Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

Explainable artificial intelligence for machine learning prediction of bandgap energies

Cite as: J. Appl. Phys. **136**, 175703 (2024); doi: 10.1063/5.0226151
Submitted: 29 June 2024 · Accepted: 18 October 2024 ·
Published Online: 4 November 2024

View Online Export Citation CrossMark

Taichi Masuda and Katsuaki Tanabe[a)]

AFFILIATIONS

Department of Chemical Engineering, Kyoto University, Nishikyo, Kyoto 615-8510, Japan

[a)]**Author to whom correspondence should be addressed:** tanabe@cheme.kyoto-u.ac.jp

ABSTRACT

The bandgap is an inherent property of semiconductors and insulators, significantly influencing their electrical and optical characteristics. However, theoretical calculations using the density functional theory (DFT) are time-consuming and underestimate bandgaps. Machine learning offers a promising approach for predicting bandgaps with high precision and high throughput, but its models face the difficulty of being hard to interpret. Hence, an application of explainable artificial intelligence techniques to the bandgap prediction models is necessary to enhance the model's explainability. In our study, we analyzed the support vector regression, gradient boosting regression, and random forest regression models for reproducing the experimental and DFT bandgaps using the permutation feature importance (PFI), the partial dependence plot (PDP), the individual conditional expectation plot, and the accumulated local effects plot. Through PFI, we identified that the average number of electrons forming covalent bonds and the average mass density of the elements within compounds are particularly important features for bandgap prediction models. Furthermore, PDP visualized the dependency relationship between the characteristics of the constituent elements of compounds and the bandgap. Particularly, we revealed that there is a dependency where the bandgap decreases as the average mass density of the elements of compounds increases. This result was then theoretically interpreted based on the atomic structure. These findings provide crucial guidance for selecting promising descriptors in developing high-precision and explainable bandgap prediction models. Furthermore, this research demonstrates the utility of explainable artificial intelligence methods in the efficient exploration of potential inorganic semiconductor materials.

I. INTRODUCTION

The bandgap is an inherent property of semiconductors and insulators, significantly influencing their electrical and optical characteristics.[1] Thus, the high-precision and high-throughput calculation of bandgaps of new materials is sought after for the development of materials for optoelectronic devices, such as light-emitting diodes, photovoltaics, and scintillators.[2] Furthermore, significant progress has been made in the calculation of bandgaps.[3] However, the conventional density functional theory (DFT) calculations using the local density approximation[4] often underestimate the bandgap by more than 30% due to the inherent defects of delocalization error and derivative discontinuity.[5] To overcome these challenges and enhance the accuracy of bandgap calculations, computational methods, such as hybrid functionals[6] incorporating partial nonlocal Hartree–Fock (HF) exchange and the *GW*[7,8] approximation, have been utilized. Although the hybrid functionals have seen some improvement in the accuracy, calculating the HF exchange in solids require significantly higher computational resources than the evaluation of local density functionals.[9] Similarly, while the *GW* approximation significantly enhances the accuracy of bandgap calculations, it demands high computational costs and resources.[10] For the development of new semiconductor materials, scientists need to select candidate compounds suitable for experiments. This requires exploring promising materials with specific bandgaps from a vast number of potential compounds. To achieve this, it is necessary to construct and screen a large-scale database of bandgaps, which necessitates calculating bandgaps at low computational cost and with high throughput. Therefore, hybrid functionals incorporating partial nonlocal HF exchange and the *GW* approximation are not practical for high-throughput studies aimed at screening promising candidates with specific bandgaps across vast material spaces, as they require significantly

long computation times.[10–15] Consequently, theoretical bandgap calculations still face significant challenges, necessitating methods that predict the bandgap accurately without high computational costs.

Recently, data-driven machine learning techniques in computational materials science have gained attention. By integrating high-throughput theoretical and experimental values into material databases, it has become possible to predict the characteristics of new materials and discover new materials through machine learning models, noted for their high predictive accuracy and time efficiency.[16–25] Research on predicting bandgaps using machine learning is progressing actively, holding the potential to address the challenges in theoretical bandgap calculations. Historically, machine learning models have used data from theoretical calculations and experimental bandgap values for predictions.[26–31] For instance, Pilania *et al.* demonstrated a machine learning model that combines low-fidelity calculations using Perdew–Burke–Ernzerhof (PBE) and high-fidelity calculations using Heyd–Scuseria–Ernzerhof (HSE06) to enable predictions of bandgaps at high-fidelity levels at a low cost.[13] Lee *et al.* applied the ordinary least squares regression, the lasso regression, and the support vector regression (SVR) to predict bandgap values computed by G_0W_0 for 270 inorganic materials, achieving a root mean square error (RMSE) of 0.24 eV with SVR.[12] For the experimental bandgap values, Zhuo *et al.* trained a machine learning model using 3896 experimental data points with SVR to predict the experimental bandgaps, substantially reducing computational costs while achieving higher accuracy than DFT-calculated values.[32] Kauwe *et al.* enabled more accurate predictions through ensemble learning combining SVR, gradient boosting regression (GBR), and random forest regression (RFR) compared to single machine learning models.[33]

However, machine learning faces the challenge of having opaque internal models and mechanisms, making it difficult to interpret the process and results of predictions. There is a trade-off between the predictive accuracy and interpretability of models. Complex models, known as black-box models, exhibit high predictive accuracy but are hard to interpret. In contrast, simple models, referred to as white-box models, are easier to interpret but often have lower predictive accuracy compared to black-box models.[34–41] Specifically, models, such as tree ensembles, for example, random forests and gradient boosting, support vector machines, and neural networks, are considered complex and difficult to interpret, thus being classified as black-box models.[42–50] To address the issue of black-box models in machine learning, research is being conducted to make the decision-making process of machine learning interpretable and explainable by utilizing the explainable artificial intelligence techniques.[51–58] Explainable artificial intelligence techniques gained significant attention after the Defense Advanced Research Projects Agency (DARPA) in the U.S. launched the explainable artificial intelligence program in 2017. DARPA defines explainable artificial intelligence as AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.[59] In the field of materials science, research incorporating explainable artificial intelligence techniques has been conducted.[60–62] By using explainable artificial intelligence methods, researchers can understand the internal workings of machine learning models, identify the model's issues and limitations, enhance the reliability of the models' predictions, and uncover unexpected correlations that lead to scientific insights.[63,64] For instance, Morita *et al.* demonstrated the high reliability of a machine learning model for predicting dielectric constants by using SHapley Additive exPlanations (SHAP) to analyze the model's results, revealing restored physical correlations.[65] Guo *et al.* discovered that the difference in molecular polarizability is a crucial factor governing diffusion selectivity by combining a light gradient boosting machine model with SHAP to predict gas molecule diffusion selectivity.[66] Rodríguez-Pérez *et al.* analyzed a model for predicting compound activity using Shapley values, identifying structural patterns that determine the probability of the predicted activity.[67] In the context of the bandgap prediction, applying explainable artificial intelligence methods can facilitate a closer integration of data-driven approaches and scientific insights, thereby advancing the resolution of challenges associated with the bandgap prediction. Obada *et al.* utilized SHAP on a model predicting the bandgaps of ABX_3 perovskite materials from a dataset obtained through DFT calculations, using features, such as the electronegativity and the first ionization energy of each element. Their findings revealed that the Pauling electronegativity of the B-site cation was the most influential feature for the bandgap prediction.[68] Hui *et al.* utilized SHAP to reveal feature importance in a GBR tree model predicting the bandgap of perovskite materials, using a dataset calculated by DFT, demonstrating that the average ground state bandgap is a significant feature.[69] Zhang *et al.* developed a Δ-machine learning model linking the bandgaps of HSE and PBE and applied SHAP to show that descriptors related to the atomic radius and the PBE functional bandgap are important features.[70] Rath *et al.* employed the SHAP method on a machine learning model using the Materials Project (MP)[71] database to classify ABX_3 perovskites into direct and indirect bandgaps, showing that transition metal defects and high electronegativity differences are significant features in the direct bandgap classification model.[72] Ogoshi *et al.* constructed a random forest model to predict the directness of bandgaps using the MP dataset and used the multivariate data explanation method to reveal that, in addition to material symmetry, the presence or absence of *d*-bands and the relative energies of atomic orbitals are crucial for classifying direct and indirect bandgaps.[73] Huo *et al.* applied permutation feature importance (PFI) and the sure independence screening and sparsifying operator method to a model predicting bandgaps from the characteristics of the constituent elements of compounds. They demonstrated that electronegativity plays a crucial role in the bandgap prediction model.[74] However, many analyses have employed SHAP, which has a substantial computational cost for calculating feature importance and the relationship between explanatory variables and model predictions.[75] This high cost makes SHAP challenging to apply in cases where there are many features or when the model is complex. Furthermore, the majority of research applying explainable artificial intelligence techniques to bandgap prediction models focuses on models that predict theoretical bandgap values, with fewer studies addressing models that predict experimentally measured bandgap values. Additionally, the model by Huo *et al.*,[74] which predicts experimental bandgap values, includes the bandgap value theoretically

calculated using PBE as a feature. Consequently, this model is limited to compounds for which PBE calculations are feasible. As a result, there has been no research that applies explainable artificial intelligence techniques to models capable of predicting the experimental bandgap values of entirely unknown compounds and analyzing these predictions.

In this study, we analyze models reproducing the experimental and DFT bandgap values using PFI, the partial dependence plot (PDP), the individual conditional expectation (ICE) plot, and the accumulated local effect (ALE) plot, which have lower computational costs compared to SHAP.[76] This research represents the first attempt to visualize the relationship between important features derived by PFI and bandgap predictions using PDP, the ICE plot, and the ALE plot, and to analyze models constructed with the experimental and DFT datasets. Furthermore, our research is the first to analyze a model capable of predicting the experimental bandgaps of entirely unknown compounds using explainable artificial intelligence techniques. First, we tune hyperparameters and compare the prediction accuracy of SVR, GBR, and RFR using elemental characteristics of compounds as features to reproduce the experimental and DFT bandgap values of 1229 inorganic compounds. Next, we use PFI to derive important features for predicting bandgaps in SVR, GBR, and RFR models for both the experimental and DFT data and compare the important features across models and datasets. Furthermore, to verify if PFI indeed identified important features for the model's predictive performance, we compare the prediction accuracy when features were reduced based on the PFI ranking order vs random feature reduction. Subsequently, we visualize the relationship between important features derived from PFI and bandgap predictions using PDP, compare models and datasets, and conduct analyses. Finally, since PDP may not accurately capture the relationship between features and bandgaps due to feature interactions and correlations, we validate the robustness of PDP by applying the ICE plot and the ALE plot. Furthermore, we examine the impact of feature selection based on PFI results on the model's predictive performance and explainability.

II. METHODS

A. Permutation feature importance (PFI)

PFI is a method for visualizing the importance of features in a model. Breiman[77] introduced the prototype method of PFI for decision tree models, and Fisher *et al.*[78] proposed a model-agnostic PFI method.[79] The method for calculating the feature importance of a specific feature F_1 using PFI involves fixing all other features and randomly shuffling the values of F_1 and then assessing the change in the model's predictive accuracy.[80,81] In this study, the feature importance of F_1 was calculated as the difference between the model's prediction accuracy with the original data and the prediction accuracy when the values of F_1 were randomly shuffled ten times.

B. Partial dependence plot (PDP)

PDP is a method for visualizing the marginal effects of features on the predicted values of a model, originally proposed by Friedman.[82] Partial dependence function is generally expressed as shown in Eq. (1),[81,83]

$$\hat{f}_{x_s}(x_s) = E_{x_c}\left[\hat{f}(x_s, x_c)\right] = \int \hat{f}(x_s, x_c) d\mathbb{R}(x_c). \quad (1)$$

In Eq. (1), x_s represents the feature to be plotted, x_c denotes the features other than x_c, and $\hat{f}$ refers to the machine learning model. Here, to estimate the partial dependence function, the Monte Carlo method is utilized, and the partial dependence function can be expressed as shown in Eq. (2),[81,84]

$$\hat{f}_{x_s}(x_s) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}\left(x_s, x_c^{(i)}\right). \quad (2)$$

Here, N refers to the number of instances included in the dataset, and PDP is a line connecting G evenly spaced grid points for x_s. In this study, we use 50 grid points to visualize the average relationship between the feature and the predicted values.

C. Challenges of PDP

PDP captures the average relationship between features and model predictions but fails to capture the relationship between features and model predictions for each instance. Consequently, PDP faces several challenges. For instance, when there are interactions between features, it cannot adequately capture the relationship between features and predictions.[85] Additionally, when there is a strong correlation between features, PDP may need to make predictions for feature values that are far from the distribution of the training data. This can lead to substantial discrepancies due to the impact of extrapolation.

D. Individual conditional expectation (ICE) plot

The ICE plot was proposed by Goldstein *et al.*[85] and visualizes the relationship between features and predicted values for each instance, with their average corresponding to PDP.[81] Therefore, ICE plots enable the confirmation of heterogeneity among instances and allow for the assessment of the interaction effects between features and predicted values, which was highlighted as a challenge of PDP.[85] In this study, we set the number of grids to 50 and visualize the relationship between features and predicted values for each instance.

E. Accumulated local effect (ALE) plot

The ALE plot, proposed by Apley and Zhu,[86] visualizes the relationship between features and predictions by dividing the feature value range into intervals and accumulating the average differences in predictions within these intervals, thereby mitigating the influence of other features. The ALE function is generally expressed as shown in Eq. (3),[81,86]

$$\hat{f}_{ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{\{i: x_{ij} \in N_j(k)\}} \left\{ f(z_{k,j}, \mathrm{x}_{i,\setminus j}) - f(z_{k-1,j}, \mathrm{x}_{i,\setminus j}) \right\}. \quad (3)$$

Here, $N_j(k)$ refers to the kth interval obtained by dividing the range of the feature x_j. $n_j(k)$ represents the number of samples

TABLE I. Variables used in the models. The chemical compositions are vectorized, as based on the properties of the elements and their total, mean and range are used as explanatory variables.

First ionization potential (kJ mol^{-1})	Allred–Rochow electronegativity	Atomic number	Atomic radius (Å)
Atomic weight (u)	Boiling point (K)	Cohesive energy (eV)	Covalent radius (Å)
Critical temperature (K)	Density (g ml^{-1})	Gordy electronegativity	Melting point (K)
Mendeleev number	Molar density (mol ml^{-1})	Molar heat capacity (J mol^{-1} K^{-1})	Nagle electronegativity
Number of unfilled *d* valence electrons	Number of unfilled *f* valence electrons	Number of unfilled *p* valence electrons	Number of unfilled *s* valence electrons
Orbital radius (pm)	Pauling electronegativity	Period	Zunger radius (Å)
Absolute value of valence	Family	Gilmore number of valence electrons	Group
Heat of atomization (kJ mol^{-1})	Heat of fusion (kJ mol^{-1})	Heat of vaporization (kJ mol^{-1})	Ionic radius (Å)
L quantum number	Number of electrons forming covalent bonds	Number of valence electrons	Number of outer shell electrons
Polarizability (Å^3)	Specific heat (J g^{-1} K^{-1})	Thermal conductivity (W m^{-1} K^{-1})	Number of *d* electrons
Number of *f* electrons	Number of *p* electrons	Number of *s* electrons	

contained in the *k*th interval $N_j(k)$. $k_j(x)$ defines the index of the interval to which the value x of x_j belongs to. $z_{k,j}$ is the partition point when the range of x_j is divided into several intervals. In this study, the range of the selected feature was divided into 50 intervals.

F. Computational details

In this study, the dataset comprises 1229 inorganic materials. The experimental values of the bandgap were obtained from the dataset in Ref. 32. These inorganic materials were selected from the dataset in Ref. 32, which are also included in the MP dataset. The theoretical values of the bandgap calculated using DFT were obtained from the MP dataset. The chemical compositions were vectorized, and their total, average, and range, based on the properties of the elements listed in Table I, were used as explanatory variables. These values were derived from Ref. 33. For SVR, the explanatory variables in the dataset were preprocessed by standardizing with scikit-learn's StandardScaler and applying L2 normalization with scikit-learn's Normalizer. The simulations were conducted in a conda environment using Anaconda version 24.1.2, Python version 3.12.2, and scikit-learn version 1.4.2, on macOS Sonoma version 14.4.1 (23E224). The chip used was an Apple M1 Pro chip with 16 GB of memory. The models employed in this study include SVR, GBR, and RFR, utilizing scikit-learn's SVR, GradientBoostingRegressor, and RandomForestRegressor. Models were constructed using both the experimental dataset and the MP dataset, with hyperparameters listed in Table II. These hyperparameters were optimized through grid search within the ranges specified in Table II, using tenfold cross-validation. Cross-validation is a method that extends the well-known holdout validation approach for model performance evaluation by repeatedly splitting the data multiple times. In cross-validation, one fold is used as test data, while the remaining k − 1 folds are used as training data. This splitting process is repeated k times, with each fold being used as test data once. The average of the k results is taken to estimate the true

TABLE II. Optimized parameters (first row) and the search range for grid tuning (second row) of SVR, GBR, and RFR constructed with the experimental values and the MP dataset. *C* in SVR refers to the regularization parameter. The strength of the regularization is inversely proportional to *C*. *gamma* in SVR refers to the kernel coefficient. *n_estimators* in GBR refers to the number of boosting stages to perform. *learning_rate* in GBR refers to the learning rate that shrinks the contribution of each tree. *max_depth* in GBR refers to the maximum depth of the individual regression estimators. *n_estimators* in RFR refers to the number of trees in the forest. *max_features* in RFR refers to the number of features to consider when looking for the best split. *max_depth* in RFR refers to the maximum depth of the tree. The CSV file containing the results of the grid search is available in the supplementary material.

	Hyperparameter	Experimental dataset	MP dataset
SVR	*C*	10 [0.1, 1, 10, 100, 1000]	10 [0.1, 1, 10, 100, 1000]
	gamma	1 [0.001, 0.01, 0.1, 1, 10]	1 [0.001, 0.01, 0.1, 1, 10]
GBR	*n_estimators*	3000 [1000, 3000, 5000, 7000, 9000]	3000 [1000, 3000, 5000, 7000, 9000]
	learning_rate	0.05 [0.01, 0.03, 0.05, 0.07, 0.09]	0.07 [0.01, 0.03, 0.05, 0.07, 0.09]
	max_depth	3 [1, 3, 5, 7, 9]	5 [1, 3, 5, 7, 9]
RFR	*n_estimators*	600 [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]	200 [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
	max_features	"sqrt" ["sqrt," "log2," 1, None]	None ["sqrt," "log2," 1, None]
	max_depth	20 [5, 10, 15, 20, 25, 30]	20 [5, 10, 15, 20, 25, 30]

performance of the model.[87] In other words, using one fold of cross-validation is essentially the equivalent of using a holdout set as a test set.[88] Holdout validation is considered statistically inefficient because it does not use much of the data for training the predictive model. Additionally, if the split between training and test data is inappropriate, holdout validation can lead to misleading performance estimates and significant uncertainty.[87] Furthermore, holdout validation is prone to overfitting, especially when the data are limited and many model variations are tested.[89] In fact, it has been previously reported that that among various model performance evaluation methods, a single holdout validation tends to yield the most biased and least stable estimates of model performance, resulting in non-reproducible outcomes, which is why researchers are advised to avoid it.[87] On the other hand, cross-validation enables unbiased estimation of model performance and tends to provide a more accurate estimate of generalization error, especially when dealing with small datasets. Moreover, since model evaluation is conducted k times, the variance of performance metrics decreases, resulting in more reliable estimates.[90] Therefore, K-fold cross-validation is generally preferred over the holdout method, and it has been shown to improve both bias and variance compared to the estimates obtained through the simple holdout method.[91] In this study, due to the limited amount of material data available, we utilized all the data for cross-validation to make the most effective use of it. The evaluation metric is RMSE. RMSE is the most commonly used evaluation metric for bandgap prediction models.[12,13,29,31–33,92] Similar to RMSE, the widely used mean absolute error (MAE) employs absolute values, making it difficult to compute the gradient or sensitivity of MAE with respect to specific model parameters. On the other hand, RMSE does not use absolute values, and the sum of squared errors is often defined as the cost function to be minimized by adjusting the model parameters. It has been shown in previous studies that adjusting the parameters through this cost function is highly effective in improving model performance.[93] Therefore, RMSE is used in many machine learning studies, including bandgap prediction models, and this study also adopts RMSE as an evaluation metric.

III. RESULTS AND DISCUSSION

A. Comparison between the experimental bandgaps and the DFT bandgaps

Figure 1 shows the relationship between the experimental bandgaps and the MP bandgaps. Table III presents the top ten compounds with the largest bandgap errors, along with their bandgaps and errors. Compounds with wide bandgaps tend to have significant discrepancies between the experimental values and the MP bandgaps, particularly in the case of fluorides. It has been previously reported that DFT calculations for fluorides often exhibit significant errors, with several potential causes identified.[94–96] One explanation is the inaccuracy of the exchange-correlation function in regions with large inhomogeneities in the local density of valence band orbitals.[94] Additionally, the self-energy potential (SEP) used for anions, which is typically derived from neutral atoms, may contribute to these errors. The high electronegativity of fluoride anions results in a weak SEP, which is considered a contributing factor to the observed discrepancies.[96]

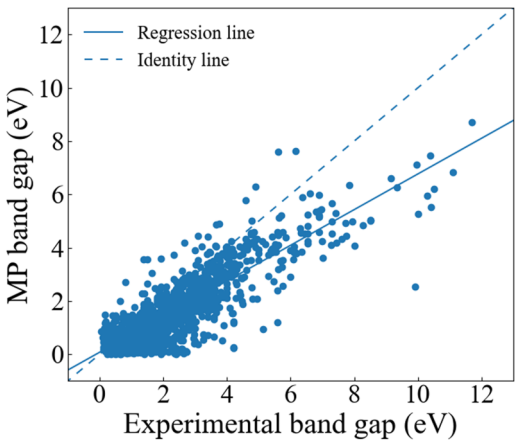

FIG. 1. Relationship between the experimental bandgaps and the MP bandgaps. The dotted line represents the identity line, indicating a perfect match between experimental and MP values. The solid line shows the regression line between these values.

B. Comparison of prediction accuracy between models and datasets

For the three models, SVR, GBR, and RFR, tuning was performed using grid search within the hyperparameter ranges listed in Table II, employing tenfold cross-validation with both the experimental dataset and the MP dataset. Table II presents the optimized parameters resulting from hyperparameter tuning. The predictive accuracy of SVR constructed with the experimental dataset is 0.701 eV in terms of RMSE. The predictive accuracy of SVR constructed with the MP dataset is 0.580 eV in terms of RMSE. The predictive accuracy of GBR constructed with the experimental dataset is 0.672 eV in terms of RMSE. The predictive accuracy of GBR constructed with the MP dataset is 0.577 eV in terms of RMSE. The predictive accuracy of RFR constructed with the experimental dataset is 0.706 eV in terms of RMSE. The predictive accuracy of RFR constructed with the MP dataset is 0.599 eV in terms

TABLE III. Top ten compounds with the largest bandgap errors, along with their bandgaps and errors.

Compound	Experimental bandgap (eV)	Absolute error (eV)
MnF_2	9.9	7.36
RbF	10.4	4.88
CsF	10.0	4.74
GeO_2	5.58	4.36
KF	10.3	4.35
NaF	10.5	4.30
MgF_2	11.1	4.28
AgCl	5.13	4.18
$InAgO_2$	4.2	3.97
Ta_2O_5	4.2	3.94

of RMSE. The predictive accuracy for the MP dataset was the highest with SVR, while GBR achieved the highest predictive accuracy for the experimental dataset. The results of predictive accuracy indicate that, across all models, the predictive accuracy with the MP dataset was higher than that with the experimental dataset. This is attributed to the stronger correlation between the characteristics of the constituent elements of compounds and the bandgaps in the MP compared to the experimental values and the closer approximation of the bandgap data to a normal distribution in the MP dataset. Figure 2 shows the top ten features with the highest correlation to the bandgaps, for both the experimental and MP datasets, along with their absolute correlation coefficients. The results from Fig. 2 indicate that the absolute correlation coefficients of the top ten features with the highest correlation are greater for the MP bandgaps than for the experimental bandgaps. Therefore, the higher predictive accuracy of the MP bandgaps compared to the experimental values is likely due to the stronger correlation between explanatory and response variables, which is advantageous for prediction. Additionally, the skewness and kurtosis of the experimental bandgaps are 1.65 and 4.51, respectively, while those for the MP bandgaps are 1.28 and 2.22, respectively. These values were calculated using SciPy, with Fisher's definition[97] applied for kurtosis. For a normal distribution, skewness and kurtosis are zero. Hence, although both the experimental and MP bandgaps deviate significantly from zero and are not close to a normal distribution, the MP bandgaps have lower skewness and kurtosis, making them closer to a normal distribution compared to the experimental bandgaps. Therefore, the higher predictive accuracy of the MP bandgaps is likely due to their closer approximation to a normal distribution.

C. Analysis of feature importance using PFI

Figure 3 shows PFI for SVR, GBR, and RFR constructed using the experimental and MP dataset. Figure 3 reveals that in all cases,

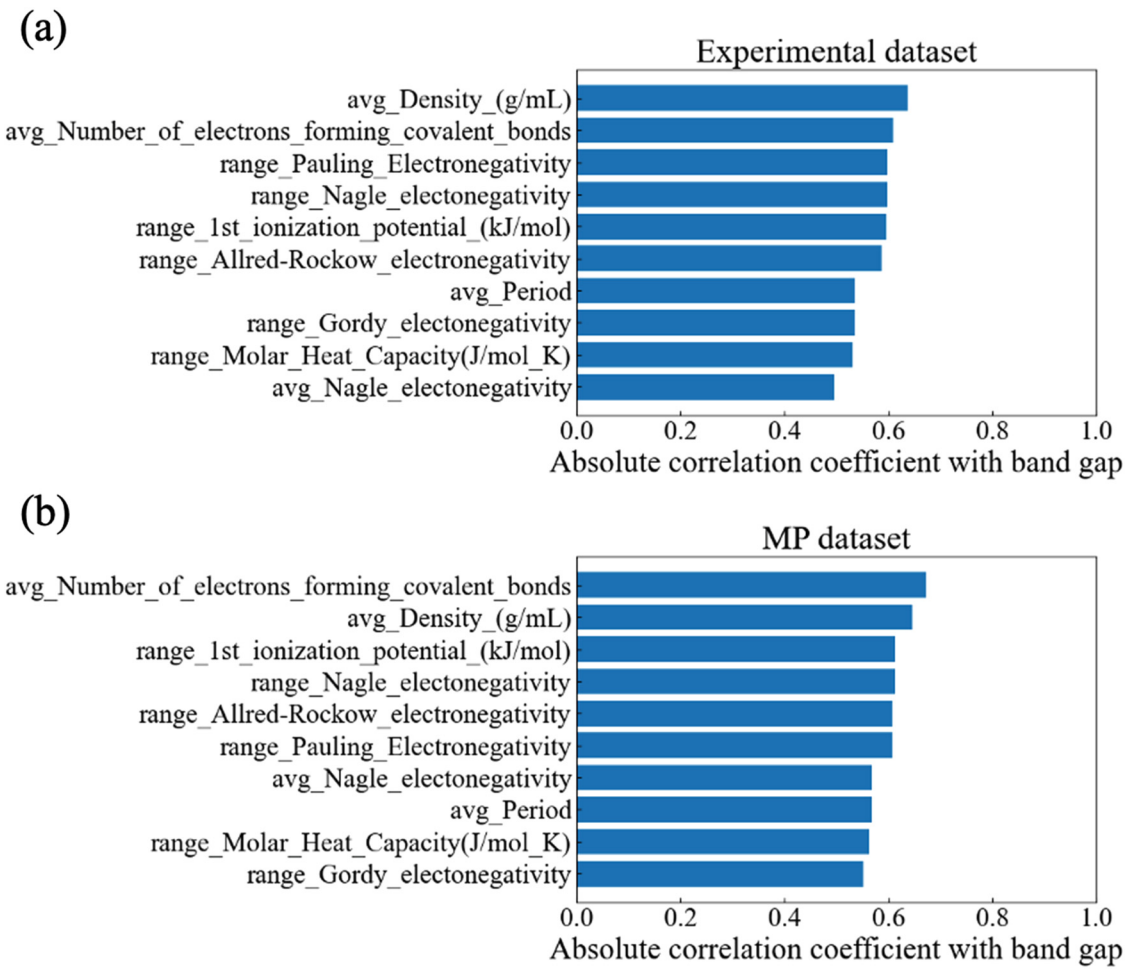

FIG. 2. Top ten features with the highest correlation to the bandgaps along with their absolute correlation coefficients for the (a) experimental and (b) MP dataset.

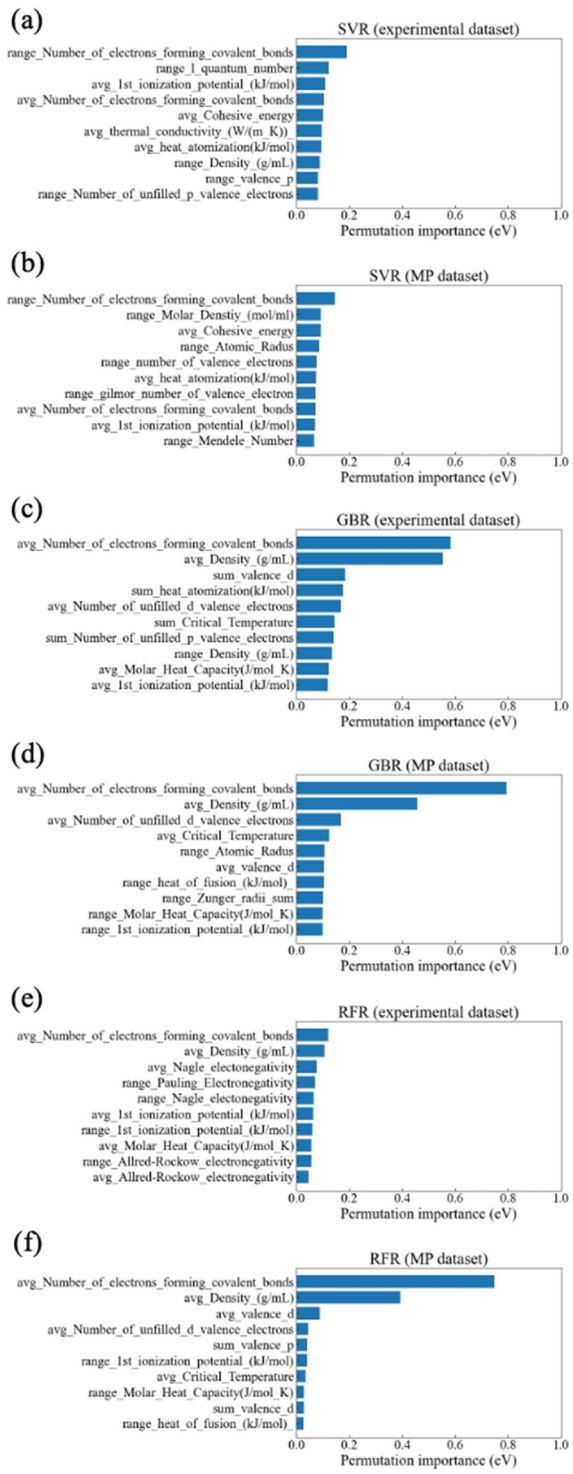

FIG. 3. PFI for SVR constructed with the (a) experimental and (b) MP dataset, GBR constructed with the (c) experimental and (d) MP dataset, and RFR constructed with the (e) experimental and (f) MP dataset.

avg_Number_of_electrons_forming_covalent_bonds holds significant importance in bandgap prediction models. The feature avg_Number_of_electrons_forming_covalent_bonds in Fig. 3 represents the average number of electrons that can form covalent bonds of the elements within compounds.[98] Additionally, Figs. 3(a) and 3(b) indicate that for SVR, range_Number_of_electrons_forming_covalent_bonds is the most important feature for predicting both the experimental and MP bandgaps. The feature range_Number_of_electrons_forming_covalent_bonds in Fig. 3 represents the difference between the maximum and minimum number of electrons that can form covalent bonds of the elements within compounds. The average number of electrons forming covalent bonds and the range of number of electrons forming covalent bonds are related to the number of electrons forming covalent bonds, which is closely associated with electron filling in the bands, thus being derived as important features. Furthermore, it is observed that for SVR predicting the experimental bandgaps, the range of number of electrons forming covalent bonds has a higher importance compared to SVR predicting the MP bandgaps, where the top features have lower importance. In other words, for the MP dataset, shuffling the feature values results in less variation in the predicted values, indicating a lower impact of specific features. This implies that PFI derived from SVR built on the experimental dataset may have identified features more important for the model's predictive performance compared to PFI derived from SVR built on the MP dataset. Figures 3(c)–3(f) also show that for tree ensemble models, such as RFR and GBR, avg_Number_of_electrons_forming_covalent_bonds and avg_Density_(g/ml) are significantly more important features compared to others for both the experimental and MP datasets. The feature avg_Density_(g/ml) represents the average mass density (g/ml) of the elements comprising compounds, which relates to the mass of atomic nuclei per unit volume and is connected to atomic bonding characteristics. Therefore, the high importance of the average mass density feature can be attributed to its relationship with atomic bonding. On the other hand, RFR predicting the experimental bandgaps shows more evenly distributed feature importance compared to other tree ensemble models, which tend to be more heavily biased toward the average number of electrons forming covalent bonds and the average mass density. This distribution in feature importance for RFR is likely due to the hyperparameter *max_features* being set to "sqrt" in the experimental dataset from Table II, meaning that the number of features sampled is the square root of the total number of features. In contrast, other tree ensemble models had *max_features* set to None from Table II, meaning that all features were used in constructing each decision tree. Thus, certain features were repeatedly chosen and used for splitting, increasing their relative importance while decreasing the importance of other features. Additionally, for PFI of RFR built on the experimental dataset, features related to electronegativity were found to be important. This is because electronegativity is closely related to atomic bonding and significantly impacts the band structure. The importance of features related to electronegativity in predicting bandgaps has been reported in the studies by Obada *et al.*,[68] Hui *et al.*,[69] and Huo *et al.*[74] Furthermore, Fig. 3(e) indicates that the features derived as important by PFI of RFR built on the experimental dataset are similar to the features strongly correlated with

bandgaps, suggesting an influence of the correlation between bandgaps and features, indicating that this model is influenced by the correlation between bandgaps and these features. However, when there is a strong correlation between features, it is possible that the model's performance may decrease significantly when exposed to unrealistic data generated by swapping features. Additionally, adding correlated features can lead to a reduction in the importance of relevant features as their importance is divided between the correlated features. Therefore, it is necessary to conduct analysis while considering the correlation between features.[81]

D. Comparison between PFI and Gini importance

The Gini importance is a method for measuring feature importance in RFR and GBR. It is calculated based on the decrease in impurity, aggregating the total decrease in impurity at each split point for each variable. In classification, impurity is usually measured using the Gini index, while in regression, a similar calculation method using the sum of squares is employed.[99] Figures 3(c)–3(f) and 4 show the feature importance for predicting bandgaps using RFR and GBR, as determined by both PFI and Gini importance, for both the experimental and MP datasets. Despite some differences, the overall shape of the graphs and the top-ranking features are similar. However, Gini importance, while computationally fast and simple, is biased toward variables with many split points, such as those with many categories in categorical variables or continuous variables. Additionally, there is a potential issue where the importance of highly correlated features is dispersed, leading to an underestimation of individual importance. On the other hand, permutation feature importance, although more computationally intensive than Gini importance, does not suffer from such biases and can calculate feature importance without relying on the model, making it applicable to any model.[99–101]

E. Influence of the reduction of features

Next, to verify whether PFI actually derived important features for the model's predictive performance, we compared the predictive accuracy when features were reduced according to the PFI ranking vs when features were randomly reduced. These comparisons were conducted across all machine learning models and datasets. Specifically, datasets were created by removing 1, 2, 4, 8, 16, 32, and 64 features in the order of the PFI ranking and at random, respectively. Using these datasets, the predictive accuracy was evaluated by RMSE (in the unit of eV) for each machine learning model and dataset through tenfold cross-validation. Figure 5 shows the relationship between the number of removed features and RMSE when features were reduced according to the PFI ranking and randomly. The results indicate that overall, RMSE tends to increase, meaning predictive accuracy decreases, when features are removed according to the PFI ranking compared to random removal of features. For SVR, RMSE consistently increased as the number of features increased. In contrast, for tree ensemble models, such as RFR and GBR, RMSE sometimes decreased as the number of features increased, particularly when the number of removed features was small. This suggests that even when highly important features were removed according to PFI, predictive accuracy could decrease. This implies that the PFI ranking derived more important features for predictive accuracy for SVR compared to tree ensemble models. Furthermore, models, such as GBR, trained on the experimental dataset and SVR exhibit a higher rate of RMSE increase when features are removed according to the PFI ranking, compared to models with a more extreme bias toward specific features, such as GBR and RFR, trained on the MP dataset. This indicates that models without an extreme bias toward specific features are better at deriving important features for predictive accuracy from the PFI ranking. Additionally, SVR constructed with the experimental datasets showed a greater increase in RMSE when features were reduced according to the PFI ranking compared to SVR constructed with MP datasets. This indicates a decrease in predictive accuracy, suggesting that PFI for SVR built with the experimental datasets derived features more important for its predictive accuracy than PFI for SVR built with MP datasets. This also explains why PFI for SVR built with the experimental datasets had overall higher scores compared to PFI for SVR built with MP datasets, as discussed in Sec. III C.

F. Analysis of the average relationship between the characteristics of the constituent elements of compounds and the bandgaps using PDP

Figure 6 visualizes the relationship between the average number of electrons forming covalent bonds, a top-ranking feature in terms of PFI across all models and datasets, and the bandgap using PDP. In all models and datasets, as the average number of electrons forming covalent bonds increases, the bandgap decreases. Figures 7(a) and 7(b) demonstrate the correlation between the average number of electrons forming covalent bonds and the bandgap, revealing a negative correlation that aligns with the PDP trend. Moreover, while PDP of SVR is a continuous curve, tree ensemble models tend to exhibit a step function. This tendency arises because PDP of RFR can only change between the threshold values that appear in the tree nodes, keeping predictions constant for all other values.[102] Consequently, tree ensemble models are deemed unsuitable for capturing the continuous relationship between explanatory variables and the target variable, whereas SVR is considered more appropriate. Additionally, the importance of the average number of electrons forming covalent bonds derived from PFI of SVR and RFR indicates that the greater the feature importance derived from PFI, the larger the variability in the PDP graph, even within the same model. On the other hand, the feature importance of the average number of electrons forming covalent bonds in PFI of GBR is comparable between the experimental values and the MP values, and the overall shapes of the bandgap graphs from the experimental values and the MP values are similar. However, PDP of GBR built on the MP dataset is shifted in the negative direction on the y axis compared to PDP of GBR built on the experimental dataset. This indicates that while the bandgap variation with respect to the average number of electrons forming covalent bonds is similar for both datasets, the bandgaps in the MP dataset are generally underestimated compared to the experimental values. In other words, regarding the issue of the underestimation of bandgap values by DFT, it was found that the degree of underestimation is not dependent on the range of the average number of electrons

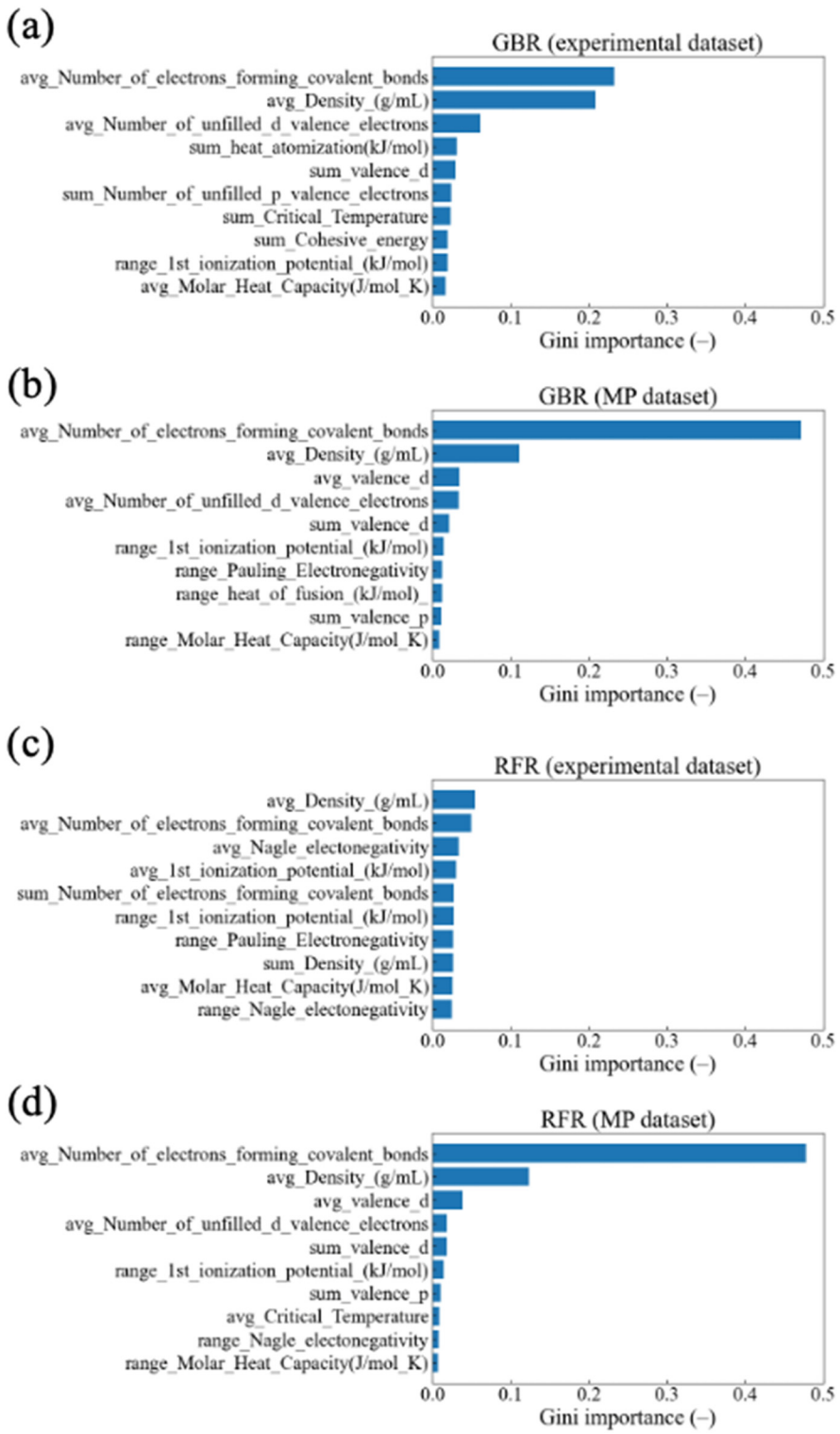

FIG. 4. Gini importance for GBR constructed with the (a) experimental and (b) MP dataset and RFR constructed with the (c) experimental and (d) MP dataset.

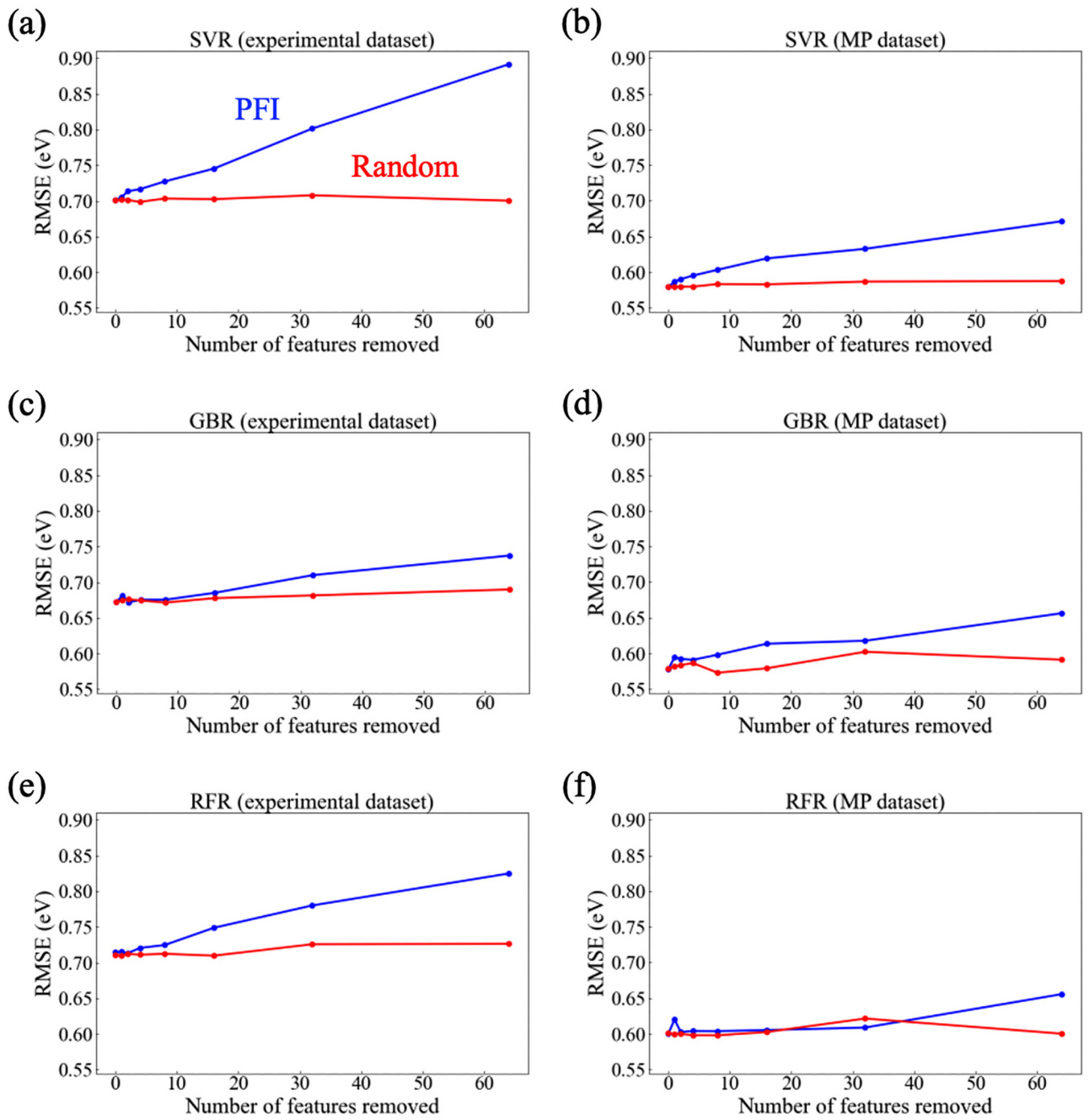

FIG. 5. Relationship between the number of removed features and RMSE (in the unit of eV) when features were reduced according to PFI ranking and randomly for SVR constructed with the (a) experimental and (b) MP dataset, GBR constructed with the (c) experimental and (d) MP dataset, and RFR constructed with the (e) experimental and (f) MP dataset.

forming covalent bonds. Instead, the bandgap is uniformly underestimated regardless of the value of the average number of electrons forming covalent bonds. Next, Fig. 8 visualizes the relationship between the average mass density, which is top-ranked in feature importance in both GBR and RFR, and the bandgap using PDP for GBR and RFR. In GBR and RFR, as the average mass density increases, the bandgap decreases. Here, Fig. 7(e) displays the average mass density and the mass density (g/ml) obtained from

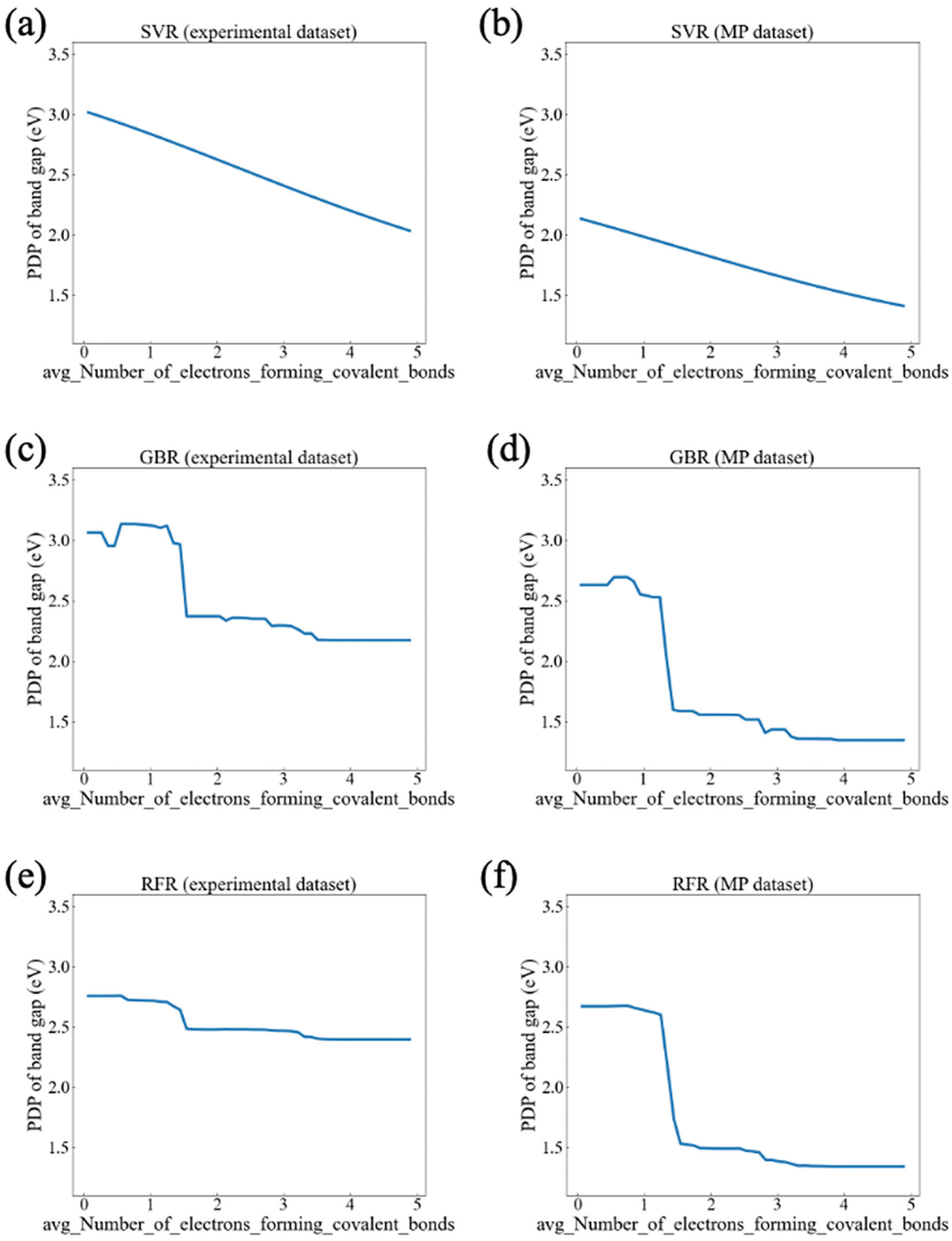

FIG. 6. Relationship between the average number of electrons forming covalent bonds and the bandgap using PDP for SVR constructed with the (a) experimental and (b) MP dataset, GBR constructed with the (c) experimental and (d) MP dataset, and RFR constructed with the (e) experimental and (f) MP dataset.

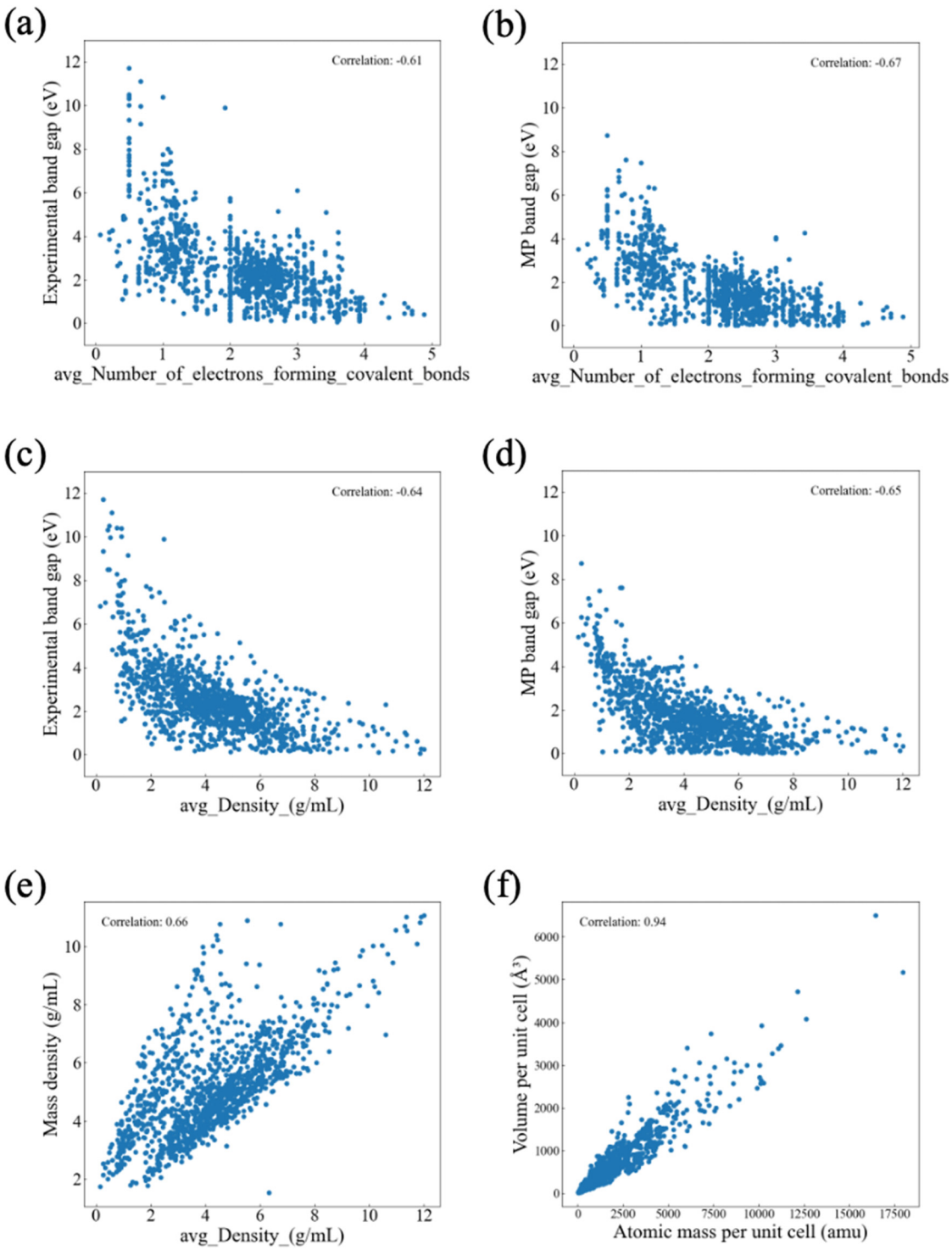

FIG. 7. Relationship between the average number of electrons forming covalent bonds and the (a) experimental and (b) MP bandgap. Relationship between the average mass density and the (c) experimental and (d) MP bandgap. (e) Relationship between the average mass density and the mass density of each compound obtained from the MP. (f) Relationship between the atomic mass per unit cell (amu) and the volume per unit cell of each compound (Å^3).

the MP, indicating a positive correlation between the average mass density and the mass density. As mass density increases, the mass of the atomic nuclei per unit volume increases. Elements with a greater mass of atomic nuclei per unit volume are presumed to have larger atomic distances due to a larger number of electron shells. Indeed, Fig. 7(f) shows a strong positive correlation between the atomic mass per unit cell (amu) and the volume per unit cell of each compound (Å3). Thus, it is inferred that there is a positive correlation between the mass density and atomic distances of each compound. Compounds with larger atomic distances have weaker interatomic bonding forces, resulting in weaker electron binding forces and, consequently, smaller bandgaps. Hence, the positive correlation between the mass density and the average mass density implies that as the average mass density increases, the bandgap decreases. Similar to the average number of electrons forming covalent bonds, PDP graphs for the average mass density show that the

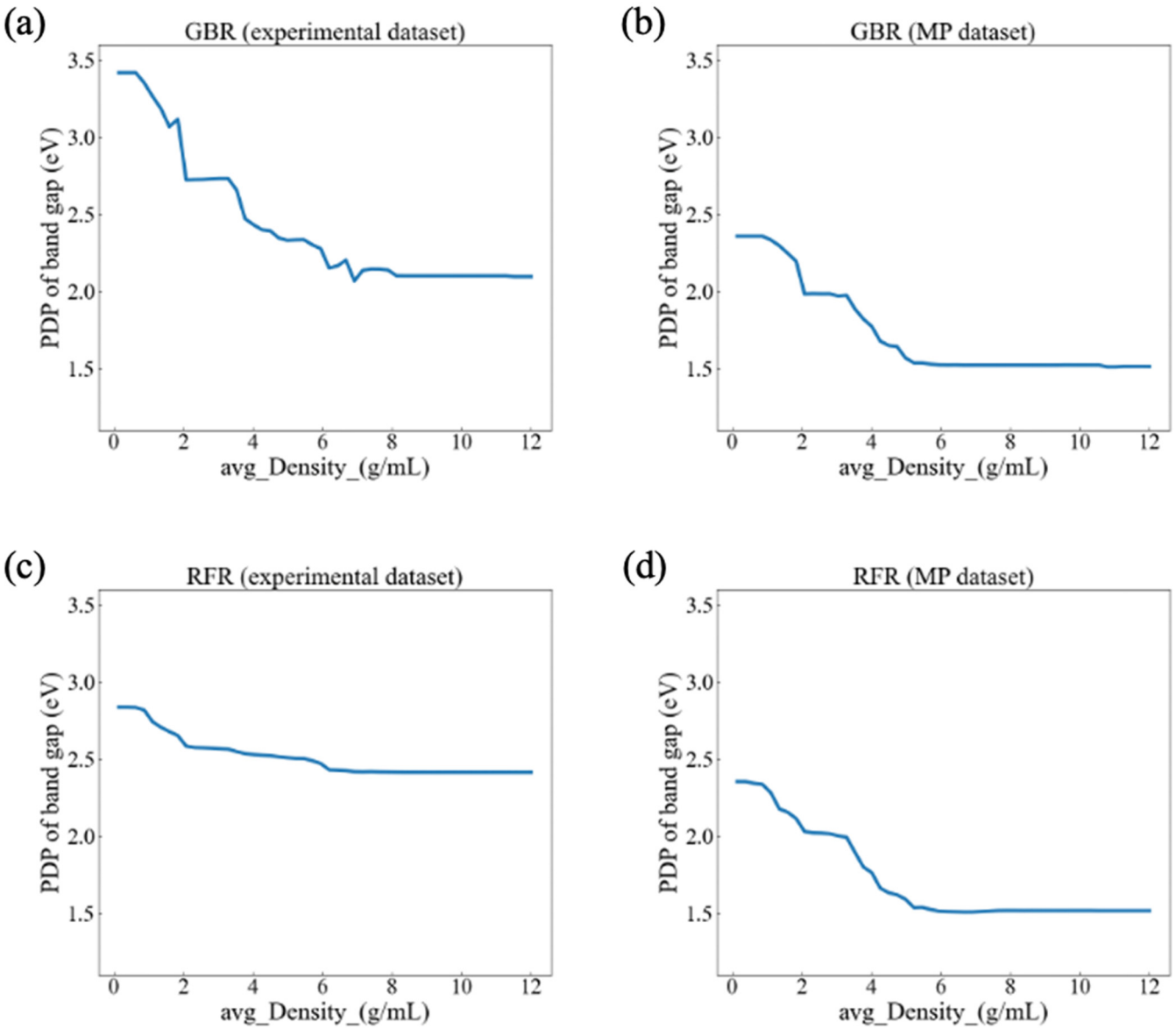

FIG. 8. Relationship between the average mass density and the bandgap using PDP for GBR constructed with the (a) experimental and (b) MP dataset and RFR constructed with the (c) experimental and (d) MP dataset.

greater the feature importance derived from PFI for the same model, the greater the variability in the PDP graph. Indeed, while PFI of GBR and RFR shows differing importances between the experimental and MP datasets, the variability patterns are similar, despite differences in variability rates. Additionally, similar to the average number of electrons forming covalent bonds, PDP graphs for MP are shifted along the y axis compared to the experimental values. This indicates that while the pattern of the bandgap variation with respect to the average mass density is similar, the bandgaps in the MP dataset are underestimated compared to the experimental values, regardless of the average mass density.

G. Validation of PDP using an ICE plot

To examine the impact of feature interactions on PDP, we applied ICE to the relationship between the average number of electrons forming covalent bonds and the bandgap for all models and datasets. Figure 9 shows ICE visualizations of the relationship between the average number of electrons forming covalent bonds and the bandgap for all models and datasets. The results from Fig. 9 indicate that the impact of feature interactions on PDP is minimal, as ICE plots for the average number of electrons forming covalent bonds and the bandgap for each instance are parallel to their average, PDP. Therefore, from the perspective of feature interactions, PDP for the average number of electrons forming covalent bonds accurately captures the relationship between the average number of electrons forming covalent bonds and the bandgap, validating its appropriateness.

H. Validation of PDP using an ALE plot

To examine the impact of feature correlations on PDP, we applied ALE to the relationship between the average number of electrons forming covalent bonds and the bandgap for all models and datasets. Figure 10 visualizes the relationship between the average number of electrons forming covalent bonds and the bandgap using ALE plots. Since ALE accumulates the differences in predicted values for each interval, the bandgap value is zero at the minimum value of the average number of electrons forming covalent bonds. Therefore, by comparing the fluctuations in the graphs of ALE plots and PDP, we can assess the impact of feature correlation on PDP. The results from Fig. 10 indicate that the variability in the ALE graph is similar to that of the PDP graph, suggesting that the impact of feature correlations on PDP is minimal. However, while PDP for SVR shows a near-linear relationship between the average number of electrons forming covalent bonds and the bandgap, ALE plots for SVR capture a non-linear relationship. Therefore, ALE plots can mitigate the influence of feature correlations compared to PDP and capture the relationship between explanatory and response variables in greater detail. Thus, from the perspective of feature correlation, PDP is considered valid in capturing the relationship between features and the bandgap. However, ALE plot is deemed more suitable as it can capture a more detailed relationship between features and the bandgap. Furthermore, with both PDP and ALE plots, there remains a significant challenge in that models with an excessive number of parameters, large-scale datasets, or complex models, such as message-passing neural networks[103] and transformer models,[104] which have been increasingly utilized in the field of materials science, are still likely to require extensive computational time.

I. Influence of the feature selection based on PFI

In recent years, it has been reported that feature selection based on explainable artificial intelligence techniques has improved model prediction accuracy.[105–109] Furthermore, while traditional feature selection methods primarily focused on improving prediction performance, without considering explainability, it has been reported that explainable artificial intelligence enhances the transparency and interpretability of the feature selection process.[110] Therefore, we used the results of PFI to perform feature selection and evaluated its impact on both model prediction performance and explainability. First, we measured the changes in model prediction accuracy as features were removed in the order of increasing importance, starting from the feature with the lowest importance as calculated by PFI. Specifically, we reduced the number of features by up to 50, starting with the least important features, and measured the prediction accuracy of the SVR model trained on the experimental data, using RMSE as the evaluation metric. Figure 11(a) shows the relationship between the number of removed features and RMSE, where features were removed in the order of increasing importance, starting from the feature with the lowest importance as calculated by PFI. It is important to note that, while Sec. III E discusses the reduction of features in the order of decreasing importance, starting from the most important feature according to PFI, in this section, features were reduced in the order of increasing importance, starting from the least important feature. From the results in Fig. 11(a), we can observe that although RMSE increased in some cases when certain features were removed, overall, RMSE decreased. In fact, RMSE before feature reduction was 0.701 eV, whereas after removing 50 features based on the PFI results, RMSE was 0.683 eV. Additionally, by removing irrelevant features and noise data, the predictive accuracy of the model improved, allowing for more reliable explanations. Therefore, this result indicates that performing feature selection based on the PFI results allows for the identification of promising descriptors, enabling the development of a more accurate bandgap prediction model. Next, based on the results of PFI, we measured PFI of the SVR model constructed using experimental data after reducing the number of features by 50. Figure 11(b) shows PFI of the top ten most important features for this model. When compared to the results in Fig. 3(a), some of the top ten important features differ, but overall, they are largely the same. Additionally, after feature reduction, the permutation importance values for all the top ten features exceeded those of the same-ranked features in the model before feature reduction. For instance, the permutation importance of the range_Number_of_electrons_forming_covalent_bonds before feature reduction was 0.191 eV, while after feature reduction, it increased to 0.206 eV. This result indicates that removing unnecessary features from the model allows the essential information for prediction to be emphasized, making it easier to explain which variables are fundamentally important for predictions. These findings are considered to provide important insights into selecting promising descriptors for constructing highly accurate and interpretable bandgap prediction models.

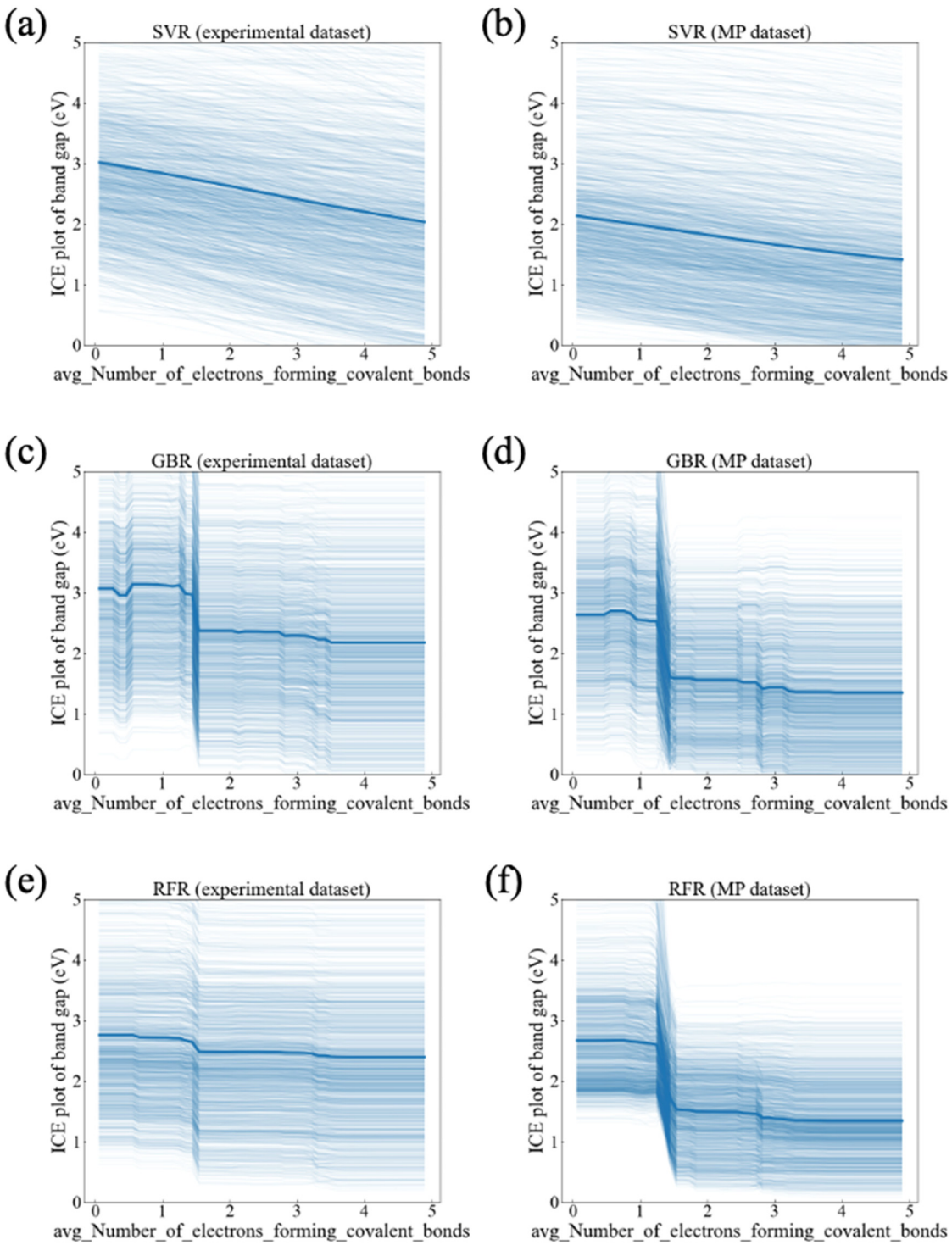

FIG. 9. Relationship between the average number of electrons forming covalent bonds and the bandgap using the ICE plot for SVR constructed with the (a) experimental and (b) MP dataset, GBR constructed with the (c) experimental and (d) MP dataset, and RFR constructed with the (e) experimental and (f) MP dataset. The thin plot represents the ICE plot, while the thick plot indicates PDP.

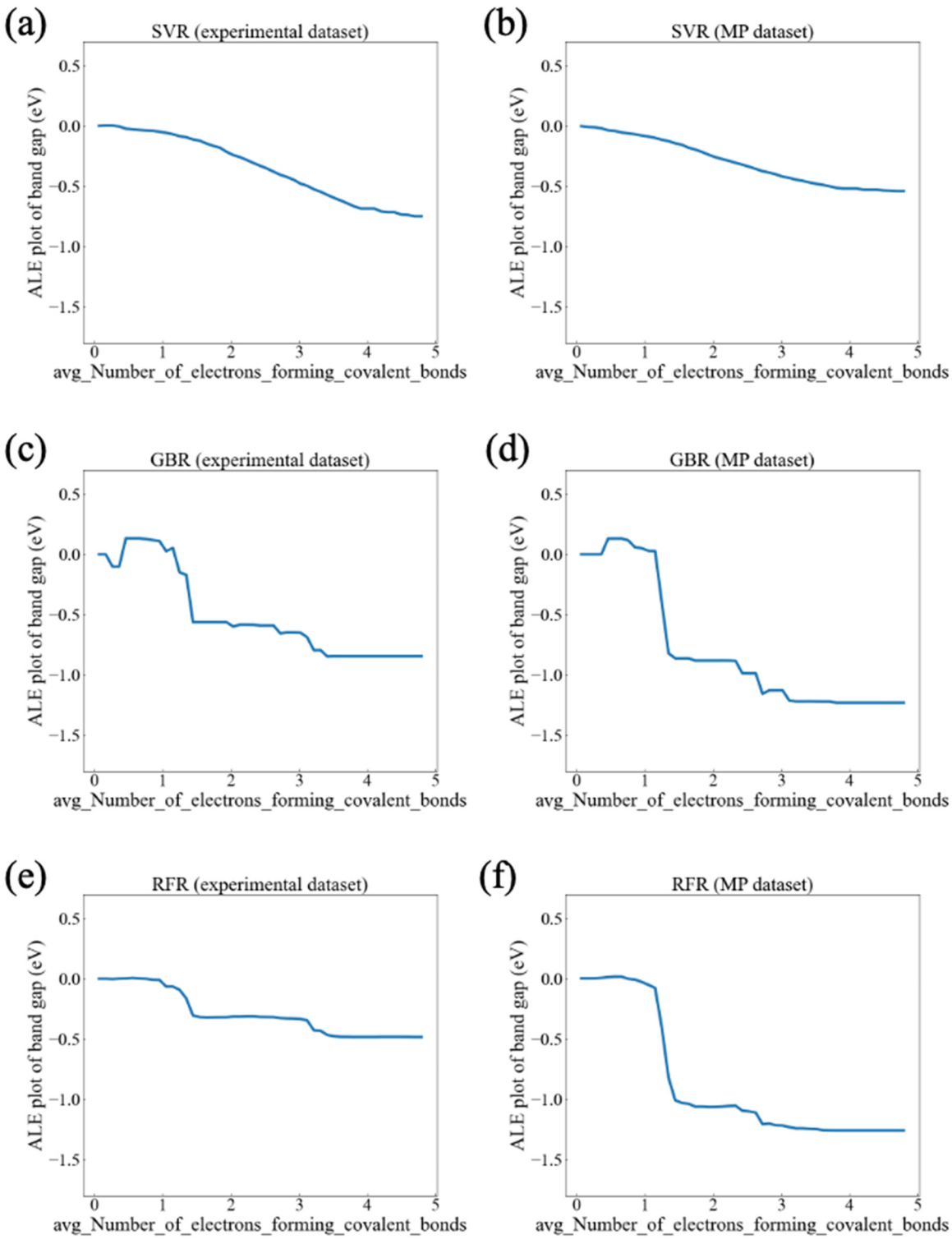

FIG. 10. Relationship between the average number of electrons forming covalent bonds and the bandgap using the ALE plot for SVR constructed using the (a) experimental and (b) MP dataset, GBR constructed using the (c) experimental and (d) MP dataset, and RFR constructed using the (e) experimental and (f) MP dataset.

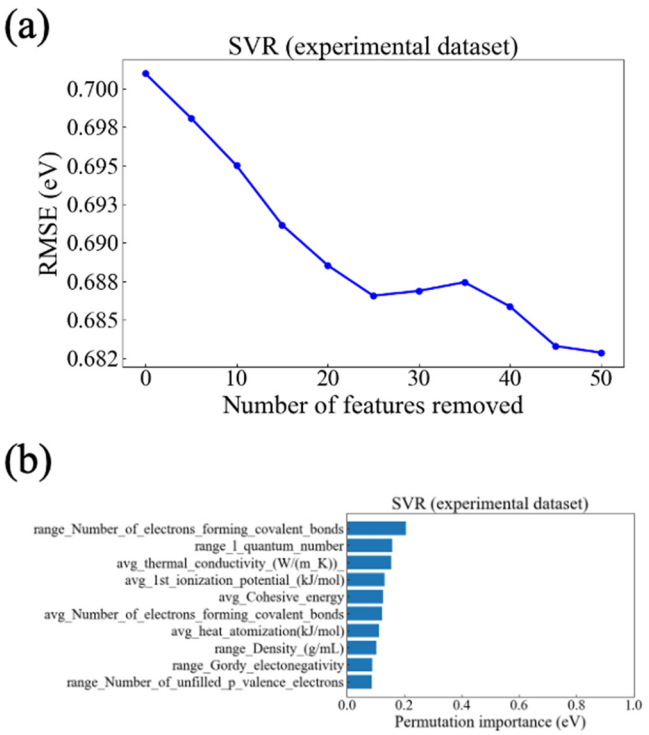

FIG. 11. (a) Relationship between the number of removed features and RMSE (in the unit of eV) for SVR constructed with the experimental dataset when features were reduced in the order of increasing importance, starting from the feature with the lowest importance as calculated by PFI. (b) PFI for SVR constructed with the experimental dataset after reducing 50 features.

IV. CONCLUSIONS

In the present study, we analyzed SVR, GBR, and RFR for reproducing the experimental and DFT bandgaps using PFI, PDP, the ICE plot, and the ALE plot. Through PFI, we identified that the average number of electrons that can form covalent bonds and the average mass density of the elements in compounds are particularly important for bandgap prediction models. Additionally, by comparing the prediction accuracy of models with features reduced based on the PFI ranking vs those reduced randomly, we demonstrated that PFI effectively identified features crucial for the model's predictive performance. Notably, the PFI rankings derived from SVR more effectively identified features crucial for predictive performance compared to those derived from tree ensemble models. Moreover, PDP visualized the dependency relationship between the characteristics of the constituent elements of compounds and the bandgap. Specifically, we discovered that the bandgap decreased as the average mass density of the elements in the compounds increased. This finding could be theoretically explained from the perspective of an atomic electronic structure. Comparing PDP of the mean valence of metals and the bandgap between the experimental and DFT values revealed that the degree of underestimation in DFT bandgap predictions did not vary with the mean valence of metals in the compounds. By applying ICE plots and ALE plots, the validity of PDP was verified in terms of the interactions and correlations between features. We also found that ALE plots can mitigate the impact of feature correlations more effectively than PDP, capturing the relationship between explanatory and response variables in greater detail. However, when applying explainable artificial intelligence methods, there are cases where proper analysis cannot be performed if there is a strong correlation between features. Additionally, complex models still present challenges due to the potentially enormous computational time required. Therefore, there is a need for methods that can calculate feature importance without being affected by feature correlations and approaches that can clarify the dependence between features and predictive outcomes while minimizing computational costs. Furthermore, we found that performing feature selection based on the results of PFI improves the model's predictive performance and enhances its explainability. Our findings in this study provide important guidelines for selecting promising descriptors for the development of highly accurate and explainable bandgap prediction models. Additionally, they offer crucial insights into the utility of explainable artificial intelligence methods for efficiently exploring promising inorganic semiconductor materials and elucidating the challenges in theoretical bandgap calculations using DFT.

SUPPLEMENTARY MATERIAL

See the supplementary material that provides the CSV file containing the results of the grid search.

ACKNOWLEDGMENTS

The authors would like to thank Jakoah Brgoch of the University of Houston for discussions. This study was financially supported, in part, by the Japan Society for the Promotion of Science (JSPS).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Taichi Masuda: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Katsuaki Tanabe:** Conceptualization (supporting); Funding acquisition (lead); Project administration (lead); Supervision (lead); Writing – original draft (supporting); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]Y. Zhang, T. Tang, C. Girit, Z. Hao, M. C. Martin, A. Zettl, M. F. Crommie, Y. Shen, and F. Wang, "Direct observation of a widely tunable bandgap in bilayer graphene," Nature **459**, 820–823 (2009).

[2]A. Talapatra, B. P. Uberuaga, C. R. Stanek, and G. Pilania, "Band gap predictions of double perovskite oxides using machine learning," Commun. Mater. **4**, 46 (2023).

[3]V. Sokolovskiy, D. Baigutlin, O. Miroshkina, and V. Buchelnikov, "Meta-GGA SCAN functional in the prediction of ground state properties of magnetic materials: Review of the current state," Metals **13**, 728 (2023).

[4]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133 (1965).

[5]Z. Wan, Q. Wang, D. Liu, and J. Liang, "Effectively improving the accuracy of PBE functional in calculating the solid band gap via machine learning," Comput. Mater. Sci. **198**, 110699 (2021).

[6]J. Heyd, G. E. Scuseria, and M. Ernzerhof, "Hybrid functionals based on a screened Coulomb potential," J. Chem. Phys. **118**, 8207–8215 (2003).

[7]F. Aryasetiawan and O. Gunnarsson, "The GW method," Rep. Prog. Phys. **61**, 237 (1998).

[8]E. L. Shirley, "Self-consistent GW and higher-order calculations of electron states in metals," Phys. Rev. B **54**, 7758 (1996).

[9]A. J. Garza and G. E. Scuseria, "Predicting band gaps with hybrid density functionals," J. Phys. Chem. Lett. **7**, 4165–4170 (2016).

[10]V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, "Machine learning for predicting the band baps of ABX_3 perovskites from elemental properties," J. Phys. Chem. C **124**, 8905–8918 (2020).

[11]G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman, "Machine learning bandgaps of double perovskites," Sci. Rep. **6**, 19375 (2016).

[12]J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, "Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques," Phys. Rev. B **93**, 115104 (2016).

[13]G. Pilania, J. E. Gubernatis, and T. Lookman, "Multi-fidelity machine learning models for accurate bandgap predictions of solids," Comput. Mater. Sci. **129**, 156–163 (2017).

[14]L. Weston and C. Stampfl, "Machine learning the band gap properties of kesterite I_2–II–IV–V_4 quaternary compounds for photovoltaics applications," Phys. Rev. Mater. **2**, 085407 (2018).

[15]H. Wang, A. Tal, T. Bischoff, P. Gono, and A. Pasquarello, "Accurate and efficient band-gap predictions for metal halide perovskites at finite temperature," npj Comput. Mater. **8**, 237 (2022).

[16]A. Agrawal and A. Choudhary, "Perspective materials informatics and big data realization of the 'fourth paradigm' of science in materials science," APL Mater. **4**, 053208 (2016).

[17]N. Nosengo, "Can artificial intelligence create the next wonder material?," Nature **533**, 22–25 (2016).

[18]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**, 16028 (2016).

[19]Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," J. Mater. **3**, 159–177 (2017).

[20]R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: Recent applications and prospect," npj Comput. Mater. **3**, 54 (2017).

[21]J. Correa-Baena, K. Hippalgaonkar, J. van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, and T. Buonassisi, "Accelerating materials development via automation, machine learning, and high-performance computing," Joule **2**, 1410–1420 (2018).

[22]K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[23]L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, "Data-driven materials science: Status, challenges, and perspectives," Adv. Sci. **6**, 1900808 (2019).

[24]S. G. Louie, Y. Chan, F. H. Jornada, Z. Li, and D. Y. Qiu, "Discovering and understanding materials through computation," Nat. Mater. **20**, 728–735 (2021).

[25]Z. Wang, A. Chen, K. Tao, J. Cai, Y. Han, J. Gao, S. Ye, S. Wang, I. Ali, and J. Li, "AlphaMat: A material informatics hub connecting data, features, models and applications," npj Comput. Mater. **9**, 130 (2023).

[26]T. Gu, W. Lu, X. Bao, and N. Chen, "Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors," Solid State Sci. **8**, 129–136 (2006).

[27]P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, "Informatics-aided bandgap engineering for solar materials," Comput. Mater. Sci. **83**, 185–195 (2014).

[28]J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Comput. Mater. **5**, 83 (2019).

[29]G. S. Na, S. Jang, Y. L. Lee, and H. Chang, "Tuplewise material representation based machine learning for accurate band gap prediction," J. Phys. Chem. A **124**, 10616–10623 (2020).

[30]Y. Tang, H. Chen, J. Wang, and X. Niu, "Machine learning-aided band gap prediction of semiconductors with low concentration doping," Phys. Chem. Chem. Phys. **25**, 18086–18094 (2023).

[31]S. Ghosh and J. Chowdhury, "Predicting band gaps of ABN_3 perovskites: An account from machine learning and first-principle DFT studies," RSC Adv. **14**, 6385–6397 (2024).

[32]Y. Zhuo, A. M. Tehrani, and J. Brgoch, "Predicting the band gaps of inorganic solids by machine learning," J. Phys. Chem. Lett. **9**, 1668–1673 (2018).

[33]S. K. Kauwe, T. Welker, and T. D. Sparks, "Extracting knowledge from DFT: Experimental band gap predictions through ensemble learning," Integr. Mater. Manuf. Innov. **9**, 213–220 (2020).

[34]T. Mori and N. Uchihira, "Balancing the trade-off between accuracy and interpretability in software defect prediction," Empir. Softw. Eng. **24**, 779–825 (2019).

[35]W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," Proc. Natl. Acad. Sci. U.S.A. **116**, 22071–22080 (2019).

[36]O. Loyola-González, "Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view," IEEE Access **7**, 154096–154113 (2019).

[37]G. Baryannis, S. Dani, and G. Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability," Future Gener. Comput. Syst. **101**, 993–1004 (2019).

[38]A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," Neural Comput. Appl. **32**, 18069–18083 (2020).

[39]B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," Hum. Brain Mapp. **41**, 1435–1444 (2020).

[40]G. Pilania, "Machine learning in materials science: From explainable predictions to autonomous design," Comput. Mater. Sci. **193**, 110360 (2021).

[41]F. Funer, "Accuracy and interpretability: Struggling with the epistemic foundations of machine learning-generated medical information and their practical implications for the doctor-patient relationship," Philos. Technol. **35**, 5 (2022).

[42]N. Barakat, A. P. Bradley, and M. N. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," IEEE Trans. Inf. Technol. Biomed. **14**, 1114–1120 (2010).

[43]P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," Inf. Sci. **225**, 1–17 (2013).

[44]J. Balfer and J. Bajorath, "Visualization and interpretation of support vector machine activity predictions," J. Chem. Inf. Model. **55**, 1136–1147 (2015).

[45]Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," Transp. Res. Part C Emerg. **58**, 308–324 (2015).

[46]J. Hur, S. Ihm, and Y. Park, "A variable impacts measurement in random forest for mobile cloud computing," Wirel. Commun. Mob. Comput. **2017**, 6817627.

[47]R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput. Surv. **51**, 1–42 (2019).
[48]A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," Knowl. Inf. Syst. **61**, 1269–1302 (2019).
[49]J. Petch, S. Di, and W. Nelson, "Opening the black box: The promise and limitations of explainable machine learning in cardiology," Can. J. Cardiol. **38**, 204–213 (2022).
[50]X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y. Han, "Explainable machine learning in materials science," npj Comput. Mater. **8**, 204 (2022).
[51]A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access **6**, 52138–52160 (2018).
[52]A. Holzinger, "From machine learning to explainable AI," in *2018 World Symposium on Digital Intelligence for Systems and Machines* (IEEE, 2018), pp. 55–66.
[53]D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable AI in intrusion detection systems," in *IECON 2018—44th Annual Conference of the IEEE Industrial Electronics Society* (IEEE, 2018), pp. 3237–3243.
[54]A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion **58**, 82–115 (2020).
[55]X. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware eXplainable AI," IEEE Trans. Knowl. Data Eng. **34**, 29–49 (2022).
[56]I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where," IEEE Trans. Industr. Inform. **18**, 5031–5042 (2022).
[57]A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. S. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," IEEE Trans. Artif. Intell. **3**, 852–866 (2022).
[58]W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," Knowl.-Based Syst. **263**, 110273 (2023).
[59]D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," AI Mag. **40**, 44–58 (2019).
[60]J. Feng, J. L. Lansford, M. A. Katsoulakis, and D. G. Vlachos, "Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences," Sci. Adv. **6**, eabc3204 (2020).
[61]R. Dybowski, "Interpretable machine learning as a tool for scientific discovery in chemistry," New J. Chem. **44**, 20914–20920 (2020).
[62]V. Korolev and P. Protsenko, "Accurate, interpretable predictions of materials properties within transformer language models," Patterns **4**, 100803 (2023).
[63]F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry," Acc. Mater. Res. **3**, 597–607 (2022).
[64]T. Liu and A. S. Barnard, "The emergent role of explainable artificial intelligence in the materials sciences," Cell Rep. Phys. Sci. **4**, 101630 (2023).
[65]K. Morita, D. W. Davies, K. T. Butler, and A. Walsh, "Modeling the dielectric constants of crystals using machine learning," J. Chem. Phys. **153**, 024503 (2020).
[66]S. Guo, X. Huang, Y. Situ, Q. Huang, K. Guan, J. Huang, W. Wang, X. Bai, Z. Liu, Y. Wu, and Z. Qiao, "Interpretable machine-learning and big data mining to predict gas diffusivity in metal-organic frameworks," Adv. Sci. **10**, 2301461 (2023).
[67]R. Rodríguez-Pérez and J. Bajorath, "Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values," J. Med. Chem. **63**, 8761–8777 (2020).
[68]D. O. Obada, E. Okafor, S. A. Abolade, A. M. Ukpong, D. Dodoo-Arhin, and A. Akande, "Explainable machine learning for predicting the band gaps of ABX_3 perovskites," Mater. Sci. Semicond. Process. **161**, 107427 (2023).
[69]Z. Hui, M. Wang, J. Wang, J. Chen, X. Yin, and Y. Yue, "Predicting the properties of perovskite materials by improved compositionally restricted attention-based networks and explainable machine learning," J. Phys. D: Appl. Phys. **57**, 315303 (2024).
[70]L. Zhang, T. Su, M. Li, F. Jia, S. Hu, P. Zhang, and W. Ren, "Accurate band gap prediction based on an interpretable Δ-machine learning," Mater. Today Commun. **33**, 104630 (2022).
[71]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary The materials project A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).
[72]S. Rath, G. S. Priyanga, N. Nagappan, and T. Thomas, "Discovery of direct band gap perovskites for light harvesting by using machine learning," Comput. Mater. Sci. **210**, 111476 (2022).
[73]E. Ogoshi, M. Popolin-Neto, C. M. Acosta, G. M. Nascimento, J. N. Rodrigues, O. N. Oliveira Jr, F. V. Paulovich, and G. M. Dalpian, "Learning from machine learning: The case of band-gap directness in semiconductors," Discov. Mater. **4**, 6 (2024).
[74]S. Huo, S. Zhang, Q. Wu, and X. Zhang, "Feature-assisted machine learning for predicting band gaps of binary semiconductors," Nanomaterials **14**, 445 (2024).
[75]I. Covert, S. Lundberg, and S. Lee, "Understanding global feature contributions with additive importance measures," arXiv:2004.00668 (2020).
[76]S. Hu, C. Xiong, P. Chen, and P. Schonfeld, "Examining nonlinearity in population inflow estimation using big data: An empirical comparison of explainable machine learning models," Transp. Res. A: Policy Pract. **174**, 103743 (2023).
[77]L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[78]A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," Mach. Learn. Res. **20**, 1–18 (2019).
[79]F. Fumagalli, M. Muschalik, E. Hüllermeier, and B. Hammer, "Incremental permutation feature importance (iPFI): Towards online explanations on data streams," Mach. Learn. **112**, 4863–4903 (2023).
[80]S. Oh, "Predictive case-based feature importance and interaction," Inf. Sci. **593**, 155–176 (2022).
[81]C. Molnar, *Interpretable Machine Learning* (2024); see https://christophm.github.io/interpretable-ml-book/.
[82]J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat. **29**, 1189–1232 (2001).
[83]H. Sarvaiya, A. Loya, C. Warke, S. Deshmukh, S. Jagnade, A. Toshniwal, and F. Kazi, "Explainable artificial intelligence (XAI): Towards malicious SCADA communications," in *ISUW 2020*, LNEE (Springer, 2022), Vol. 847, pp. 151–162.
[84]C. Molnar, T. Freiesleben, G. König, J. Herbinger, T. Reisinger, G. Casalicchio, M. N. Wright, and B. Bischl, "Relating the partial dependence plot and permutation feature importance to the data generating process," in *Explainable Artificial Intelligence. xAI 2023, CCIS* (Springer, 2023), Vol. 1901, pp. 456–479.
[85]A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," J. Comput. Graph. Stat. **24**, 44–65 (2015).
[86]D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," J. R. Stat. Soc. Ser. B Methodol. **82**, 1059–1086 (2020).
[87]C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "An empirical comparison of model validation techniques for defect prediction models," IEEE Trans. Softw. Eng. **43**, 1–18 (2017).
[88]J. Eertink, M. Heymans, G. Zwezerijnen, J. Zijlstra, H. de Vet, and R. Boellaard, "External validation: A simulation study to compare cross-validation versus holdout or external testing to assess the performance of clinical prediction models using PET data from DLBCL patients," EJNMMI Res. **12**, 58 (2022).
[89]Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," J. Mach. Learn. Res. **5**, 1089–1105 (2004).

[90]F. Maleki, N. Muthukrishnan, K. Ovens, C. Reinhold, MD, and R. Forghani, "Machine learning algorithm validation," Neuroimaging Clin. N. Am. **30**, 433–445 (2020).
[91]C. Chavez-Chong, "Cross-validation for spatial data," (2024), hal-04605503; see https://hal.science/hal-04605503v1.
[92]A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K. R. Lee, and A. K. Singh, "Machine-learning-assisted accurate band gap predictions of functionalized MXene," Chem. Mater. **30**, 4031–4038 (2018).
[93]T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature," Geosci. Model Dev. **7**, 1247–1250 (2014).
[94]S. C. Erwin and C. C. Lin, "The self-interaction-corrected electronic band structure of six alkali fluoride and chloride crystals," J. Phys. C: Solid State Phys. **21**, 4285 (1988).
[95]F. Matusalem, M. Marques, L. K. Teles, A. Filippetti, and G. Cappellini, "Electronic properties of fluorides by efficient approximated quasiparticle DFT-1/2 and PSIC methods: BaF_2, CaF_2 and CdF_2 as test cases," J. Phys.: Condens. Matter. **30**, 365501 (2018).
[96]Z. Ai, S. Yang, K. Xue, W. Yang, J. Huang, and X. Miao, "DFT-1/2 for ionic insulators: Impact of self-energy potential on band gap correction," Comput. Mater. Sci. **239**, 112978 (2024).
[97]R. A. Fisher, "Moments and product moments of sampling distributions," Proc. Lond. Math. Soc. **s2–30**, 199–238 (1930).
[98]L. Pauling and B. Kamb, "A revised set of values of single-bond radii derived from the observed interatomic distances in metals by correction for bond number and resonance energy," Proc. Natl. Acad. Sci. U.S.A. **83**, 3569–3571 (1986).
[99]S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?," Bioinformatics **34**, 3711–3718 (2018).
[100]A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," Bioinformatics **26**, 1340–1347 (2010).
[101]J. Gómez-Ramírez, M. Ávila-Villanueva, and MÁ Fernández-Blázquez, "Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods," Sci. Rep. **10**, 20630 (2020).
[102]J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (Association for Computing Machinery, 2016), pp. 5686–5697.
[103]R. E. A. Goodall and A. A. Lee, "Predicting materials properties without crystal structure: Deep representation learning from stoichiometry," Nat. Commun. **11**, 6280 (2020).
[104]A. Y. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, "Compositionally restricted attention-based network for materials property predictions," npj Comput. Mater. **7**, 77 (2021).
[105]G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, "Evaluating explainable artificial intelligence methods based on feature elimination: A functionality-grounded approach," Electronics **12**, 1670 (2023).
[106]P. A. Kowalski and M. Walczak, "Feature selection for regression tasks base on explainable artificial intelligence procedures," in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2023), pp. 1–8.
[107]Z. Yang, Z. Wang, C. Huang, and X. Yao, "An explainable feature selection approach for fair machine learning," in *Artificial Neural Networks and Machine Learning—ICANN 2023* (Springer, 2023), Vol. 14261, pp. 75–86.
[108]X. Chen, M. Liu, Z. Wang, and Y. Wang, "Explainable deep learning-based feature selection and intrusion detection method on the internet of things," Sensors **24**, 5223 (2024).
[109]A. Hinterleitner, T. Bartz-Beielstein, R. Schulz, S. Spengler, T. Winter, and C. Leitenmeier, "Enhancing feature selection and interpretability in AI regression tasks through feature attribution," arXiv:2409.16787 (2024).
[110]J. Zacharias, M. von Zahn, J. Chen, and O. Hinz, "Designing a feature selection method based on explainable artificial intelligence," Electron. Mark. **32**, 2159–2184 (2022).