JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

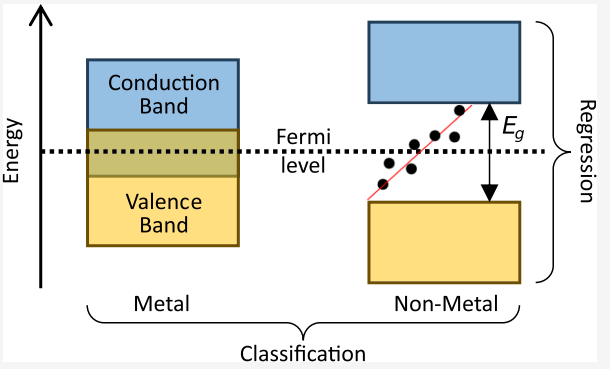
pubs.acs.org/jcim Article

Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow

Son Gyo Jung, Guwon Jung, and Jacqueline M. Cole*

Cite This: J. Chem. Inf. Model. 2024, 64, 1187−1200 Read Online

ABSTRACT: Machine learning (ML) methods can train a model to predict material properties by exploiting patterns in materials databases that arise from structure−property relationships. However, the importance of ML-based feature analysis and selection is often neglected when creating such models. Such analysis and selection are especially important when dealing with multifidelity data because they afford a complex feature space. This work shows how a gradient-boosted statistical feature-selection workflow can be used to train predictive models that classify materials by their metallicity and predict their band gap against experimental measurements, as well as computational data that are derived from electronic-structure calculations. These models are fine-tuned via Bayesian optimization, using solely the features that are derived from chemical compositions of the materials data. We test these models against experimental, computational, and a combination of experimental and computational data. We find that the multifidelity modeling option can reduce the number of features required to train a model. The performance of our workflow is benchmarked against state-of-the-art algorithms, the results of which demonstrate that our approach is either comparable to or superior to them. The classification model realized an accuracy score of 0.943, a macro-averaged F1-score of 0.940, area under the curve of the receiver operating characteristic curve of 0.985, and an average precision of 0.977, while the regression model achieved a mean absolute error of 0.246, a root-mean squared error of 0.402, and R^2 of 0.937. This illustrates the efficacy of our modeling approach and highlights the importance of thorough feature analysis and judicious selection over a "black-box" approach to feature engineering in ML-based modeling.



1. INTRODUCTION

The analysis of band gaps (E_g) of functional inorganic materials is pivotal to the design of many applications, including light-emitting diodes, photovoltaic cells, and transistors.[1−8] There are well-established *ab initio* approaches that are used to predict E_g. Theoretical methods, such as high-throughput electronic-structure calculations based on density functional theory (DFT), have played a vital role in accelerating the discovery of novel chemical materials in these fields of research.[9−17] The process characterizing materials and their band gap properties has been streamlined via *ab initio* methods that facilitate computational simulations of material properties. This advancement has accelerated the exploration of diverse chemical landscapes across multiple research fields, a pace unattainable through conventional design-to-device processes.

While DFT calculations offer significant capabilities, they are often inaccurate or are not general enough, owing to inherent errors; these stem from their approximate nature and their requirement of additional chemical information that is neither standardized nor is readily available.[18−21] A notable systematic discrepancy is observed in DFT-based calculations of E_g, in that they frequently underestimate E_g relative to their cognate experimental values. These errors are attributed to approximations in the exchange−correlation functional and a derivative discontinuity term in the true density functional. Improved calculations for E_g can be afforded using hybrid functionals and GW-type methods.[22−24] Yet, their high computational cost makes them unsuitable for rapid chemical property predictions. Additional functionals exist that can afford the accurate prediction of E_g without such an increase in the computational requirement. These include the Becke-Johnson (mBJ) potential and the generalization of Δ-self-consistent field (Δ-SCF) to solids.[25−27] However, there are limitations associated with them. For instance, the mBJ functional is highly effective for many semiconductors and insulators, but it struggles with ferromagnetic metals, while the Δ-SCF method relies on the dielectric screening properties of

Received: November 27, 2023
Revised: January 26, 2024
Accepted: January 26, 2024
Published: February 6, 2024

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

electrons.[27−29] It is also important to highlight that DFT calculations are mostly restricted to ordered crystal structures, and their accuracy falters for highly correlated systems. However, the integration of DFT + U can ameliorate such limitations.[30]

These efforts within computational materials science have led to the creation of data repositories with extensive sets of computed chemical structures and their properties, such as the Materials Project (MP).[31,32] The accessibility of these chemical data, coupled with the rise of big-data initiatives, has resulted in a growing interest in data-driven methods owing to their proficiency in processing and analyzing large-scale, high-dimensional data sets.

In materials science, data-driven approaches leverage materials informatics and machine learning (ML),[33−37] which may include electronic structure calculations. A typical materials informatics workflow involves transforming the *ab initio* chemical information into a machine-readable format using feature descriptors.[38−43] The generated features are then used for model training, which facilitates the statistical prediction of: (i) properties of unseen chemical materials in a regression problem or (ii) the specific class or category the materials are associated with in a classification problem. The rationale is to empower ML models to deduce chemical structure−property relationships that exceed the capabilities of manual analysis. These techniques have showcased their prowess in accurately predicting chemical structures and properties, including the use of a multifidelity modeling strategy that harnesses both DFT calculations and experimental measurements.[44−46] This exemplifies the effectiveness of materials screening for the realization of novel materials within highly complex feature spaces for various applications.

Various ML techniques have been employed in the prediction of E_g against the DFT calculations. For instance, Gladkikh et al.[47] demonstrated the use of kernel ridge regression (KRR), extremely randomized trees, and alternating conditional expectations to predict E_g of ABX3 perovskites from elemental properties. Pilania et al.[48] leveraged a KRR model to estimate E_g for double perovskites. Pilania et al.[29] also explored a multifidelity ML modeling approach, where a multifidelity cokriging statistical learning framework is used to amalgamate variable-fidelity quantum mechanical calculations, to generate an ML model based on a Gaussian process regression. A support vector regression (SVR) with a radial kernel was used by Huang et al.[49] to predict both the band offset and E_g of nitride-based semiconductors. Similarly, Lee et al.[50] employed SVR for the prediction of E_g of inorganic compounds. Other approaches include the use of crystal graph convolutional neural networks[51] and gradient boosting decision trees (GBDTs).[52] These studies demonstrate the applicability of ML in computational material science. However, the models in these studies had been trained on E_g values that were derived from diverse DFT calculations using different functionals. Considering the inherent imprecision of these calculations, achieving a close alignment with experimental values poses a challenge for the models unless additional adjustments or corrections are made, which, in turn, will incur a high computational cost.

There have also been efforts to develop ML-based solutions to predict E_g against experimental measurements. An artificial neural network-based solution was proposed by Zhaochun et al.[53] to predict E_g and the melting point of binary and ternary compound semiconductors. An SVR technique was employed by Gu et al.[54] for the prediction of E_g of binary and ternary compound semiconductors using a set of experimental data, consisting of 25 binary compounds and 31 ternary compounds. Other regression approaches were explored, which involved ordinary least-squares (OLS) or least absolute shrinkage and selection operator (Lasso), or even the elastic net, which linearly combines the penalties of the Lasso and ridge regression methods.[55] One notable study is by Zhou et al.,[56] who trained a support vector classification and a SVR on experimental measurements to classify and predict E_g of inorganic solids; this data set has become a part of the Matbench test suite v0.1.[57] Zhou et al. trained these models using 136 features, or variables, that were generated solely from the chemical composition of a material. This meant that only chemical composition is required to compute an estimate of E_g against experimental values. When examining the area under the curve of the receiver operating characteristic curve (AUC-ROC), they realized an AUC-ROC of 0.97 for the classification of materials by metallicity, while a coefficient of determination (R^2) of 0.90 and a root-mean squared error (RMSE) of 0.45 eV were achieved for the regression analysis of E_g against experimental measurements. Other algorithms have been applied to such a benchmark set to evaluate their efficacy. The range of mean absolute errors (MAE) realized using alternative algorithms is 0.3310−0.4461 eV.[57−63] See Section 2 Methods for the definition of these performance metrics.

Various studies within this domain showcase a diverse range of methodologies. While certain studies depend upon a restricted set of experimental data, others harness sophisticated algorithms, frequently incorporating a substantial number of input features to attain the previously mentioned model performance. In general, there is a noticeable deficiency in addressing comprehensive statistical feature analysis and selection, mitigating multicollinearity, and conducting permutation analysis, among other considerations. Moreover, exploration of optimization strategies aimed at enhancing model generalization appears to be inadequately discussed.

In this study, we employ the gradient boosted and statistical feature selection (GBFS) workflow, which we have designed for materials-property predictions.[44] The GBFS workflow integrates a distributed gradient boosting framework, in conjunction with exploratory data and statistical analyses and multicollinearity treatments, to discern a subset of features that are highly relevant to the target variable or class within a complex feature space; this affords minimal feature redundancy and maximal relevance to the target variable or classes. The efficacy and the efficiency of the workflow has been showcased in previous materials-property predictions against DFT calculations.[44]

Here, we extend our research into the domain of prediction, utilizing experimental data. Specifically, we implement the GBFS workflow to predict E_g against experimental measurements and explore a multifidelity modeling strategy by augmenting these experimental data with DFT calculations from the MP. Our objective is to showcase the versatility of the proposed workflow as a general-purpose tool, extending beyond the confines of specific data types such as DFT calculations. Additionally, we sought to comprehend the impact of enhancing the predictive model by incorporating data from various streams. The performance of our models is compared to state-of-the-art reports from the literature.

For a like-for-like comparison to the work of Zhou et al., we confine our descriptor sets to those based on chemical

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

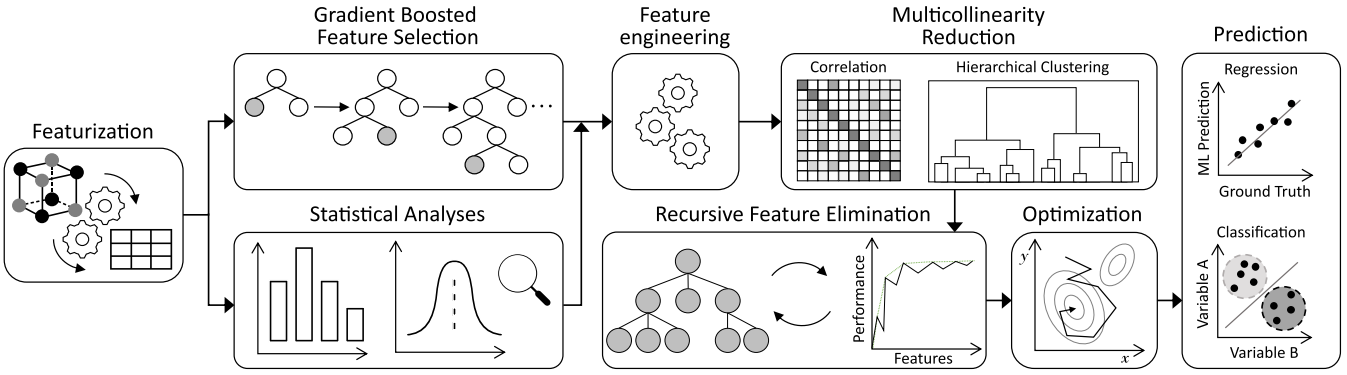


Figure 1. Overview of our operational workflow as described in Section 2—Methods. See ref 44 for a more detailed description.

composition alone, understanding that, most typically, experimental E_g records in the literature lack comprehensive crystallographic information. Later, we extend our analysis to another set of experimental measurements, namely, by Kiselyova et al.[64] Our method highlights the importance of thorough feature analysis and judicious selection over merely complex modeling, as a simple tree-based model trained on features selected and engineered by the GBFS workflow can yield results that are comparable or superior to those reported in the literature. The workflow additionally provides insights into feature interactions and their relevance to the target variable. Furthermore, we apply our final classification and regression models to chemical compositions in Pearson's Crystal Structure Database (94,095) and the MP (105,583). The results are made available as a part of the Supporting Information that serves as a resource for researchers in inorganic material design.

2. METHODS

The experimental measurements employed in this study were compiled from diverse literature sources, as referenced.[56,64−67] The 154,718 DFT calculations utilized for the multifidelity modeling were obtained from MP.[31,32] The results presented in Sections 3.1 and 3.2 were generated using a data set of 6354 chemical compositions. This is identical to the data set utilized by Zhou et al.,[56] facilitating a like-for-like comparison. The additional regression analysis in Section 3.3 considers 7588 chemical compositions, primarily from the work of Kiselyova et al.[64]

A high-dimensional feature vector was generated by leveraging a set of composition-based descriptors. These include the composition featurizer modules from Matminer[68] and Pymatgen.[69] Further features were created using statistics taken over elemental attributes for a particular chemical composition. These calculations are based on data sources, which includes Magpie,[61] Pymatgen,[69] Deml,[70] and the neural network embeddings of elements generated using the MatErials Graph Network.[71] Moreover, the GBFS workflow combines the following: (i) a gradient boosting framework to determine a subset of features that maximize their relevance to the target variable or class, (ii) statistical analyses of the exploratory features to identify those statistically significant to the target variable or class, (iii) a feature engineering step to generate additional features, (iv) a two-step multicollinearity reduction to obtain minimal feature redundancy, which involves a correlation and hierarchical cluster analyses, (v) a recursive feature elimination (RFE), and (vi) a Bayesian optimization to determine the architecture of the final predictive ML model. See Figure 1 for the schematic diagram of the workflow. More details of the methodological workflow can be found in ref 44.

For the classification analysis, we consider the accuracy and F1-score, where the latter is defined as the harmonic mean of the precision and recall as follows

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

where TP and TN are true positive and true negative, and FP and FN are false positive and false negative, respectively. For the regression analysis, we consider the MAE, the mean squared error (MSE), and the coefficient of determination that is defined as the square of the Pearson correlation coefficient, R

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i) \tag{6}$$

$$R = \frac{\text{Covar}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} \tag{7}$$

where y and $\hat{y}$ are the true and predicted values, respectively, over N number of samples; $\text{Covar}(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$ is the covariance between x and y; $\text{Var}(x) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$ and $\text{Var}(y) = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$ are the variance of x and y; $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ are the mean of x and y, respectively. The range of R is $[-1, 1]$, and its value indicates the linear tendency of a quantity to change as the values of another is varied.

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

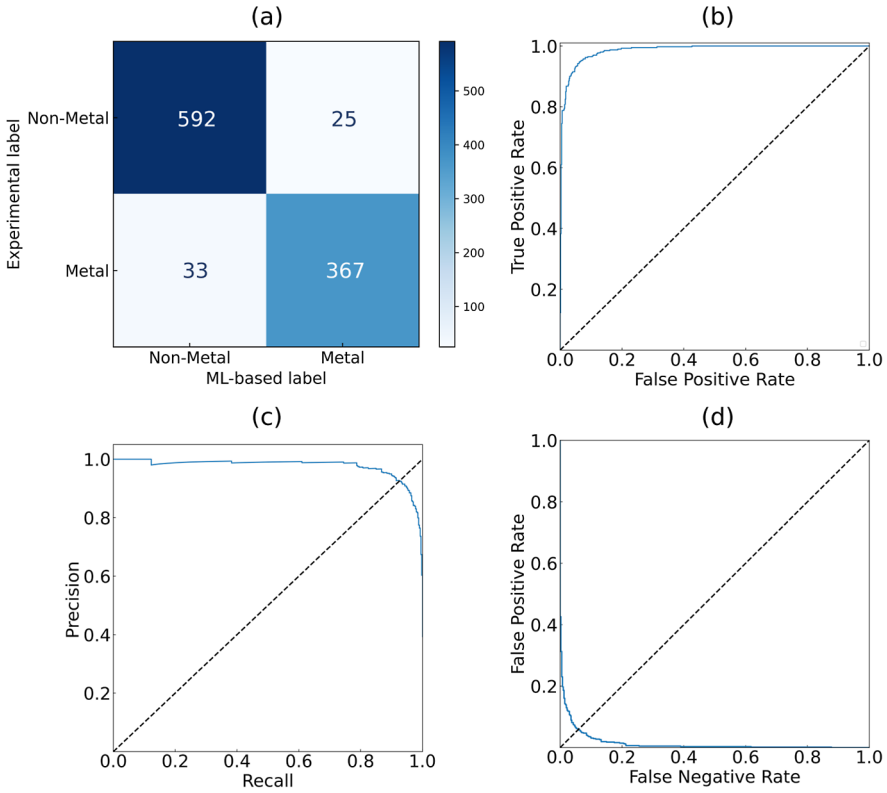


Figure 2. Classification performance on the test set. (a) Confusion matrix, (b) ROC, (c) PR, and (d) DET curves. The macro-averaged AUC-ROC of 0.985, AP of 0.977, and EER of 6% were achieved.

3. RESULTS AND DISCUSSION

3.1. Classification of Materials by Metallicity. *3.1.1. Performance Results.* The classification of materials by their metallicity was performed by using a gradient boosting algorithm using 27 (out of 827) features selected via the GBFS workflow. The model performance is illustrated in Figure 2, along with the performance metrics that are summarized in Table 1. An accuracy score (eq 1) of 0.943 and a balanced accuracy score of 0.938 were achieved. The ROC curve illustrates a diagnostic ability of the model toward the target classes, as the classification threshold is varied by depicting the variation of the true positive rate against the false positive rate (FPR). A macro-averaged AUC-ROC of 0.985 was realized. This indicates that the final classifier is highly discriminative toward the two target classes; this is consistent with the output class-probability distribution (in Supporting Information 1), which is illustrative of an almost binary outcome. In comparison, Zhou et al. achieved an accuracy of 0.92 and an AUC-ROC of 0.97.

Table 1. Summary of the Performance Metrics for the Final Classifier When Applied to the Test Set

	precision	recall	F1-score
metal	0.947	0.959	0.953
nonmetal	0.936	0.917	0.927
macro average	0.942	0.938	0.940
weighted average	0.943	0.943	0.943

The detection error trade-off (DET) shows the change in the FPR relative to the false negative rate (FNR) as the classification threshold is varied. The point at which the FPR and FNR cross the diagonal, known as the equal error rate (EER), is approximately 0.06. This result suggests that a low error rate can be obtained in the model predictions when using selected features. The fact that the plot profiles are fairly symmetric about their respective diagonals appears to indicate that the class imbalance has been well corrected. Moreover, the precision−recall (PR) curve was plotted to assess the effectiveness of the model, owing to the presence of an imbalanced data set. The PR curve demonstrates the quality of the model performance by illustrating the trade-off between precision and recall as the classification threshold is varied. This is important because imbalanced classes can lead to the FPR becoming less informative as a large number of negatives (i.e., false positives and true negatives) yields a low FPR, while still allowing poor precision. An average precision (AP) of 0.977 was achieved; this is the average of the precision at each classification threshold that is weighted by the change in the recall from the prior threshold. The result indicates that both high precision and recall are realized, where a low FPR and a low FNR are attained. This implies that the model is capable of returning a high percentage of positives (i.e., true positives and false negatives) that are mostly classified correctly.

The overall model performance can be evaluated via the F1-score (see eqs 2−4). The macro- and weighted-averaged F1-scores were 0.940 and 0.943, respectively. Zhou et al. did not state their F1-score. The benchmark scores from the Matbench test suite v0.1 show the highest F1-score and balanced accuracy score as 0.920 and 0.921, respectively. This further supports that our classifier is highly discriminative toward the two target classes.

Upon closer examination of the predictions, clear trends are discernible in the characteristics of misclassified chemical

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

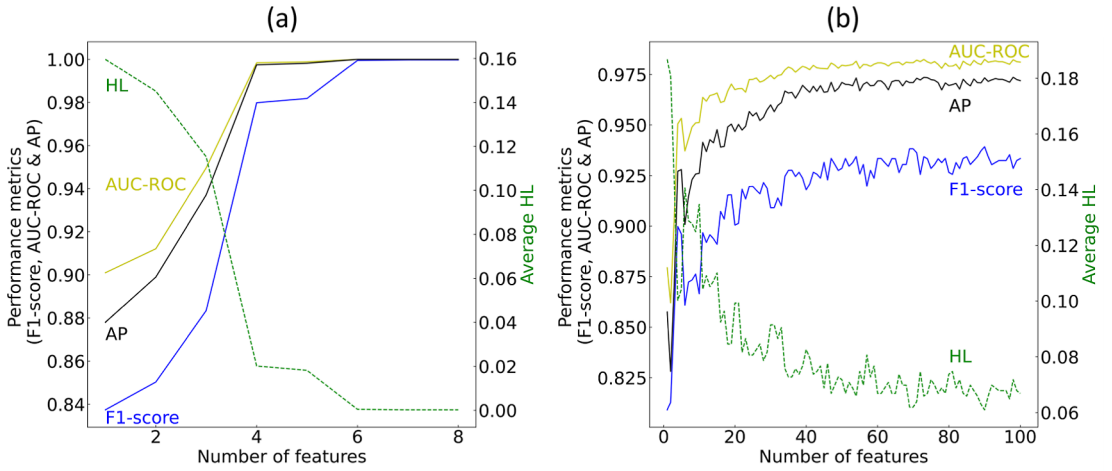


Figure 3. Gradient boosting feature selection results of the classification of materials by metallicity. Performance of GBDTs on (a) the training set and (b) the validation set, where classification models are trained recursively with an increasing subset of features, beginning from the most relevant feature based on the realized total loss reduction.

compounds. An analysis of the two most influential features, based on the total loss reduction, reveals distributions that differ from those observed in the training set. In the scaled distribution of the one-hot-encoding of the highest occupied molecular orbital (HOMO) character corresponding to the p-orbital, we observed that in the training set, the average feature values for nonmetals and metals are ca. 0.82 and 0.26, respectively. This finding indicates a stronger association of nonmetals with HOMO with p-orbital characteristics. A similar feature distribution is noted among the correctly classified compounds in the test set. However, among misclassified compounds in the test set, the average feature values deviate to 0.40 for nonmetals and 0.67 for metals, which goes against the general trend.

Similar observations hold true when the mean number of filled valence p-orbitals among elements in the chemical composition. The scaled distributions between correctly classified chemical compounds and those in the training set are consistent, hovering around 0.50 and 0.26 for nonmetals and metals, respectively. This suggests that nonmetals tend to have a higher mean number of filled valence p-orbitals. In contrast, misclassified compounds exhibit feature values of ca. 0.46 for nonmetals and ca. and 0.42 for metals. Once more, this observation contradicts the overarching pattern. Notably, compounds misclassified as metals manifest a distribution that is closer to nonmetals in both cases. It seems that the model tends to classify chemical compounds as nonmetals when relatively higher feature values are observed for these two features, which are associated with p-orbitals.

These observations align with the fundamental principles of chemistry, as the majority of nonmetal and metalloid elements are situated in the p-block, encompassing groups 13−18 of the periodic table. The p-block comprises chemical elements in which np orbitals are filled, resulting in distinctive chemical properties that distinguish them from those of elements in other blocks of the periodic table. Consequently, understanding the statistical metrics related to the number of filled p-valence orbitals appears to offer insights into chemical elements that are highly discriminatory toward the target classes. This leads to specific chemical elements being associated with having a greater likelihood of being a nonmetal, particularly for elements such as Si ($[Ne]3s^2 3p^2$), Ge ($[Ar]4s^2 3d^{10} 4p^2$), and As ($[Ar]4s^2 3d^{10} 4p^3$), which are common in semiconductors and belong to the p-block in periods 3 and 4 and in groups 14 and 15.

3.1.2. Oversampling. The benchmark data set used to achieve the results above consists of 6354 compounds, which partition into 2458 metal ($E_g > 0$ eV) and 3896 nonmetal ($E_g = 0$ eV) materials, and a train-to-test split ratio of 4:1. A total of 827 composition-based features were computed (see Methods for more details). The imbalanced data set was treated by applying the smoothed random oversampling (smoothed-ROS) method with a shrinkage of 0.35, which is an extension of ROS with the introduction of Gaussian noise. The noise is stochastic. Therefore, it prevented an overtraining of the model on particular values of a feature due to the increased intraclass sample variability, and it appeared to improve model generalization. We found the smoothed-ROS method to be the most effective in alleviating potential learning biases among other oversampling techniques.

3.1.3. Gradient Boosted and Statistical Feature Selection Workflow. The recursive training of GBDTs with an increasing subset of features showed the convergence of AUC-ROC, AP, F1-score, and the average Hamming loss on the training set using ca. 8 of the most relevant features; the feature relevance ranking was initially obtained by observing the loss reduction achieved by each feature when training a GBDT with the entire features. On the validation set, convergence of all four performance metrics was observed once the first ca. 60 features were considered in the model training. The results are shown in Figure 3.

3.1.4. Feature Analyses and Feature Engineering. With these 60 features selected, we performed a generalization of the one-way analysis of variance F-test, Pearson's chi-squared test, mutual information (MI) analysis, and discriminant analysis using logistic regression. Examples of features found to be statistically significant were as follows: the maximum number of filled p-valence orbitals, the average number of p-valence electrons, the mean group of the periodic table, the thermal conductivity, the minimum coefficient of linear thermal expansion, the mean Mendeleev number, the HOMO energy, the mean melting point, and the one-hot-encoding of HOMO character corresponding to the p-orbital. These are consistent with our previous analysis,[44] where a full discussion of the

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

feature interaction and interpretation is made. The features selected by GBFS and statistical analyses were used to engineer new features via the brute-force method. This yielded an additional 159 new features, leading to a total of 219 features that comprised the preliminary subset of features for the classification analysis.

3.1.5. Multicollinearity Reduction and Recursive Feature Elimination. Multicollinearity effects within the data set were reduced by removing features with a correlation coefficient of 0.8 or higher. This reduced the number of features to 105. Further treatment of multicollinearity effects was carried out via a hierarchical cluster analysis, which uses the Spearman rank-order correlation with 1.5 units of Ward's linkage distance as the threshold; this led to the retention of 41 features since only a single feature from each cluster was kept, where the optimal distance threshold was identified using the Elbow method. The corresponding dendrogram of the hierarchical agglomerative clustering, which illustrates the formation of clusters moving up the dendrogram, and the 10-fold permutation feature-importance analysis can be found in Supporting Information 2. Subsequently, the optimal subset of features was identified by eliminating 14 further features via the 10-fold RFE, using a weighted F1-score as the performance metric (see Supporting Information 3 for the RFE plot). This resulted in the final subset of 27 features, identified from among 827 original and 159 engineered features, to be most relevant to the target classes without any prior knowledge of the domain. In contrast, Zhou et al. employed a total of 136 features in the final model.

3.1.6. Model Optimization. A two-step optimization process was followed to determine the architecture of the final classifier. The hyperparameters of the model were optimized by using a combination of grid search and Bayesian optimization by using Gaussian processes. An initial hyperparameter tuning process was performed by scanning the hyperparameter space by using the grid-search method. This subsequently identified the region in which Bayesian optimization was to be applied. Such an optimization strategy proves to be particularly effective for an objective function that has no closed form and is expensive to evaluate and in cases when the evaluations result in noisy responses. The convergence, partial dependence, and evaluation plots from the Bayesian optimization results are shown in Supporting Information 4, and the total loss reduction (i.e., the feature-relevance ranking) realized by the final set of features is shown in Supporting Information 5. In general, the types of features that were selected for the final classification analysis are as anticipated. This finding is consistent with the results from the statistical analyses and those obtained using the SHapley Additive exPlanations (SHAP) framework,[72] which is a game theoretic approach to explain the output of an ML model. See Supporting Information 6 for the average contribution and beeswarm plots from the SHAP analysis.

3.2. Regression Analysis of Band Gap. *3.2.1. Performance Results.* The regression analysis of E_g was performed using a gradient boosting algorithm with 46 features selected via the GBFS workflow. The model performance and error distributions are listed in Figure 4. The solid blue line is the line of best fit between the experimental measurements and ML-based predictions, generated using the OLS method. The linear fit has a gradient of 0.93 and a y-intercept of 0.16 eV. The y-intercept may indicate a small systematic bias for small values of E_g. Nevertheless, the linear fit illustrates the close

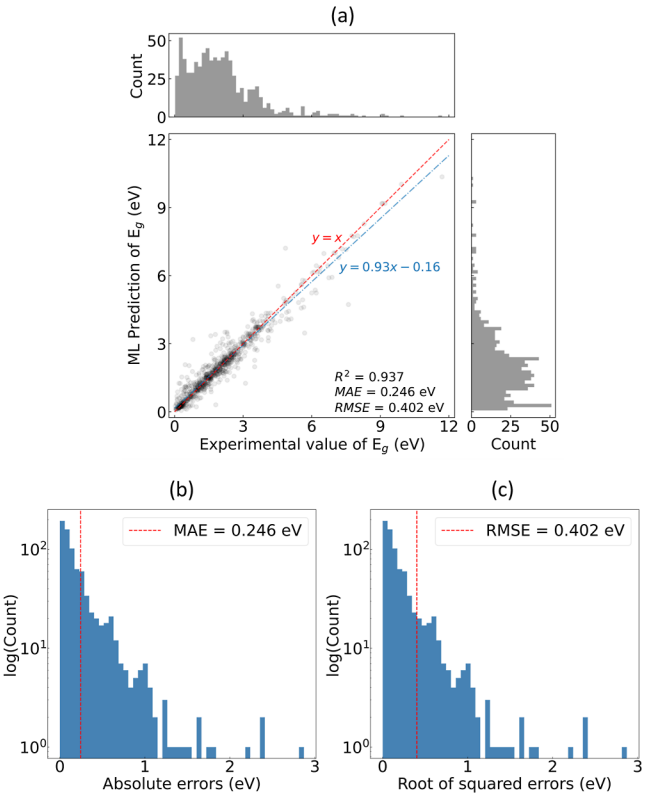


Figure 4. (a) Regression of the ML-based predictions of E_g against experimental measurements, where the regression models are trained on the final subset of features selected by the GBFS workflow. The dashed red line is drawn to represent the hypothetical case, where the ML-based prediction would equal the experimental measurement. The solid blue line is a linear fit generated using the OLS method. Distribution of the (b) absolute errors and (c) root of squared errors of the ML-based prediction of E_g are shown, where the dashed red line indicates the MAE and RMSE, respectively.

correspondence between the experimental measurements and ML-based predictions. This is further supported by the R^2 value of 0.937, which indicates a high correlation between the prediction and the ground truth. Moreover, MAE and RMSE values of 0.246 and 0.402 eV were realized, respectively, where a greater degree of error in the latter is due to the greater penalization of predictions with larger deviations from their true values (see eqs 5−7). The distributions of absolute errors and root of squared errors are shown in Figure 4b,c, respectively. In comparison, Zhou et al. achieved an R^2 of 0.90 and RMSE of 0.45 eV, which also demonstrates the robustness of their SVR model. Our model performance is comparable or superior to those achieved with alternative methods (cf. the aforementioned range of MAE in Introduction).

A closer examination of the training set reveals that the chemical composition with the smallest E_g of 0.02 eV corresponds to $Pb_{0.87}Sn_{0.13}Se$ and the largest E_g of 11.1 eV corresponds to MgF_2. The E_g value is less than 3 eV for the majority of the chemical compositions of materials in the training set, as illustrated by the population distribution plot in Figure 5. This explains the relatively larger deviations in the prediction for semiconductors with larger E_g values. Specifically, we observe an underestimation of E_g where its value is large, as depicted by the line of best fit (in blue) in Figure 4a, which sits below the diagonal line in red. We attribute such an

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

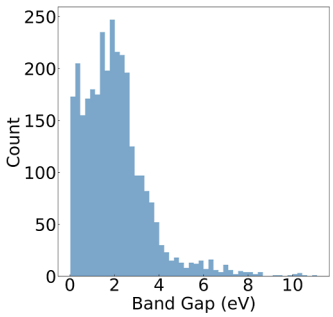


Figure 5. Distribution of band gap (E_g) values of nonmetals in the training set.

observation to the lack of wide or ultrawide band gap semiconductors in the training data. Nevertheless, the ML modeling approach has demonstrated an accurate prediction of E_g in the absence of structural information. This is despite the fact that our methodology does not involve specific adjustments or treatments to accommodate different crystal forms of the same chemical compound, a phenomenon known as polymorphism. Our approach strictly trains the model to be agnostic to polymorphs, aligning with the methodology employed by Zhou et al. This ensures a like-for-like comparison in our study. This is to say that the models demonstrate good performance, even though they are unaware of the different crystal forms that the sample compound may adopt, owing to the various ways in which chemical elements can arrange within unit cells of the crystalline lattice of each crystal form. See Section 3.3 for the additional regression analysis where we evaluate the model predictions against the median and mean experimental E_g values. Taking our attention back to the error distribution plots in Figure 4, which has a log scale in the *y*-axis, we see that the majority of predictions have an error below ca. 1 eV. At a closer examination, out of 780 chemical compositions in the test set, 670 predictions (ca. 86%) have an absolute error below 0.5 eV and 545 predictions (ca. 70%) have an absolute error below 0.25 eV.

The results discussed thus far present predictions of E_g values against experimental measurements with promising statistical figures-of-merit. Nonetheless, it is important to validate these results by considering how these predictions fare against well-studied inorganic compounds rather than simply demonstrating their collective statistical quality in an anonymized form. A comparative analysis is, therefore, conducted against six unseen compounds that are extensively researched both experimentally and at various theoretical levels. The results are summarized in Table 2.

The aforementioned underestimation of DFT-based E_g values that have been computed using a PBE functional is clearly apparent among the results, with a negative percentage difference being shown across all compounds. Although relatively smaller in magnitude, our model (labeled GBFS E_g) exhibits a negative percentage difference beyond an E_g value of ca. 3 eV. This is consistent with our previous discussion in the regression analysis, where we observed a systematic positive bias of ca. 0.16 eV at low values of E_g and an underestimation at higher E_g values, with a gradient of 0.93 between the ML prediction and the ground truth—an effect that becomes exaggerated at higher values of E_g. This explains the largest deviations that are observed, for example, in LiF, which predicts an E_g value of 14.2 eV. We attribute this anomaly to the lack of wide or ultrawide band gap semiconductors within the training set, as previously discussed. Nevertheless, ML-based predictions (both GBFS and SVR E_g) yielded lower MAE and RMSE values when compared with high-throughput calculations that incorporated a PBE functional. The result pertaining to GW-type calculations realized the lowest errors among the methods. However, such an approach incurs the greatest computational cost and cannot provide an efficient and automated approach to the E_g prediction. The second lowest errors originated from the high-throughput calculations that employed hybrid functionals (HSE); this is another computationally expensive approach. The MAE achieved using HSE is comparable to the value obtained by our method, while the RMSE is ca. 1 eV lower for the former. Again, this difference stems from the error associated with wide and ultrawide band gap semiconductors. We anticipate a significant improvement of ML models pending the availability of a significantly greater number of ultrawide band gap semiconductors.

We now delve into a more detailed examination of the feature interactions that contributed to the aforementioned results in the regression analysis. The highest relevance, as indicated by the realized total loss reduction, is associated with the HOMO energy (feature number m = 593). This is succeeded by the standard deviation of the periodic group among elements in the chemical composition (m = 258), the fraction of p-valence electrons (m = 586), the average Mendeleev number among elements in the chemical composition (m = 115), the average deviation of electronegativity among elements in the chemical composition (m = 158), and the average deviation of the number of filled valence p-orbitals among elements in the composition (m = 170).

The selection of HOMO energy by the GBFS workflow as the feature with the highest relevance was anticipated given that it is a parameter directly involved in E_g estimation. The correlation between the HOMO energy and E_g is ca. −0.58 in the training data, indicating that lower HOMO energies correspond to larger E_g values. This correlation is logical as the

Table 2. Summary of the Band Gap Predictions (E_g in eV) against Experimental Measurements

composition	experimental E_g	GBFS E_g	SVR E_g[56]	PBE E_g	GW E_g	HSE E_g
PbTe	0.19[73]	0.215 (13%)	0.2 (5%)	0 (−100%)[73]	0.26 (36%)[73]	0.19 (0%)[74]
$CuSbS_2$	1.38[75]	1.40 (1%)	1.39 (1%)	0.9 (−35%)[75]	1.1 (−20%)[75]	1.69 (22%)[75]
GaN	3.2[73]	3.00 (−6%)	4.45 (39%)	1.62 (−49%)[73]	3.32 (4%)[73]	3.14 (−2%)[26]
TiO_2	3.42[24]	3.25 (−5%)	3.99 (16%)	2.13 (−37%)[24]	3.73 (9%)[24]	3.67 (7%)[24]
ZnS	3.91[73]	3.41 (−13%)	3.12 (−20%)	2.07 (−47%)[73]	4.15 (6%)[73]	3.49 (−11%)[26]
LiF	14.2[73]	10.53 (−26%)	9.87 (−30%)	9.2 (−35%)[73]	15.1 (6%)[73]	11.47 (−19%)[76]
MAE		0.76	1.16	1.73	0.32	0.63
RMSE		2.29	3.54	5.47	0.18	1.30

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

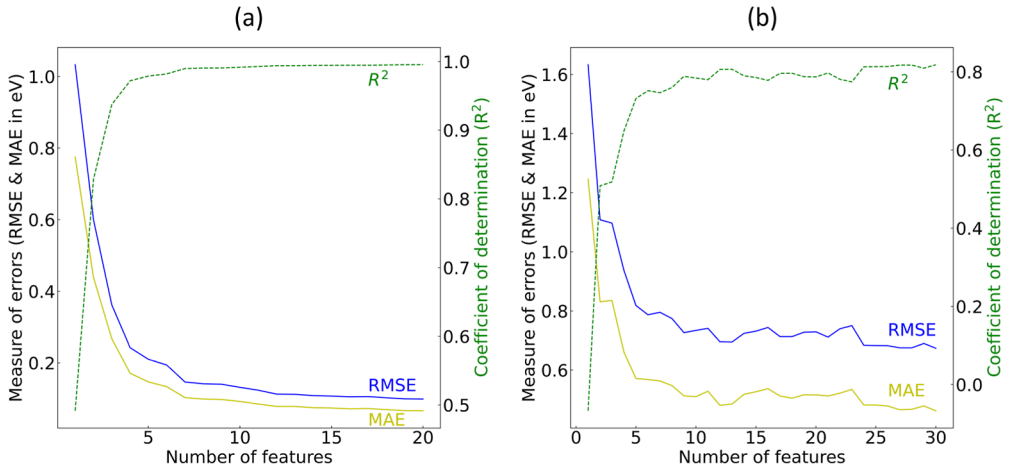


Figure 6. Gradient boosting feature selection (GBFS) result of the regression analysis of E_g. Model performance of GBDTs on (a) the training set and (b) the validation set, where regression models are trained recursively with an increasing subset of features, beginning from the most relevant feature based on the realized total loss reduction.

HOMO represents the highest energy molecular orbital that contains electrons, akin to the valence band in Band theory. The energy of the lowest unoccupied molecular orbital represents the minimum energy level into which an electron can be excited. While its inclusion was expected, its correlation with the E_g values in the training data is not as pronounced as that of the HOMO energy, registering a correlation magnitude of only 0.17.

The subsequent set of features that are deemed significant in predicting E_g are associated with statistical values based on the period group number and Mendeleev order number among elements in the chemical composition. These features are directly linked to the chemical compositions of the compounds. The standard deviation of the periodic group among elements in the chemical composition exhibits a correlation of ca. 0.38 with the target variable. This suggests a discernible trend between the two, indicating that higher deviations in the periodic group number among elements are associated with elevated values of E_g. This aligns with the observation that some of the largest E_g values in the training data are found in chemical compositions such as MgF_2, NaF, RbF, and BeO. These compositions involve *s*-block metals in groups 1 and 2 paired with p-block nonmetals and halogens in groups 16 and 17 (e.g., oxygen and fluorine), generating some of the largest standard deviations of the periodic group among elements in the chemical composition. Moreover, the Mendeleev number, distinct from the atomic numbering system, is an ordering assigned to each chemical element in the periodic system. Its purpose is to arrange elements so that those with similar behaviors are placed consecutively. Similar to the previous feature, a clear trend is apparent, wherein a lower mean Mendeleev number among elements in the composition correlates with higher values of E_g, with a correlation of ca. −0.53. In the training data, chemical compounds with some of the highest E_g values exhibit a mean Mendeleev number of ca. 10.

Another important set of features to be acknowledged involves p-valence electrons. Two relevant features are identified by the GBFS workflow, which are (i) the fraction of p-valence electrons and (ii) the average deviation of the number of filled valence p-orbitals among elements in the chemical composition. As discussed in the classification problem, these observations align with fundamental principles of chemistry, given that the majority of nonmetal and metalloid elements reside in the p-block, spanning groups 13−18 of the periodic table. The p-block encompasses chemical elements in which np orbitals are filled, resulting in distinctive chemical properties that set them apart from elements in other blocks of the periodic table. Notably, elements such as Si, Ge, and As, which are well-known in semiconductor applications, fall within this category. A correlation of 0.61 is observed against E_g values in the training data for the fraction of p-valence electrons.

The final among the most pertinent features to be discussed is the average deviation of electronegativity among elements in the chemical composition. This feature demonstrated the highest correlation with the E_g values in the training set, registering a value of ca. 0.68. The elements with the highest electronegativity are typically found in groups 16 and 17 (e.g., fluorine, oxygen, and chlorine), while those with the lowest electronegativity are situated in groups 1 and 2 (e.g., sodium, lithium, potassium, and magnesium). Consequently, the highest average deviation of electronegativity is generally observed between these two regions of the periodic table. The rationale behind selecting this feature lies in the fact that pairing a metal with a nonmetal element corresponds to a large difference in orbital energy. This phenomenon becomes more pronounced as the disparity in electronegativity between the pair of elements increases. In other words, a substantial difference in the electronegativity of two elements in a compound leads to an increase in its ionic properties, which in turn reduces the overlap of orbitals and elevates E_g of a material. Consequently, the probability of a material exhibiting a larger E_g increases when elements with higher electronegativity, concentrated in the p-block, are incorporated into the material composition. This is intuitive, as the strength of ionic bonding or electrostatic interaction is directly determined by the difference in electronegativity between neighboring ions. The result implies that nonmetal elements, such as oxygen and fluorine, are important attributes to consider when distinguishing metallic bonding from other types of bonding. This stands to reason since a metal can exist in the form of oxides, while ionic and metallic characteristics are distinct; an

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

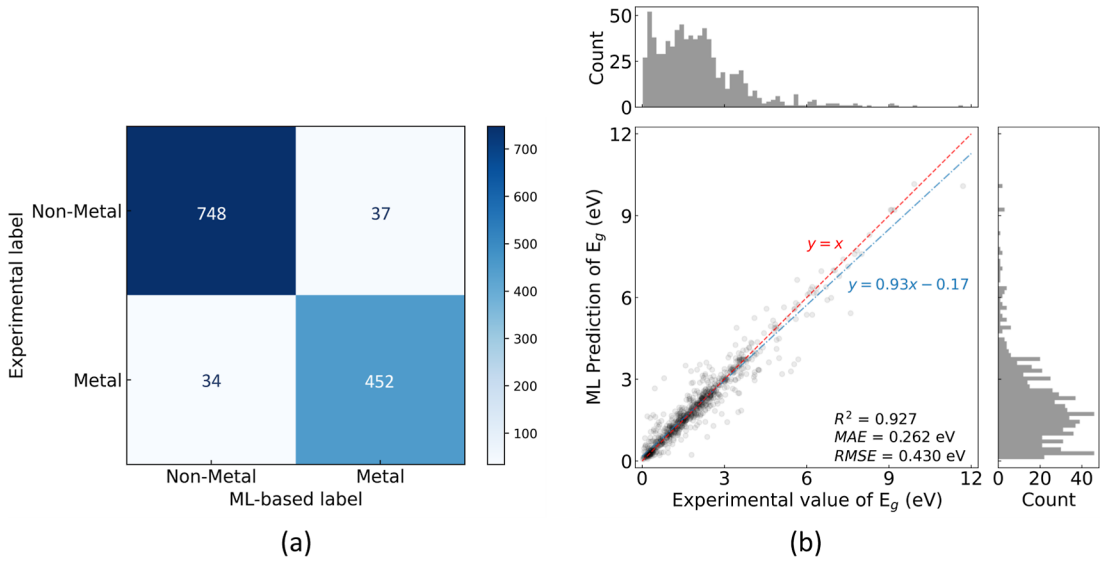


Figure 7. Multifidelity model performance against experimental measurements and DFT-based computed results for the (a) classification of materials by metallicity and (b) regression analysis of E_g in the test set, where the classification model was trained with 22 features and the regression model was trained on 25 features, both selected via the GBFS workflow. For the classification process, a macro-weighted AUC-ROC of 0.987, AP of 0.982, F1-score of 0.941, and a balanced accuracy of 0.941 were achieved.

increase in the former contributes to an increase in E_g as orbital overlap decreases.

3.2.2. Gradient Boosted and Statistical Feature Selection Workflow. The regression analysis considered 3896 compounds ($E_g > 0$), with 2458 materials of unique chemical compositions and a train-to-test split ratio of 4:1. In common with the classification problem, we employed only composition-based descriptors and processed the features via the GBFS workflow. This led to the selection of 46 features as the final subset of features among more than 820 exploratory features. The performance of the regression models during feature selection, on the training set and the validation set, is shown in Figure 6. The performance metrics of interest are R^2, MAE, and RMSE. For both the training and the validation sets, the performance metrics plateaued before the first ca. 30 features, with a relatively worse performance on the out-of-sample validation set as anticipated. Subsequently, we performed the aforementioned statistical feature analyses and engineering, multicollinearity reduction, and permutation-importance analysis (Supporting Information 7), RFE (Supporting Information 8), and Bayesian optimization of the final regression model (Supporting Information 9), and we evaluated the total loss reduction achieved by the selected 46 features (Supporting Information 10). An independent feature analysis was conducted using the SHAP framework, which is in agreement with the features selected by the GBFS workflow (Supporting Information 11).

3.3. Multifidelity Modeling. *3.3.1. Multifidelity Model.* Another direction explored in this research was to develop a multifidelity model using 154,718 DFT results from the MP, with 105,584 unique chemical compositions. An auxiliary model was trained on these chemical compositions of materials for the classification process and on 69,219 chemical compositions with $E_g > 0.35$ eV for the regression process, both using the GBFS workflow and a train-to-test split ratio of 4:1 (see Supporting Information 12 for the performance of the auxiliary models). An energy cutoff of 0.35 eV was applied to the DFT data set in order to mitigate inherent errors in DFT calculations of E_g. We anticipated that the multifidelity approach could account for the lack of wide band gap semiconductors. However, we observed a comparable performance to that of the models that had been constructed without the multifidelity strategy. The results are listed in Figure 7. Closer examination of the feature interactions showed that the incorporation of DFT-based auxiliary models led to a smaller final subset of features selected via the GBFS workflow: 22 features for the classification and 25 features for the regression. These correspond to a reduction of 5 features in the former and a reduction of 21 features for the latter. Moreover, DFT-aware features are among the top three in the feature relevance ranking, which is based on the loss reduction realized when training the predictive models (Supporting Information 13). This demonstrates the benefit associated with such a modeling approach, which has gained popularity within computational material science.

It is crucial to distinguish our work from the multifidelity ML modeling approach conducted by Pilania et al.[29] Their methodology can be classified as multifidelity because it incorporates calculations on 600 chemical compounds using different functionals for each compound, which have varying levels of exchange correlation within DFT. This entails the amalgamation of low- and high-fidelity DFT calculations, enabling cost-effective predictions at a higher fidelity level. Given the multifidelity modeling approach in our work arises from the fact that we have used computational and experimental data sources as input, a comparison of our work with theirs presents challenges, as their exclusive focus on computational data in the context of multifidelity ML models differs from our approach. Nevertheless, the use of a multifidelity approach in their work yielded improvements, while our study demonstrated a singular effect—a substantial reduction in the feature space. Despite being the sole observed outcome, this reduction remains a crucial and noteworthy result, as it can lead to a reduction in the feature space by up to ca. 46%; this renders many of the essential features identified in Section 3.2 no longer necessary.

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

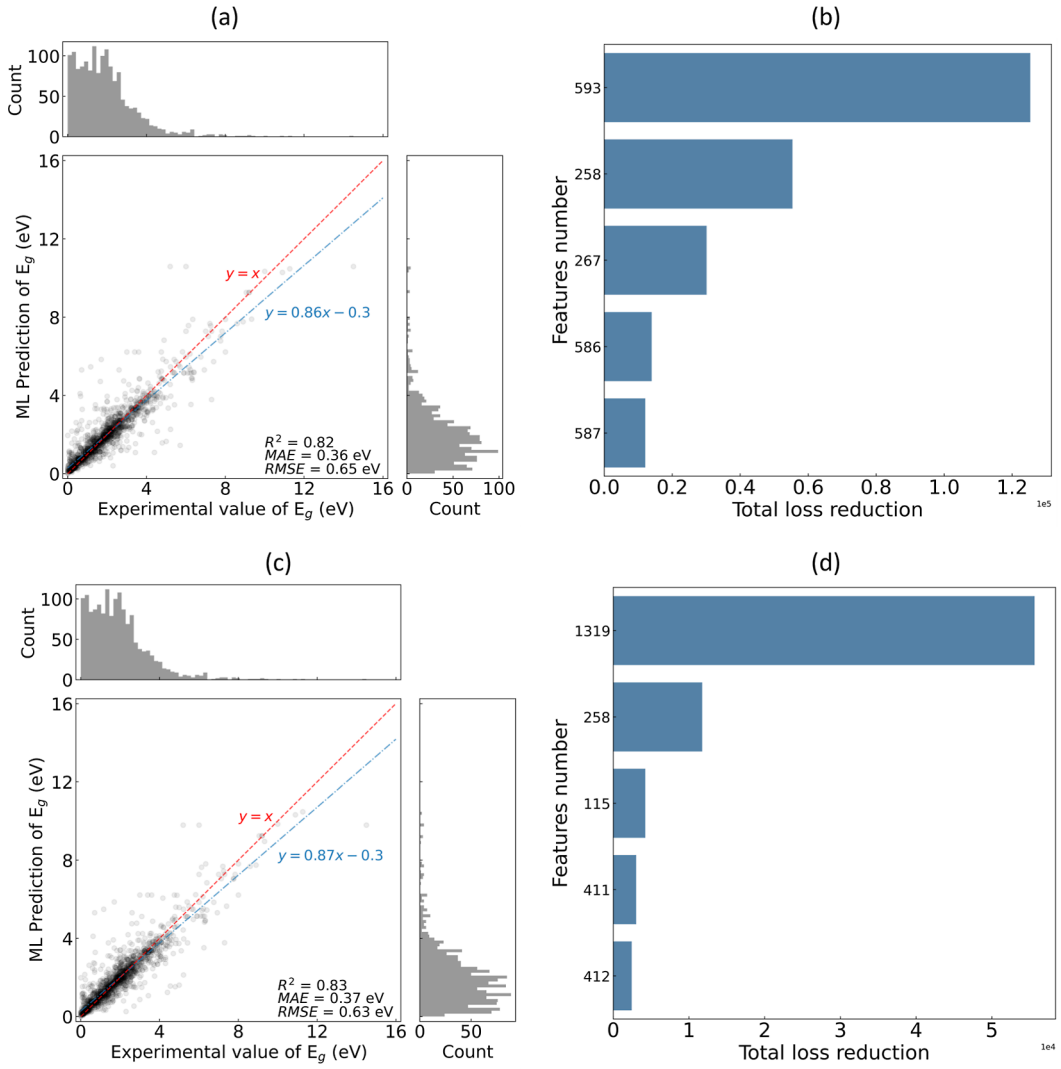


Figure 8. Regression analysis of E_g in the test set against (a) experimental measurements with 42 features, with the 5 most relevant features shown in (b), and against (c) experimental measurements and DFT calculations (i.e., using the multifidelity modeling approach) with 16 features, with the five most relevant features shown in (d). The features were selected via the GBFS workflow. Feature number 1319 corresponds to the DFT-based prediction of E_g values, which realized the largest total loss reduction. See *feature_list.csv* for the full list of feature names.

It is unsurprising to encounter this difference, considering the limited data set of 600 chemical compounds in the study by Pilania et al.[29] In such cases, the amalgamation of multiple data sources in the training process would prove advantageous. This is because the augmented training data for the multifidelity model can either facilitate the exploration of a larger chemical space, often mitigating the necessity for extreme extrapolation of learned relationships into unseen territories, or strengthen existing relationships with greater statistical certainty.

To demonstrate this, we employed a pseudocomparison multifidelity modeling approach utilizing only 600 experimental measurements that were randomly selected, in conjunction with ca. 150,000 DFT calculations. More specifically, we conducted two separate regression analyses: the first involved 600 experimental measurements, and the second utilized the same 600 experimental measurements alongside ca. 150,000 DFT calculations, employing a multifidelity modeling strategy. In the former scenario, we attained an R^2 of 0.80, MAE of 0.50 eV, and RMSE of 0.77 eV. Conversely, in the latter scenario, we achieved an R^2 of 0.86, MAE of 0.44 eV, and RMSE of 0.65 eV. As anticipated, the implementation of the multifidelity modeling strategy led to improvements in the quantified statistical figures-of-merit compared to a standard regression model; this is despite the recognized errors inherent in DFT calculations. Our analysis suggests a trade-off between the broader chemical coverage achieved by incorporating DFT calculations and the inherent uncertainties associated with them. Notably, the latter serves as a limiting factor in this study.

3.3.2. Another Experimental Data Set. We extended our regression analysis to another set of experimental measurements, namely, those reported by Kiselyova et al.[64] Their data set consists of 7588 chemical compositions of materials, with 3233 unique chemical compositions. The result with and without implementing a multifidelity modeling strategy is shown in Figure 8. In common with the findings noted above, we find that the model performances are comparable to each other. The use of the multifidelity approach resulted in a reduction of 26 features, with a predominant portion of the total loss reduction being attributed to the DFT-based prediction of E_g values. We infer that while DFT calculations

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

Table 3. Examples of Input Chemical Composition and the Corresponding Prediction of Band Gap (E_g), Sorted by the Absolute Percentage Difference from the Mean Experimental Values in Ascending Order[a]

composition	experimental E_g (eV)					pred. E_g (eV)	% diff. from mean	% diff. from median
	n	min	max	mean	median			
$Ga_2(TeO_3)_3$	2	4.14	4.15	4.15	4.15	4.11	−1.0	−1.0
$Hg_3(SI)_2$	1	2.25	2.25	2.25	2.25	2.22	−1.3	−1.3
$BaGeS_3$	1	2.46	2.46	2.46	2.46	2.49	1.2	1.2
HgCl	3	2.84	3.80	3.38	3.50	3.43	1.5	−2.0
$K_3Th_2Cu_3S_7$	1	2.49	2.49	2.49	2.49	2.45	−1.6	−1.6
Bi_2SO_2	2	1.12	1.50	1.31	1.31	1.33	1.5	1.5
$K_2Hg_3(GeS_4)_2$	2	2.64	2.70	2.67	2.67	2.71	1.5	1.5
$La_3Mo_4O_{16}F$	1	3.70	3.70	3.70	3.70	3.64	−1.6	−1.6
$Ba_3Cd(SnS_4)_2$	1	2.75	2.75	2.75	2.75	2.80	1.8	1.8
$Cs_2Ba_3(P_2O_7)_2$	2	5.06	6.31	5.69	5.69	5.79	1.8	1.8
Ga_2PbS_4	4	2.38	2.55	2.46	2.46	2.41	−2.0	−2.0
$La_4Fe(SbS_5)_2$	1	1.00	1.00	1.00	1.00	1.03	3.0	3.0
$Zn(InTe_2)_2$	7	1.13	1.95	1.74	1.87	1.68	−3.4	−10.2
$BaCu_2GeS_4$	8	2.29	2.49	2.43	2.47	2.53	4.1	2.4
Li_2TeMoO_6	1	3.50	3.50	3.50	3.50	3.34	−4.6	−4.6
$PHPb_2N_2O_9$	1	3.81	3.81	3.81	3.81	3.62	−5.0	−5.0
$NbIn(TeO_4)_2$	1	3.50	3.50	3.50	3.50	3.32	−5.1	−5.1
In_5AgS_8	3	1.60	1.79	1.72	1.76	1.63	−5.2	−7.4
$Ba_2DyGaSe_5$	2	2.35	2.35	2.35	2.35	2.22	−5.5	−5.5
$Na_4Mg(SiSe_3)_2$	1	2.85	2.85	2.85	2.85	2.69	−5.6	−5.6
$Na_2Cd(GeS_3)_2$	3	2.57	3.21	2.79	2.60	2.96	6.1	13.8
$Ba_2ErGaSe_5$	2	1.95	1.95	1.95	1.95	2.07	6.2	6.2
$InHg_7S_6Cl_5$	1	2.54	2.54	2.54	2.54	2.37	−6.7	−6.7
$BaAl_4S_7$	1	3.74	3.74	3.74	3.74	3.48	−7.0	−7.0
$Rb_2Cd_3(B_4O_7)_4$	1	4.76	4.76	4.76	4.76	4.42	−7.1	−7.1
$Cd(GaSe_2)_2$	18	2.10	2.75	2.39	2.41	2.56	7.1	6.2
$KThCuS_3$	1	2.95	2.95	2.95	2.95	2.72	−7.8	−7.8
$Sb_8I_2O_{11}$	5	2.72	2.72	2.72	2.72	2.50	−8.1	−8.1
KCu_2BiS_3	1	1.29	1.29	1.29	1.29	1.40	8.5	8.5
$K_2Mn(SnSe_3)_2$	1	2.00	2.00	2.00	2.00	1.82	−9.0	−9.0
Rb_2VAgS_4	1	1.83	1.83	1.83	1.83	2.00	9.3	9.3
$PrMoO_4F$	1	3.64	3.64	3.64	3.64	3.30	−9.3	−9.3
$ZnSi(AgS_2)_2$	1	3.28	3.28	3.28	3.28	2.96	−9.8	−9.8
$K_2Sn_2Hg_3S_8$	2	2.40	2.50	2.45	2.45	2.19	−10.6	−10.6
PPdS	12	0.70	1.40	1.16	1.38	1.30	12.1	−5.8
$RbH_2(IO_3)_3$	2	4.07	5.08	4.58	4.58	4.00	−12.7	−12.7
$In_4Bi_3S_{10}$	1	1.42	1.42	1.42	1.42	1.24	−12.7	−12.7
$ZnCu_2SiS_4$	3	3.00	3.25	3.17	3.25	2.75	−13.2	−15.4
Te_3As_2	28	0.48	1.88	1.22	1.35	1.40	14.8	3.7
$NbInBi_2O_7$	1	2.70	2.70	2.70	2.70	2.30	−14.8	−14.8
$InCuGeSe_4$	1	1.30	1.30	1.30	1.30	1.50	15.4	15.4
$BaBiBS_4$	1	2.34	2.34	2.34	2.34	1.98	−15.4	−15.4
$NaSc(SeO_3)_2$	1	5.50	5.50	5.50	5.50	4.56	−17.1	−17.1
$RbBaPS_4$	1	3.30	3.30	3.30	3.30	2.70	−18.2	−18.2
$Ba_3Er_2(PS_4)_4$	1	3.30	3.30	3.30	3.30	2.69	−18.5	−18.5
$SbPbIO_2$	1	2.48	2.48	2.48	2.48	1.98	−20.2	−20.2
$Zn_3(PS_4)_2$	2	3.07	3.19	3.13	3.13	2.49	−20.4	−20.4
$AgBi(PSe_3)_2$	1	1.40	1.40	1.40	1.40	1.11	−20.7	−20.7
$NaB_5(H_2O_5)_2$	2	5.61	6.13	5.87	5.87	7.13	21.5	21.5
$Cs_2Mg_2(WO_4)_3$	1	4.53	4.53	4.53	4.53	3.46	−23.6	−23.6
$KP(HO_2)_2$	3	3.20	7.00	5.72	6.95	7.09	24.0	2.0
$Ti(Bi_3O_5)_4$	1	3.09	3.09	3.09	3.09	2.33	−24.6	−24.6
$LiZnPS_4$	1	3.44	3.44	3.44	3.44	2.57	−25.3	−25.3
$NaYb(PS_3)_2$	1	1.85	1.85	1.85	1.85	2.37	28.1	28.1
FeS_2	9	0.02	1.25	0.83	0.92	1.17	41.0	27.2
$Bi_4Pb_7Se_{13}$	4	0.23	0.29	0.26	0.26	0.14	−46.2	−46.2
$Hg_8Bi_3As_4Cl_{13}$	1	4.30	4.30	4.30	4.30	2.04	−52.6	−52.6
$SrBe_2(BO_3)_2$	1	4.69	4.69	4.69	4.69	7.32	56.1	56.1

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

Table 3. continued

composition	experimental E_g (eV)					pred. E_g (eV)	% diff. from mean	% diff. from median
	n	min	max	mean	median			
$Ba_8U_2PdSe_{16}$	2	0.18	1.60	0.89	0.89	1.43	60.7	60.7
As	10	0.17	1.25	0.80	1.14	1.38	72.5	21.1

[a]Data from ref 64. Here, n is the number of experimental values sampled; min and max represents their range; mean and median are their corresponding descriptive statistics.

facilitate the creation of less complex models in terms of the number of input features, they do not necessarily improve the model accuracy, possibly because of the inherent limitations associated with these calculations.

Moreover, this experimental data set is used to assess the predictions made on the chemical compositions within Pearson's Crystal Structure Database (94,095) and the MP (105,583). A sample of 60 chemical compositions, previously unseen by the model, was randomly chosen, and the predictions were compared against the experimental measurements. The results are summarized in Table 3. As anticipated, the model has a tendency to underestimate the experimental E_g values, with 37 and 39 (out of 60) chemical compositions exhibiting a negative percentage difference from the mean and from the median E_g values, respectively. The average absolute percentage differences from the mean and from the median are ca. 15 and 13%, respectively. Notably, a lower absolute percentage difference is observed when considering the median E_g value for specific compounds such as $BaCu_2GeS_4$, PPdS, $KP(HO_2)_2$, FeS_2, and As. This suggests that a lower deviation can be realized when considering the median as the measure of central tendency in the presence of multiple experimental measurements, as outliers exert a relatively minimal effect on the median of a given data set compared to the mean. These predicted E_g values are included as Supporting Information, potentially serving as a valuable resource for researchers engaged in the study of inorganic materials and their band gaps.

4. CONCLUSIONS

This study has employed a ML-based feature selection and statistical feature analysis workflow to train predictive models that classify materials by their metallicity and predict their band gap (E_g). Our feature-selection workflow integrates a distributed gradient boosting framework, in conjunction with exploratory data and statistical analyses and multicollinearity treatments, to identify and select a subset of features that is highly relevant to the target variable or class within a complex feature space; this affords minimal feature redundancy and maximal relevance to the target variable or classes. Gradient boosting trees are subsequently trained with the selected features, which are solely based on the chemical composition of a material. The classification model realized a macro-averaged F1-score of 0.940, AP of 0.977, and AUC-ROC of 0.985, while the regression model achieved an MAE of 0.246, RMSE of 0.402, and R^2 of 0.937. The results are superior to high-throughput DFT calculations that employ a PBE functional while being either superior to or comparable to complex algorithms reported in the literature. This exemplifies the efficacy of our modeling approach and highlights the importance of thorough feature analysis and judicious selection over merely complex modeling. We further explored the multifidelity modeling strategy and found that such an approach can reduce the number features required to train a model. We applied our models to chemical compositions in Pearson's Crystal Structure Database (94,095) and the MP (105,583). The results are made available as a part of the Supporting Information, serving as a resource for researchers in the development of novel inorganic materials.

■ ASSOCIATED CONTENT

Data Availability Statement

We have made the code for the feature selection, statistical analyses, multicollinearity reduction, RFE, and Bayesian optimization available at https://github.com/Songyosk/BGML. The data sets used in this work are available from the MP v2022.10.28[31,32] and the Matbench test suite v0.1.[57] The aggregated nonmetal data and our model predictions are provided in *bandgap_data.xlsx* as a part of the Supporting Information.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c01897.

Class-probability distributions, hierarchical cluster and permutation importance analyses, recursive feature elimination, convergence, partial dependence, and evaluation plots obtained during Bayesian optimization, feature-relevance ranking, SHAP analysis, for both the regression and the classification problem, and band gap predictions against DFT calculations and the feature-relevance score for the multifidelity models (PDF)

Full list of feature names and their corresponding feature numbers (XLSX)

Aggregated nonmetal data and model predictions (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Jacqueline M. Cole − *Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K.; ISIS Neutron and Muon Source, STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0QX, U.K.; Research Complex at Harwell, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA, U.K.;* orcid.org/0000-0002-1552-8743; Email: jmc61@cam.ac.uk

Authors

Son Gyo Jung − *Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K.; ISIS Neutron and Muon Source, STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0QX, U.K.; Research Complex at Harwell, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA, U.K.*

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200

Guwon Jung − *Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge CB3 0HE, U.K.; Research Complex at Harwell, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0FA, U.K.; Scientific Computing Department, STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0QX, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01897

Author Contributions

J.M.C. conceived the overarching project. S.G.J. and J.M.C. designed the study. S.G.J. developed the workflow, performed the data acquisition and featurization, the statistical analyses, the model pretraining and fine-tuning, and analyzed the data under the PhD supervision of J.M.C. G.J. assisted with the data gathering, the model development, and model optimization. G.J. further contributed to the analysis of the results. S.G.J. drafted the manuscript with assistance from J.M.C. All authors read and approved the final agreed manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.M.C. is grateful for the BASF/Royal Academy of Engineering Research Chair in Data-Driven Molecular Engineering of Functional Materials, which is partly sponsored by the Science and Technology Facilities Council (STFC) via the ISIS Neutron and Muon Source; this chair is supported by a PhD studentship (for S.G.J.). STFC is also thanked for a PhD studentship that is sponsored by its Scientific Computing Department (for G.J.).

■ REFERENCES

(1) Schubert, E. F.; Kim, J. K. Solid-state light sources getting smart. *Science* **2005**, *308*, 1274−1278.
(2) Zhang, L.; Yang, X.; Jiang, Q.; Wang, P.; Yin, Z.; Zhang, X.; Tan, H.; Yang, Y.; Wei, M.; Sutherland, B. R.; et al. Ultra-bright and highly efficient inorganic based perovskite light-emitting diodes. *Nat. Commun.* **2017**, *8*, 15640.
(3) Nadarajah, A.; Word, R. C.; Meiss, J.; Könenkamp, R. Flexible inorganic nanowire light-emitting diode. *Nano Lett.* **2008**, *8*, 534−537.
(4) Chen, L.; Lin, C.-C.; Yeh, C.-W.; Liu, R.-S. Light converting inorganic phosphors for white light-emitting diodes. *Materials* **2010**, *3*, 2172−2195.
(5) Radisavljevic, B.; Radenovic, A.; Brivio, J.; Giacometti, V.; Kis, A. Single-layer MoS2 transistors. *Nat. Nanotechnol.* **2011**, *6*, 147−150.
(6) Polman, A.; Knight, M.; Garnett, E. C.; Ehrler, B.; Sinke, W. C. Photovoltaic materials: Present efficiencies and future challenges. *Science* **2016**, *352*, aad4424.
(7) Miles, R. W.; Zoppi, G.; Forbes, I. Inorganic photovoltaic cells. *Mater. Today* **2007**, *10*, 20−27.
(8) Liu, F.; Zeng, Q.; Li, J.; Hao, X.; Ho-Baillie, A.; Tang, J.; Green, M. A. Emerging inorganic compound thin film photovoltaic materials: Progress, challenges and strategies. *Mater. Today* **2020**, *41*, 120−142.
(9) Urban, A.; Seo, D.-H.; Ceder, G. Computational Understanding of Li-Ion Batteries. *npj Comput. Mater.* **2016**, *2*, 16002.
(10) Jónsson, E.; Johansson, P. Modern Battery Electrolytes: Ion−Ion Interactions in Li^+/Na^+ Conductors from DFT Calculations. *Phys. Chem. Chem. Phys.* **2012**, *14*, 10774−10779.
(11) Aykol, M.; Kim, S.; Hegde, V. I.; Snydacker, D.; Lu, Z.; Hao, S.; Kirklin, S.; Morgan, D.; Wolverton, C. High-Throughput Computational Design of Cathode Coatings for Li-Ion Batteries. *Nat. Commun.* **2016**, *7*, 13779.
(12) Shi, S.; Gao, J.; Liu, Y.; Zhao, Y.; Wu, Q.; Ju, W.; Ouyang, C.; Xiao, R. Multi-Scale Computation Methods: Their Applications in Lithium-Ion Battery Research and Development. *Chin. Phys. B* **2016**, *25*, 018212.
(13) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational Screening of Perovskite Metal Oxides for Optimal Solar Light Capture. *Energy Environ. Sci.* **2012**, *5*, 5814−5819.
(14) Kuhar, K.; Crovetto, A.; Pandey, M.; Thygesen, K. S.; Seger, B.; Vesborg, P. C.; Hansen, O.; Chorkendorff, I.; Jacobsen, K. W. Sulfide Perovskites for Solar Energy Conversion Applications: Computational Screening and Synthesis of the Selected Compound LaYS3. *Energy Environ. Sci.* **2017**, *10*, 2579−2593.
(15) Ceder, G.; Chiang, Y.-M.; Sadoway, D.; Aydinol, M.; Jang, Y.-I.; Huang, B. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature* **1998**, *392*, 694−696.
(16) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191−201.
(17) Setyawan, W.; Gaume, R. M.; Lam, S.; Feigelson, R. S.; Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Comb. Sci.* **2011**, *13*, 382−390.
(18) Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295−2310.
(19) Perdew, J. P. Density functional theory and the band gap problem. *Int. J. Quantum Chem.* **2009**, *28*, 497−523.
(20) Seidl, A.; Görling, A.; Vogl, P.; Majewski, J. A.; Levy, M. Generalized Kohn-Sham schemes and the band-gap problem. *Phys. Rev. B* **1996**, *53*, 3764−3774.
(21) Jain, A.; Hautier, G.; Ong, S. P.; Moore, C. J.; Fischer, C. C.; Persson, K. A.; Ceder, G. Formation Enthalpies by Mixing GGA and GGA + U Calculations. *Phys. Rev. B* **2011**, *84*, 045115.
(22) Heyd, J.; Scuseria, G. E. Efficient hybrid density functional calculations in solids: Assessment of the Heyd−Scuseria−Ernzerhof screened Coulomb hybrid functional. *J. Chem. Phys.* **2004**, *121*, 1187−1192.
(23) Garza, A. J.; Scuseria, G. E. Predicting band gaps with hybrid density functionals. *J. Phys. Chem. Lett.* **2016**, *7*, 4165−4170.
(24) Gerosa, M.; Bottani, C. E.; Caramella, L.; Onida, G.; Di Valentin, C.; Pacchioni, G. Electronic structure and phase stability of oxide semiconductors: Performance of dielectric-dependent hybrid functional DFT, benchmarked against G W band structure calculations and experiments. *Phys. Rev. B* **2015**, *91*, 155201.
(25) Becke, A. D.; Johnson, E. R. A simple effective potential for exchange. *J. Chem. Phys.* **2006**, *124*, 221101.
(26) Tran, F.; Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Phys. Rev. Lett.* **2009**, *102*, 226401.
(27) Chan, M. K.; Ceder, G. Efficient band gap prediction for solids. *Phys. Rev. Lett.* **2010**, *105*, 196403.
(28) Koller, D.; Tran, F.; Blaha, P. Merits and limits of the modified Becke-Johnson exchange potential. *Phys. Rev. B* **2011**, *83*, 195134.
(29) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156−163.
(30) Himmetoglu, B.; Floris, A.; De Gironcoli, S.; Cococcioni, M. Hubbard-corrected DFT energy functionals: The LDA + U description of correlated systems. *Int. J. Quantum Chem.* **2014**, *114*, 14−49.
(31) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(32) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based on REpresentational State Transfer (REST) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209−215.
(33) Jain, A.; Hautier, G.; Ong, S. P.; Persson, K. New Opportunities for Materials Informatics: Resources and Data Mining Techniques for Uncovering Hidden Relationships. *J. Mater. Res.* **2016**, *31*, 977−994.
(34) Rajan, K. Materials Informatics. *Mater. Today* **2005**, *8*, 38−45.
(35) Wei, J.; Chu, X.; Sun, X.-Y.; Xu, K.; Deng, H.-X.; Chen, J.; Wei, Z.; Lei, M. Machine Learning in Materials Science. *InfoMat* **2019**, *1*, 338−358.
(36) Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Materiomics* **2017**, *3*, 159−177.
(37) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 83.
(38) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.
(39) Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *J. Chem. Phys.* **2011**, *134*, 074106.
(40) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.
(41) Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies With Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(42) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Muller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(43) Choudhary, K.; DeCost, B.; Tavazza, F. Machine Learning With Force-Field-Inspired Descriptors for Materials: Fast Screening and Mapping Energy Landscape. *Phys. Rev. Mater.* **2018**, *2*, 083801.
(44) Jung, S. G.; Jung, G.; Cole, J. M. Gradient boosted and statistical feature selection workflow for materials property predictions. *J. Chem. Phys.* **2023**, *159*, 194106.
(45) Jung, G.; Jung, S. G.; Cole, J. M. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chem. Sci.* **2023**, *14*, 3600−3609.
(46) Jung, S. G.; Jung, G.; Cole, J. M. *Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks*. submitted 2023.
(47) Gladkikh, V.; Kim, D. Y.; Hajibabaei, A.; Jana, A.; Myung, C. W.; Kim, K. S. Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties. *J. Phys. Chem. C* **2020**, *124*, 8905−8918.
(48) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375.
(49) Huang, Y.; Yu, C.; Chen, W.; Liu, Y.; Li, C.; Niu, C.; Wang, F.; Jia, Y. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *J. Mater. Chem. C* **2019**, *7*, 3238−3245.
(50) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **2016**, *93*, 115104.
(51) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(52) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679.
(53) Zhaochun, Z.; Ruiwu, P.; Nianyi, C. Artificial neural network prediction of the band gap and melting point of binary and ternary compound semiconductors. *Mater. Sci. Eng. B* **1998**, *54*, 149−152.
(54) Gu, T.; Lu, W.; Bao, X.; Chen, N. Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sci.* **2006**, *8*, 129−136.
(55) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **2014**, *83*, 185−195.
(56) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **2018**, *9*, 1668−1673.
(57) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **2020**, *6*, 138.
(58) Wang, A. Y.-T.; Kauwe, S. K.; Murdock, R. J.; Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *npj Comput. Mater.* **2021**, *7*, 77.
(59) De Breuck, P. P.; Evans, M. L.; Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys.: Condens. Matter* **2021**, *33*, 404002.
(60) De Breuck, P.-P.; Hautier, G.; Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Comput. Mater.* **2021**, *7*, 83.
(61) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.
(62) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(63) Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **2015**, *115*, 1094−1101.
(64) Kiselyova, N. N.; Dudarev, V. A.; Korzhuyev, M. A. Database on the bandgap of inorganic substances and materials. *Inorg. Mater. Appl. Res.* **2016**, *7*, 34−39.
(65) Strehlow, W. H.; Cook, E. L. Compilation of energy band gaps in elemental and binary compound semiconductors and insulators. *J. Phys. Chem. Ref. Data* **1973**, *2*, 163−200.
(66) Joshi, N. *Photoconductivity: Art: Science & Technology*; Marcel Dekker: New York, 1990.
(67) Madelung, O. *Semiconductors: Data Handbook*; Springer Science & Business Media, 2004.
(68) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(69) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(70) Deml, A. M.; O'Hayre, R.; Wolverton, C.; Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Phys. Rev. B* **2016**, *93*, 085142.
(71) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.
(72) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017; pp 4768−4777.
(73) Shishkin, M.; Kresse, G. Self-consistent GW calculations for semiconductors and insulators. *Phys. Rev. B* **2007**, *75*, 235102.
(74) Skelton, J. M.; Parker, S. C.; Togo, A.; Tanaka, I.; Walsh, A. Thermal physics of the lead chalcogenides PbS, PbSe, and PbTe from first principles. *Phys. Rev. B* **2014**, *89*, 205203.
(75) Yu, L.; Kokenyesi, R. S.; Keszler, D. A.; Zunger, A. Inverse Design of High Absorption Thin-Film Photovoltaic Materials. *Adv. Energy Mater.* **2013**, *3*, 43−48.
(76) Schimka, L.; Harl, J.; Kresse, G. Improved hybrid functional for solids: The HSEsol functional. *J. Chem. Phys.* **2011**, *134*, 024116.

https://doi.org/10.1021/acs.jcim.3c01897
J. Chem. Inf. Model. 2024, 64, 1187−1200