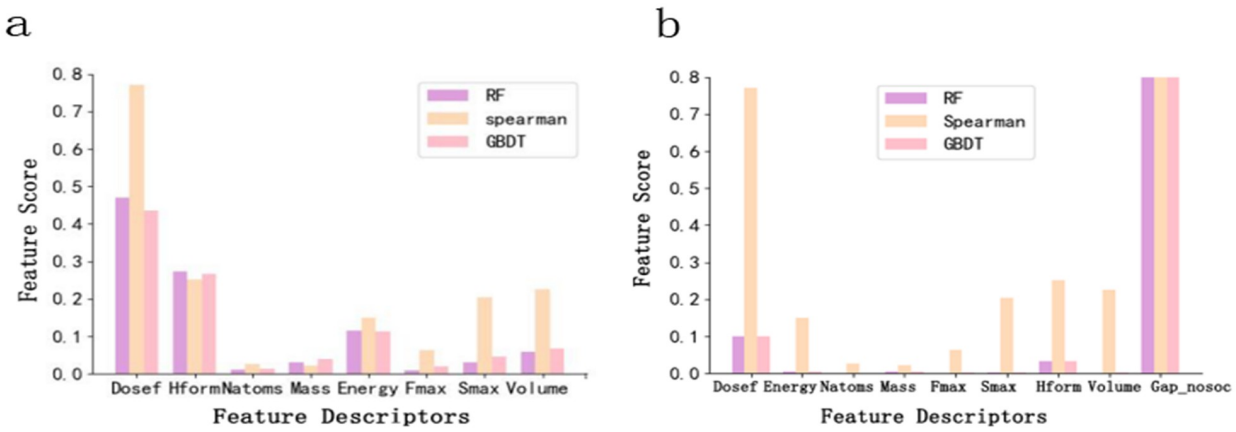

Fig 5. Feature importance evaluation. (a) 8-dimensional feature space (b) 9-dimensional feature space.

https://doi.org/10.1371/journal.pone.0255637.g005

challenged by the root. Although all the data used here are available from the C2DB database, feature selection is more important. It is easy to train the models to obtain high prediction accuracy in other fields owing to the availability of a large amount of data; however, in the field of materials, data are very few and limited. Determining the highly relevant features from the limited datasets and training the models to obtain a high prediction accuracy is the key issue. In this study, four ML models trained on a common 2D material database were utilized to successfully predict the bandgap of 2D crystals. When Gap_nosoc is not added to the feature space, the GBDT and RF models yield prediction accuracy >90%, whereas the SVR and MLP models show lower prediction accuracy (70%). The RMSE of the GBDT is ~0.24 eV, whereas that of the SVR and MLP is ~0.41 eV. After Gap_nosoc is added to the feature space, the prediction accuracy of the four ML models is significantly improved. The RMSE of the four ML models decreases significantly to 0.09 eV, 0.10 eV, 0.17 eV, and 0.12 eV, respectively. The R^2 of all the models is >94%. The results indicate that features highly relevant to the target have a major influence on the performance of the training model. In contrast to three-dimensional materials, the electrons in 2D materials are inherently constrained in the direction perpendicular to the material, which make 2D materials naturally have many unique properties. The bandgap is an essential electronic property of the materials. The width of the bandgap determines whether the material is a semiconductor or an insulator. Therefore, we first focused on predicting the bandgaps of 2D materials. In the future, we will investigate the prediction of other properties of 2D materials, such as magnetic properties, electron mobility, and mechanical stability.

Supporting information

S1 Fig. Scatter plot of the feature correlation.
(DOCX)

S2 Fig. Pearson linear correlation coefficient map.
(DOCX)

S1 Data. Availability of data and material.
(RAR)

S1 Code. Code used to build the model.
(DOC)

Author Contributions

Funding acquisition: Yu Zhang, Meng Li.

Investigation: Wenjing Xu.

Methodology: Yu Zhang.

Resources: Guangjie Liu.

Software: Wenjing Xu, Jinlong Zhu.

Supervision: Yu Zhang.

Writing – original draft: Yu Zhang.

Writing – review & editing: Yu Zhang, Guangjie Liu, Zhiyong Zhang, Meng Li.

References

1. Waldrop MM. The chips are down for Moore's law. Nature. 2016; 530(7589):144–147. https://doi.org/10.1038/530144a PMID: 26863965
2. Ghazanfar N, Muhammad K, Sikandar A, Amir A, Ghulam D, Malik R, et al. Gate tunable transport in Graphene/MoS2/(Cr/Au) vertical field-effect transistors. Nanomaterials. 2018; 8(01):14. https://doi.org/10.3390/nano8010014 PMID: 29283377
3. Nekrasov N, Kireev D, Omerovi N, Emelianov A, Bobrinetskiy I. Photo-induced doping in a graphene field-effect transistor with inkjet-printed organic semiconducting molecules. Nanomaterials. 2019; 9 (12):1753. https://doi.org/10.3390/nano9121753 PMID: 31835474
4. Das S, Pandey D, Thomas J, Roy T. 2D Materials: The role of graphene and other 2D materials in solar photovoltaics. Advanced Materials. 2019; 31(01):1970006. https://doi.org/10.1002/adma.201970006
5. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, et al. Electric Field Effect in Atomically Thin Carbon Films. Science. 2004; 306(5696):666–669. https://doi.org/10.1126/science.1102896 PMID: 15499015
6. Deng YX, Chen SZ, Zhang Y, Yu X, Xie ZX, Chen KQ, et al. Penta-hexa-graphene nanoribbons: intrinsic magnetism and edge effect induce spin-gapless semiconducting and half-metallic properties. ACS Applied Materials And Interfaces. 2020; 12(47): 53088–53095. https://doi.org/10.1021/acsami.0c14768 PMID: 33197167
7. Wang XX, Shen NF, Wu J, Wang BL, Wan JG. Predicting quantum spin hall effect in graphene/gasb and normal strain-controlled band structures. Applied Surface Science. 2020; 526:146704. https://doi.org/10.1016/j.apsusc.2020.146704
8. Dragoman M, Dinescu A, Nastase F, Modovan A, Dragoman D. Graphene bandgap induced by ferroelectric HfO2 doped with Zr (HfZrO). Nanotechnology. 2020; 31:275202. https://doi.org/10.1088/1361-6528/ab814b PMID: 32191931
9. Mak KF, Lee C, Hone J, Shan J, Heinz TF. Atomically Thin MoS2: A New Direct-Gap Semiconductor. 2010; 105(13):136805. https://doi.org/10.1103/PhysRevLett.105.136805 PMID: 21230799
10. Yoon A, Kim JH, Yoon JC, Lee YD, Lee ZH. van der Waals Epitaxial Formation of Atomic Layered α-MoO3 on MoS2 by Oxidation. ACS Applied Materials & Interfaces. 2020; 12(19):22029–22036. https://doi.org/10.1021/acsami.0c03032 PMID: 32298075
11. Yang K, Liu H, Wang S, Li W, Han T. A horizontal-gate monolayer MoS2 transistor based on image force barrier reduction. Nanomaterials. 2019; 9(09):1245. https://doi.org/10.3390/nano9091245 PMID: 31480685
12. Jaiswal HN, Liu MM, Shahi S, Wei SC, Lee J. Chakravarty A, et al. Diode-like selective enhancement of carrier transport through metal-semiconductor interface decorated by monolayer boron nitride. Advanced Materials. 2020; 32(36):2002716. https://doi.org/10.1002/adma.202002716 PMID: 32725788
13. Lin ZY, Wang C, Chai Y. Emerging group-vi elemental 2d materials: preparations, properties, and device applications. Small. 2020; 16(41):2003319. https://doi.org/10.1002/smll.202003319 PMID: 32797721
14. Zhao N, Zhu YF, Jiang Q. Novel electronic properties of two-dimensional asxsby alloys studied using DFT. Journal of Materials Chemistry C. 2018; 6(11):2854–2861. https://doi.org/10.1039/C8TC00079D

15. Ghasemi R, Jamilpanah L, Faridi E, Hajiali MR, Shafei M, Mohseni SM, et al. Electrical and magneto-optical characterization of Py/MoS2 bilayer: A facile growth of magnetic-metal/semiconductor heterostructure. Materials Letters. 2020; 265:127454. https://doi.org/10.1016/j.matlet.2020.127454

16. Ebadi M, Marchiori C, Mindemark J, Brandell D, Araujo CM. Assessing structure and stability of polymer/lithium-metal interfaces from first-principles calculations. Journal of Materials Chemistry A. 2019; 7(14):8394–8404. https://doi.org/10.1039/C8TA12147H

17. Luo BC, Wang XH, Tian EK, Song HZ, Qu HQ, Cai ZM, et al. Mechanism of ferroelectric properties of (BaCa)(ZrTi)O3 from first-principles calculations. Ceramics International. 2018; 44(08):9684–9688. https://doi.org/10.1016/j.ceramint.2018.02.197

18. Shahrokhi M, Raybaud P, Bahers TL. On the understanding of the optoelectronic properties of S-doped MoO3 and O-doped MoS2 bulk systems: a DFT perspective. Journal of Materials Chemistry C. 2020; 8(26):9064–9074. https://doi.org/10.1039/D0TC02066D

19. Sukhanova EV, Kvashnin DG, Popov ZI. Induced spin polarization in graphene via interactions with halogen doped MoS2 and MoSe2 monolayers by dft calculations. Nanoscale. 2020; 12(45):23248–23258. https://doi.org/10.1039/d0nr06287a PMID: 33206100

20. Wang WD, Yang CG, Bai LW, Li ML, Li WB. First-principles study on the structural and electronic properties of monolayer MoS2 with s-vacancy under uniaxial tensile strain. Nanomaterials. 2018; 8(02):74. https://doi.org/10.3390/nano8020074 PMID: 29382182

21. Rad AS. First principles study of al-doped graphene as nanostructure adsorbent for NO2 and N2O: DFT calculations. Applied Surface Science. 2015; 357:1217–1224. https://doi.org/10.1016/j.apsusc.2015.09.168

22. Dai ZH, Liu LQ, Zhang Z. Strain Engineering of 2D Materials: Issues and Opportunities at the Interface. Advanced Materials. 2019; 31(45):1805417. https://doi.org/10.1002/adma.201805417 PMID: 30650204

23. Gao YX, Zhang YY, Du SX. Recovery of the Dirac states of graphene by intercalating two-dimensional traditional semiconductors. Journal of Physics Condensed Matter. 2019; 31(19):194001. https://doi.org/10.1088/1361-648X/ab05a6 PMID: 30736029

24. Niggas A, Schwestka J, Creutzburg S, Aumayr F, Wilhelm RA. The role of contaminations on the interaction of highly charged ions with 2D materials. Journal of Physics: Conference Series. 2020; 1412 (20):202011. https://doi.org/10.1088/1742-6596/1412/20/202011

25. Fernández L, Alves VS, Nascimento LO, Peña F,. Gomes M, Marino EC. Renormalization of the band gap in 2D materials through the competition between electromagnetic and four-fermion interactions in large N expansion. Physical Review D. 2020; 102(01):016020. https://doi.org/10.1103/PhysRevD.102.016020

26. Esteban-Puyuelo R, Sarma DD, Sanyal B. Complexity of mixed allotropes of MoS2 unraveled by first-principles theory. Physical Review B. 2020; 102(16):165412. https://doi.org/10.1103/PhysRevB.102.165412

27. Jung JH, Kim SH, Park Y, Lee D, Lee JS. Metal-halide perovskite design for next-generation memories: first-principles screening and experimental verification. Advanced Science. 2020; 7(16):2001367. https://doi.org/10.1002/advs.202001367 PMID: 32832372

28. Hao J, Zhang HJ, Li JW, Wang T, Wan LH, Wei YD, et al. Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. Journal of Physical Chemistry Letters. 2020; 11 (08):3075–3081. https://doi.org/10.1021/acs.jpclett.0c00721 PMID: 32239944

29. De Luna P, Wei J, Bengio Y, Aspuru-Guzik A, Sargent E. Use machine learning to find energy materials. Nature. 2017; 552(7683):23–27. https://doi.org/10.1038/d41586-017-07820-6

30. Sparks TD, Kauwe SK, Parry ME, Tehrani AM, Brgoch J. Machine learning for structural materials. Annual Review of Materials Research. 2020; 50(01): 27–48. https://doi.org/10.1146/annurev-matsci-110519-094700

31. Wang T, Zhang C, Snoussi H, Zhang G. Machine learning approaches for thermoelectric materials research. Advanced Functional Materials. 2020; 30(05): 906041. https://doi.org/10.1002/adfm.201906041

32. Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid- state materials science. npj Computational Materials. 2019; 5:83. https://doi.org/10.1038/s41524-019-0221-0

33. Oda H, Kiyohara S, Mizoguchi T. Machine learning for structure determination and investigating the structure-property relationships of interfaces. Journal Physics Materials. 2019; 2(03):034005. https://doi.org/10.1088/2515-7639/ab15c8

34. Schleder GR, Padilha ACM, Acosta CM, Costa M, Fazzio A. From dft to machine learning: recent approaches to materials science–a review. Journal of Physics Materials. 2019; 2(03):032001. https://doi.org/10.1088/2515-7639/ab084b

35. Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid- state materials science. npj Computational Materials. 2019; 5:83. https://doi.org/10.1038/s41524-019-0221-0

36. Carleo G, Cirac I, Cranmer K, Daudet L, Zdeborová L. Machine learning and the physical sciences. Review Modern Physics. 2019; 91(04):045002. https://doi.org/10.1103/RevModPhys.91.045002

37. Han BN, Lin YX, Yang YF, Mao NN, Li WY, Wang HZ, et al. Deep-learning-enabled fast optical identifification and characterization of two-dimensional materials. Advanced Materials. 2019; 32(29):200953. https://doi.org/10.1002/adma.202000953 PMID: 32519397

38. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine Learning for Molecular and Materials Science. Nature. 2018; 559(7715):547–555. https://doi.org/10.1038/s41586-018-0337-2 PMID: 30046072

39. Cherukara MJ, Narayanan B, Kinaci A, Sasikumar K, Gray SK, Chan MKY, et al. Ab-initio based bond order potential to investigate low thermal conductivity of stanene nanostructures. Journal of Physical Chemistry Letters. 2016; 7(19):3752–3759. https://doi.org/10.1021/acs.jpclett.6b01562 PMID: 27569053

40. Dieb MT, Hou ZF, Tsuda K. Structure prediction of boron-doped graphene by machine learning. Journal of Chemical Physics. 2018; 148(24):241716. https://doi.org/10.1063/1.5018065 PMID: 29960333

41. Wan J, Jiang JW, Park HS. Machine learning-based design of porous graphene with low thermal conductivity. Carbon. 2020; 157:262–269. https://doi.org/10.1016/j.carbon.2019.10.037

42. Dong Y, Wu CH, Zhang C, Liu YD, Cheng JL, Lin J. Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. Npj Computational Materials. 2019; 5:26. https://doi.org/10.1038/s41524-019-0165-4

43. Baboukani BS, Ye ZJ, Reyes KG, Nalam PC. Prediction of Nanoscale Friction for Two-Dimensional Materials Using a Machine Learning Approach. Tribology Letters. 2020; 68:57. https://doi.org/10.1007/s11249-020-01294-w

44. Gubaev K, Podryabinkin EV, Hart GLW, Shapeev AV. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. Computational Materials Science. 2018; 156:148–156. https://doi.org/10.1016/j.commatsci.2018.09.031

45. Podryabinkin EV, Tikhonov EV, Shapeev AV, Oganov AR. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. Physical Review B. 2019; 99(06):064114. https://doi.org/10.1103/PhysRevB.99.064114

46. Ladygin VV, Korotaev PY, Yanilkin AV, Shapeev AV. Lattice dynamics simulation using machine learning interatomic potentials. Computational Materials Science. 2020; 172:109333. https://doi.org/10.1016/j.commatsci.2019.109333

47. Mortazavi B, Podryabinkin EV, Nvikovb IS, Roche S, Shapeev AV. Efficient machine-learning based interatomic potentials for exploring thermal conductivity in two-dimensional materials. Journal of Physics Materials. 2020; 3(02):02LT02. https://doi.org/10.1088/2515-7639/ab7cbb

48. Korotaev P, Novoselov I, Yanilkin A, Shapeev A. Accessing thermal conductivity of complex compounds by machine learning interatomic potentials. Physical Review B. 2019; 100(14):144308. https://doi.org/10.1103/PhysRevB.100.144308

49. Mortazavi B, Novikov IS, Podryabinkin EV, Roche S, Rabczuk T, Shapeev AV, et al. Exploring phononic properties of two-dimensional materials using machine learning interatomic potentials. Applied Materials Today. 2020; 20:100685. https://doi.org/10.1016/j.apmt.2020.100685

50. Haastrup S, Strange M, Pandey M, Deilmann T, Schmidt PS, Hinsche NF, et al. The Computational 2D Materials Database: High-Throughput Modeling and Discovery of Atomically Thin Crystals. 2D Materials. 2018; 5:042002. https://doi.org/10.1088/2053-1583/aacfc1

51. Hauke J, Kossowski T. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae 2011; 30(02):87–93. https://doi.org/10.2478/v10117-011-0021-1