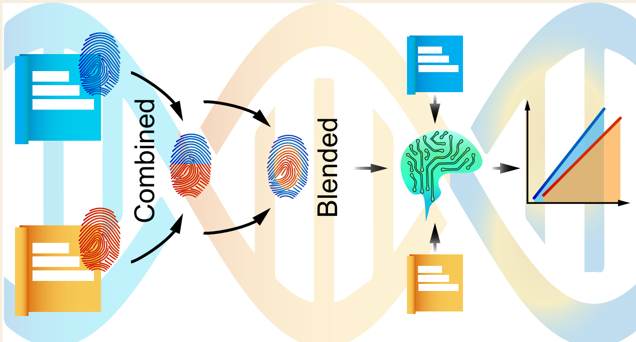
ACS PHYSICAL CHEMISTRY Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/physchemau Article

Feature Blending: An Approach toward Generalized Machine Learning Models for Property Prediction

Swanti Satsangi,‡ Avanish Mishra,‡ and Abhishek K. Singh*

Cite This: ACS Phys. Chem Au 2022, 2, 16−22 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: From studying the atomic structure and chemical behavior to the discovery of new materials and investigating properties of existing materials, machine learning (ML) has been employed in realms that are arduous to probe experimentally. While numerous highly accurate models, specifically for property prediction, have been reported in the literature, there has been a lack of a generalized framework. Herein we propose a novel feature selection approach that enables the development of a unified ML model for property prediction for several classes of materials. It involves an ingenious blending of selected features from various classes of data such that the resultant feature set equips the model with global data descriptors capturing both class-specific as well as global traits. We took accurate band gaps of three distinct classes of 2D materials as our target property to develop the proposed feature blending approach. Using Gaussian process regression (GPR) with the blended features, the ML model developed here resulted in an average root-mean-squared error of 0.12 eV for unseen data belonging to any of the participating classes. The feature blending approach proposed here can be extended to additional classes of materials and also to predict other properties.

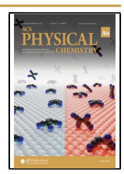
KEYWORDS: *2D materials, empirical model, Gaussian process regression, feature blending, bandgap, property prediction*

INTRODUCTION

The three paradigms of science namely empirical, theoretical, and computational have not only been contributing to and benefiting from each other through decades, but have also resulted in the generation of a huge amount of data over these years leading to a fourth paradigm. This fourth paradigm shift in materials science with computational methods leading to material discovery and property predictions is now driving the era of materials-informatics.[1,2] The workflow of material-informatics involves extraction of knowledge via data-driven machine learning (ML) methods from large amounts of unexplored computational and experimental data. ML has been utilized in the field of material science for various applications such as the discovery of new materials,[3−5] force-field generation,[6−10] microstructure analysis,[11−15] and property prediction.[16−30] Despite immense progress in the application of ML in material science, the applicability of all of these models thus far has been for a specific family/class of materials. Several attempts to develop multiclass generalized property prediction models surpassing the barrier of descriptor dependence have also been made. These include development of crystal graph convolutional neural networks,[31] universal graph networks,[32] SchNet,[9] and message passing neural network,[33,34] etc. Similarly, a general-purpose ML framework using a Random Forest method was proposed by Ward et al.[35] in which the data set was partitioned into sets of similar materials based on their composition, followed by the designing of several models for the various classes of materials.

Since ML-based predictions depend primarily on the availability of a pristine data set employed for training, the resultant models focused on a particular class of materials show poor transferability across different classes. Increasing the variability in the data results in high prediction errors compared to specific class data since the descriptors are incapable of describing global data trends. This insufficiency can only be compensated by increasing the amount of training data of any new incoming class, which is not always feasible. Therefore, to make predictions on a class of materials with insufficient training data, it would be highly desirable to have a pre-existing generalized ML model that has previously been trained on the desired class using a bigger data set of the same class, belonging, however, to a different database. The lack of such a generalized ML framework that works across data sets

Received: July 16, 2021
Published: September 17, 2021



© 2021 The Authors. Published by American Chemical Society

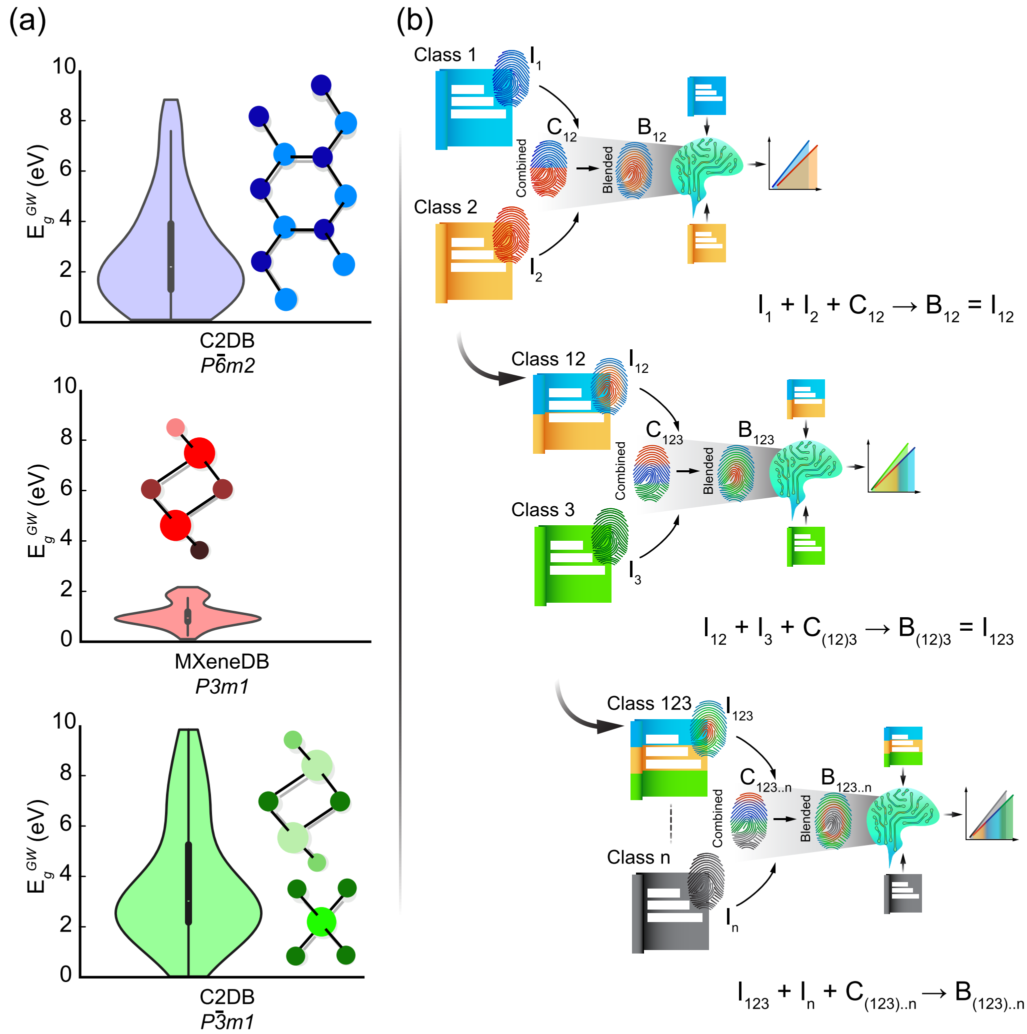


Figure 1. (a) Voilin plot showing the data distribution for model development with the GW band gap range, along with the template structure for different space groups. (b) Feature blending schematic: Stepwise selection and blending of various features to obtain global feature set for all participating classes. I_i denotes the individual feature for i^{th} class, C_{ij} denotes combined class features for i^{th} and j^{th} class and B_{ij} is the blended feature set for these classes.

poses a serious challenge in realizing the full potential of data-driven approaches and calls for a scheme that enables the development of generalized multiclass unified ML models.

In this work, we propose a feature blending approach to develop one such unified ML model generalized to unseen data from multiple participating classes of materials. A data set of 272 materials belonging to two different databases was utilized by splitting it into three different classes based on their spacegroups. The GW gap of these materials served as the target property. Blended features capture the patterns and trends elicited by the individual as well as mixed class data. The best GPR model using this feature set resulted in an rmse of 0.14/0.14 eV and R^2 of 0.99/0.99 for train/test sets and an rmse of 0.14/0.09/0.13 eV (average of 0.12 eV) and R^2 of 0.99/0.96/0.99 for validation sets belonging to the three considered classes $P\bar{6}m2$, $P3m1$ and $P\bar{3}m1$, respectively. The unprecedented accuracy can be attributed to the judicious selection of features using this approach, that represent each focused class and at the same time can also capture the complex details of the combined data set. Furthermore, some of the features belonging to the final blended set have been shown to display a universal empirical relationship with the target property that can be utilized to accelerate the estimation of the bandgap for materials with no training data.

METHODOLOGY

For developing the proposed generalized feature selection approach, structure, first-principle properties, and the property of interest (GW band gap) are extracted from open material databases.[29,36,37] Our initial feature set comprised DFT calculated properties[38] and standard deviation and mean of elemental properties,[39,40] resulting in 44 primary features. These properties, along with their corresponding symbols, have been listed in Table S1 in the Supporting Information. Other than the computed properties, elemental features (Table S1) are considered in the feature set due to their ease of availability and their role in structure formation (bonding). Selection of relevant features is performed using least absolute selection and shrinkage operator (LASSO) and neighborhood component analysis (NCA)(section 3 in Supporting Information).

Once the reduced set of features is obtained, to develop the prediction models, Gaussian process regression (GPR) is implemented along with automatic relevance determination (ARD) kernel. ARD kernel provides an opportunity to find the relevance of different features (proportion of contribution in the model) by selecting a different length scale for each. The relevance of features for the given output is obtained by calculating the inverse of the length scale for each feature. A larger length scale suggests smaller variation over the distribution of the function; hence, a smaller effect on prediction or lesser contribution and vice versa. GPR has been discussed further in the Supporting Information.

■ RESULTS AND DISCUSSION

The 2D-materials data set was collected from two online repositories, that is, the computational 2D-materials database (C2DB)[36,37] and aNANt.[41] C2DB hosts 2D-materials of numerous structural, thermodynamic and electronic properties from 30 different crystal classes, whose properties are calculated using DFT within Perdew−Burke−Ernzerhof (PBE) approximation.[36] This C2DB contains well converged many-body perturbation theory (GW) quasi-particle band gaps for ∼232 2D-materials. The majority of data belonged to classes $P\bar{6}m2$ and $P\bar{3}m1$ comprising 109 and 79 compounds, respectively. Around 44 other compounds belonging to space groups such as P1, *Pm*, *P4*/*nmm*, *Pma*2, *Pmmn*, *P*3*m*1, and *Pmn*21 were collected and only utilized for testing the generalizability of the derived empirical model. The GW band gap of C2DB spans from 1 to 10 eV for different classes. MXenes selected from the aNANt[41] database are a relatively new class of 2D-materials with chemical formula $M_{n+1}X_nT_2$ (n = 4−1), where M is early transition metals, X is either C or N, and T is functional groups attached to the top and bottom surface of MXene. This family of layered materials belongs to either space group $P3m1$ or $P\bar{3}m1$[42−44] and the GW band gap value ranges from 0 to 3 eV (shown in Figure 1a). In this data set, 84 compounds belonging to $P3m1$ space group are randomly selected. Thus, the final data set used here for ML comprised 272 compounds belonging to classes $P\bar{6}m2$, $P\bar{3}m1$, and $P3m1$.

The feature blending algorithm for the three classes was performed in three stages and can be understood pictorially from Figure 1b. Here, for any given class i with dataset D_i its individual class feature set derived after feature selection is represented by I_i. Similarly, C_{ij} denotes combined class features set obtained after applying feature selection on mixed data of ith and jth classes. B_{ij} on the other hand is the feature set which is obtained when three feature sets, individual sets I_i, I_j, and combined feature set (C_{ij}), are simply appended together and feature reduction is performed. Thus, $C_{ij} = F(D_i + D_j)$ while $B_{ij} = F(C_{ij} + I_i + I_j)$ where $F(.)$ represents feature selection. When a third class is to be included, B_{ij} is taken as individual class feature set I_{ij} for the first two classes and process of finding combined class and blended feature sets is replicated with a new class as the second class. This procedure can be generalized for any n number of classes.

Individual Class Models

At the first stage, each class was considered individually. The first data set to be utilized was $P\bar{6}m2$. Feature selection using LASSO followed by NCA resulted in eight *individual-class* features namely T_b^{mean}, C_g^{mean}, C_g^{std}, C_{mol}^{std}, H_e^{std}, T_m^{std}, r_{cov}^{std} and E_g^{PBE}. During the process of feature selection for any class, it was ensured that the features had a low correlation with each other and, at the same time, have moderate to high Pearson correlation with the response variable, E_g^{GW} (PCC (E_g^{GW})). For class $P\bar{6}m2$ this has been shown in Figure 2a and Figure 2c, respectively. From Figure 2d, it can be noticed that E_g^{PBE} is the most important feature in the model, also having the highest correlation with the target property. T_b^{mean} is another feature that has a high correlation with E_g^{GW}; however, it has moderate relevance in the model. C_{mol}^{std} has the least relevance for this GPR model.

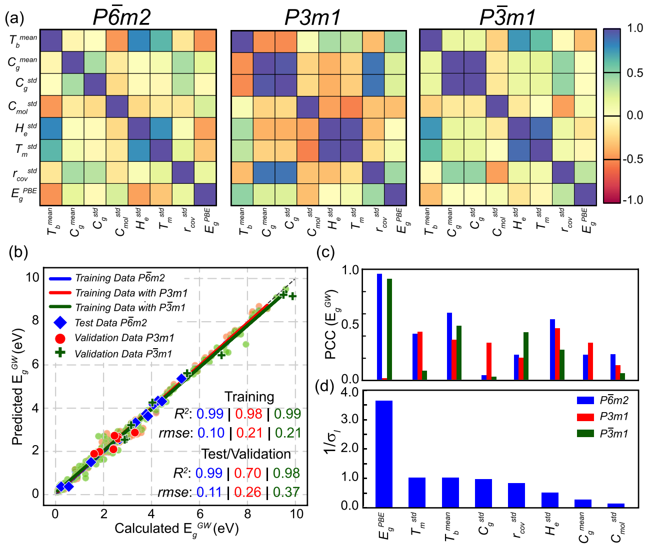


Figure 2. (a) Correlation heatmap for feature selected using $P\bar{6}m2$ data set with all three data sets; (b) scatter plot for the model developed using $P\bar{6}m2$ data set; (c) Pearson correlation coefficient of features with GW band gap (PCC (E_g^{GW})); and (d) feature importance (inverse of length scale).

After removing the 10% validation data, the remaining 90% data set was further split into a 90−10% ratio and used for the training-testing purpose. This process of splitting the train-test data was performed 2000 times, and a GPR model was built iteratively for each train-test combination by optimizing the kernel hyperparameters. The best-optimized hyperparameters were selected based on model performance on test data. This best model gave an rmse and R^2 of 0.10/0.11 eV and 0.99/0.99, respectively, for train/validation sets.

At this stage, the transferability of the model is ascertained, which is done by utilizing the optimized hyperparameters obtained for class $P\bar{6}m2$ and training with additional 90% data from the remaining two classes, one at a time, along with the original $P\bar{6}m2$ training data. On combining class $P3m1$ data while training, the model gave an rmse and R^2 of 0.21/0.26 eV and 0.98/0.70, respectively, for train/validation sets. Likewise, when trained with additional $P\bar{3}m1$ data, model resulted in an rmse and R^2 of 0.21/0.37 eV and 0.99/0.98, respectively, for train/validation sets. The performance of this model for all three classes has been shown in Figure 2b. Evidently, this model only performed well for its parent class.

Similarly, the features selection process was repeated for class $P3m1$, and the above-mentioned validation process was performed on the 10% validation data. The features obtained were P^{mean}, C_g^{mean}, H_e^{std}, χ_p^{std}, m^{std}, r_{atom}^{std}, κ_l^{std}, and E_g^{PBE}. Using these *individual-class* features, the rmse and R^2 were 0.08/0.08 eV and 0.96/0.94 for train and validation data, respectively. The correlation plot (Figure S2c) for these features with E_g^{GW} showed a low correlation between E_g^{PBE} and E_g^{GW}; however, the relevance of E_g^{PBE} in the model was quite significant (Figure S2d). Feature selection on the third class $P\bar{3}m1$ resulted in features namely, T_b^{mean}, EA^{mean}, ρ^{mean}, χ_p^{std} and E_g^{PBE}. The model

https://doi.org/10.1021/acsphyschemau.1c00017
ACS Phys. Chem Au 2022, 2, 16−22

for this class gave an rmse and R^2 of 0.10/0.10 eV and 0.99/0.99, respectively, on train/validation sets. Next, transferability of both $P3m1$ and $P\bar{3}m1$ individual class models was tested on the remaining two classes. The resultant performance has been displayed in Figure S3 in the Supporting Information and the class features have also been listed in Table S2 there. In Table S3, we can see that all of the above-developed models using *individual-class* features were highly efficient in making predictions for the parent class, however, they were not sufficient for predictions on another class despite the inclusion of data of new class while retraining the models. Thus, the localized behavior of the class features impedes generalization to other classes.

Two-Class Combined and Blended Models

At the second stage, we mixed the data of two classes and performed feature selection to obtain combined class features. The first combination comprised classes $P\bar{6}m2$ and $P3m1$ (Figure S4). The selected features of combined data set included H_e^{mean}, G^{mean}, r_{atom}^{mean}, κ_l^{mean}, T_b^{std}, EA^{std}, G^{std}, r_{atom}^{std}, r_{cov}^{std}, E_{CBM}, and E_g^{PBE}. The hyperparameters for the best-optimized model were obtained for this set of features, and the resultant model gave an rmse and R^2 of 0.17/0.17 eV and 0.98/0.97, respectively, for the train/test sets. When this model was tested on the validation data of the two classes, it gave an rmse and R^2 of 0.20/0.18 eV and 0.99/0.87 for $P\bar{6}m2/P3m1$ data as shown in Figure 3a. The prediction accuracy using these two-class

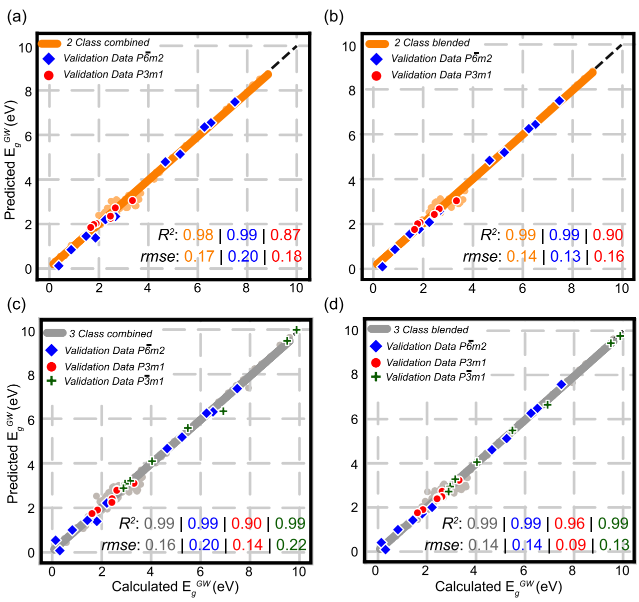


Figure 3. Scatter plots for 2 class (a) combined and (b) blended feature, and 3 class (c) combined and (d) blended features, respectively.

combined features was better than the accuracies obtained for validation using any of the individual-class features. A comparison of $P\bar{6}m2$ and $P3m1$ individual-class features, with these features revealed that other than a few common ones, the combined feature set comprised several additional features. The improvement in the performance of the model for both classes can be attributed to these additional descriptors that provided a better representation of the latent behavior of this mixed data.

To harness the capabilities of both individual and combined class features, feature blending for the two classes was performed. For the two classes $P\bar{6}m2$ and $P3m1$, when their individual features were simply appended to their combined feature set, 23 features were obtained, which were reduced using LASSO and NCA. The new set termed as *two-class blended* feature set consisted of 10 features including E_g^{PBE}, χ_p^{std}, κ_l^{std}, r_{cov}^{std}, C_g^{std}, H_e^{std}, T_b^{mean}, G^{mean}, r_{atom}^{mean}, and T_b^{std}. The best model using these blended features gave an rmse and R^2 of 0.14/0.14 eV and 0.99/0.98 for combined train/test sets, respectively. Further, on the 10% validation data, the rmse and R^2 were 0.13/0.16 eV and 0.99/0.90 for $P\bar{6}m2/P3m1$ data as shown in Figure 3b. This model with blended features gave a better accuracy than the models using individual or combined class features, thus improving generalizability. Two-class combined and blended features were next derived for pairs $P\bar{6}m2$ and $P\bar{3}m1$, as well as $P3m1$ and $P\bar{3}m1$, and similar observations were made. The features and the results for these combinations have been tabulated in Tables S2 and S3 and displayed in Figure S4. The observation that generalizability has been achieved with improvement in prediction accuracy at the same time, can be credited to the fact that this feature set reflects a balanced blend of both localized and group behavior.

Three-Class Combined and Blended Models

The feature blending scheme was next extended to all three classes, that is, $P\bar{6}m2$, $P3m1$, and $P\bar{3}m1$. For this, we begin by generating the combined class features for all three class mixed data as done in the case of two classes. A set of 13 three-class combined features were obtained. This combined feature set consisted of H_e^{mean}, IE_1^{mean}, χ_A^{mean}, T_m^{mean}, κ_l^{mean}, C_g^{std}, EA^{std}, IE_1^{std}, m^{std}, r_{atom}^{std}, r_{vdw}^{std}, E_{VBM} and E_g^{PBE}. The best model using these combined features gave an rmse of 0.16/0.16 eV and R^2 of 0.99/0.99 for combined train/test data. For validation data, rmse of 0.20/0.14/0.22 eV and R^2 of 0.99/0.90/0.99 was obtained for classes $P\bar{6}m2/P3m1/P\bar{3}m1$, respectively, as shown in Figure 3c. This again was a significant improvement compared to the performance of individual class models on validation sets.

For obtaining blended features for any number of classes, all we need is two sets of individiual class features and a set of their combined features. Thus, a remarkable facet of the feature blending approach is that it can always be reduced to a two-class problem whenever a new class needs to be included. For the three-class scenerio, this was achieved by utilizing the 10 two-class blended features obtained in stage two as the first individual feature set (for $P\bar{6}m2$ and $P3m1$ combined data) along with the seven individual-class features of the new incoming class $P\bar{3}m1$ (obtained at stage one) as second individual feature set and finally the 13 three-class combined features. Feature reduction on this set gave eight features, namely, χ_P^{std}, T_b^{mean}, r_{cov}^{std}, E_g^{PBE}, G^{mean}, r_{atom}^{mean}, H_e^{mean}, and κ_l^{mean}. The unified model designed using these three-class blended features gave an rmse of 0.14/0.14 eV and R^2 of 0.99/0.99 for combined train/test data, respectively, whereas an unprecedented rmse of 0.14/0.09/0.13 eV and R^2 0.99/0.96/0.99 were obtained for 10% validation data of $P\bar{6}m2/P3m1/P\bar{3}m1$, respectively, as shown in Figure 3d.

Similar calculations were performed by changing the sequence in which the classes are selected for inclusion. When we begin with $P\bar{6}m2$ and $P\bar{3}m1$ and consider $P3m1$ as the third incoming class, the three-class blended feature set obtained comprised T_b^{mean}, H_e^{mean}, κ_l^{mean}, EA^{std}, IE_1^{std}, χ_P^{std}, m^{std}, E_{VBM}, and E_g^{PBE}. The model using these features gave an rmse of 0.12/0.14 eV and R^2 of 0.99/0.99 for combined train/test data, whereas an rmse of 0.13/0.14/0.12 eV and R^2 0.99/0.91/0.99 were obtained for 10% validation data of $P\bar{6}m2/P3m1/P\bar{3}m1$, respectively. With $P\bar{3}m1$ and $P3m1$ as the first two classes and $P\bar{6}m2$ as the third class, the feature set $\mathrm{T_b^{mean}}$, $\mathrm{H_e^{mean}}$, $\mathrm{IE_1^{mean}}$, κ_1^{mean}, $\mathrm{C_g^{std}}$, $\chi_\mathrm{P}^{\mathrm{std}}$, $\mathrm{r_{cov}^{std}}$, and $\mathrm{E_g^{PBE}}$ resulted in an rmse of 0.14/0.15 eV and R^2 of 0.99/0.99 for combined train/test data and an rmse of 0.14/0.15/0.14 eV and R^2 of 0.99/0.90/0.99 for 10% validation data of $P\bar{6}m2/P3m1/P\bar{3}m1$, respectively. It was interesting to note that the order in which the classes were included for blending, did not have a major impact on the performance of the final unified three-class blended model. This was due to the fact that all three blended feature sets had majority features namely $\mathrm{T_b^{mean}}$, $\mathrm{H_e^{mean}}$, κ_1^{mean}, $\chi_\mathrm{P}^{\mathrm{std}}$, and $\mathrm{E_g^{PBE}}$ in common. The performance of the above-discussed models has been shown in Figure S5.

Two intriguing observations emerged as a result of the feature blending process. First, an analysis of the individual, combined, and blended models showed that the relevance of each feature of the blended feature set in the unified model, whether belonging to an individual or the combined feature set, was almost the same as its relevance in its parent model, that is, individual or combined class model. This was true for two as well as three-class models. This relevance or importance of the selected features in the regression model for the first set of three-class blended features, shown in Figure S6(a,b), can be seen to be comparable to those in their parent model. This conservation of relevance of each feature in the unified model plays a significant role in the consistent performance of blended features for all its participating classes.

Second, apart from promoting generalizability in prediction models, the blending process emanated a few features that recurred in almost all feature sets. Some of these features not only had high relevance in the statistical model, but they also displayed a notable relationship with each other and with the $\mathrm{E_g^{GW}}$. Features, namely $\chi_\mathrm{P}^{\mathrm{std}}$, $\mathrm{T_b^{mean}}$, and $\mathrm{E_g^{PBE}}$, present in all the three final three-class blended feature sets, were found to not only have high Pearson correlations with $\mathrm{E_g^{GW}}$ but also captured the right physics in the trends in the band gap. From Figure 4a two conclusions were drawn: first, increase/decrease in bandgap at PBE levels shows a similar shift at the GW level, and second, $\mathrm{E_g^{GW}}$ varies with $\chi_\mathrm{P}^{\mathrm{std}}$ and $\mathrm{T_b^{mean}}$, that is, increase in the standard deviation of elemental electronegativity, and subsequent decrease in the mean elemental boiling temperature leads to increase in the bandgap. The increase in $\chi_\mathrm{P}^{\mathrm{std}}$ indicates a broader spread (larger difference) in the electronegativity value from the mean, hence, strong interaction among the atoms, such firm overlap causing a larger splitting and consequently a wider bandgap. Besides, the boiling temperature of elements with larger electronegativity (>2.6) is comparably low (<1000 K); however, for elements with smaller χ_P (<2.6), the boiling temperature varies (up to 5000 K). Therefore, materials with a larger $\chi_\mathrm{P}^{\mathrm{std}}$ usually have lower $\mathrm{T_b^{mean}}$. Nevertheless, depending upon the constituent elements, the mean boiling temperature for a system could be higher with the larger value of $\chi_\mathrm{P}^{\mathrm{std}}$. For example, a material with the combination of C and F would have both larger $\chi_\mathrm{P}^{\mathrm{std}}$ and $\mathrm{T_b^{mean}}$. Furthermore, other pairwise distribution plots of the individual class feature provide further physical insights between different features of the particular class. Additional discussion on this is included in the Supporting Information and pictorially explained in Figure S7.

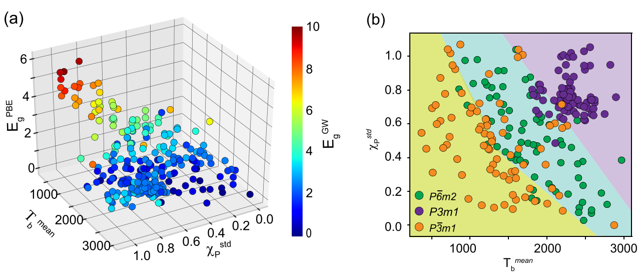


Figure 4. (a) The three-dimensional color map showing the variation of $\mathrm{E_g^{PBE}}$ with $\chi_\mathrm{P}{}^{\mathrm{std}}$ and $\mathrm{T_b^{mean}}$, and variation in their $\mathrm{E_g^{GW}}$, shown by colorbar. (b) Pair distribution plot between $\chi_\mathrm{P}^{\mathrm{std}}$ and $\mathrm{T_b^{mean}}$, with colored zone for different class of materials.

Moreover, the relationship between $\chi_\mathrm{P}^{\mathrm{std}}$ and $\mathrm{T_b^{mean}}$ was also found to partition different classes of materials and can thus be exploited for expediting selection of 2D materials in general, having bandgap in the desired range as shown in Figure 4b. $P3m1$ lies at the top right with large electronegativity and boiling temperatures, $P\bar{6}m2$ lies in the middle, followed by $P\bar{3}m1$ at the bottom of the plot. The interpretation of these three relationships provided an opportunity to model an empirical relation between them for estimating the GW band gap. This relation was derived by generating a set of compound features by applying various mathematical operations on these three selected features and correlating them with the calculated GW bandgap.[26] Some of the mathematical operations used here were x^2, x^3, $1/x$, log (x), exp (x), etc., where x is any feature. As the dimensionality increases, x will include more than one feature. Combination of these mathematical operations on these three features gave an empirical formula represented as

$$E_g^{\mathrm{estimated}} = 18.5 \times \left(\frac{\sqrt{\chi_\mathrm{P}^{\mathrm{std}}} \times E_g^{\mathrm{PBE}}}{\log(T_b^{\mathrm{mean}}) + E_g^{\mathrm{PBE}}} \right) + 0.82 \tag{1}$$

The estimated value $\mathrm{E_g^{estimated}}$ using this empirical formula showed a strong correlation of 0.93 with $\mathrm{E_g^{GW}}$ and the fitted linear equation results in a high coefficient of correlation R^2 of 0.83, as shown in Figure S8a. This equation was also utilized for estimating bandgaps of the 44 compounds belonging to space groups $P1$, Pm, $P4/nmm$, $Pma2$, $Pmmn$, $Pmn21$, and $P3m1$ from the CMR database that were neither sufficient to build an individual class model nor for feature blending. As shown in Figure S8b, the small residuals obtained exhibit the potential of this relationship in making predictions on unseen classes and corroborate the significance of the selection of the right features.

■ CONCLUSION

In summary, we attempted to extend the transferability of any pretrained machine learning model across different databases by introducing the concept of feature blending. Feature blending is shown to enable the selection of features with global relevance in contrast to local scope offered by individual class features, by iteratively mixing and selecting features obtained for various classes at different stages. Thorough validation of models at different stages of development has been performed and significance of judicious selection of features for a general-purpose machine learning framework has been instantiated. The reported unified model developed for bandgap prediction, has been shown to perform on previously unseen data from all the participating classes with comparable accuracy. The most prominent outcome of the scheme was identification of certain universal features which exhibited vital influence on the band gap of all the participating classes in a similar manner. A detailed study of physical relationships between these cross cutting features made it possible to develop a generic empirical relation for the approximation of E_g^{GW} for 2D materials. This developed equation using the three important features was found to be applicable on all 9 classes of 2D materials available in the initial data set with reasonable accuracy. Although classes in this work were formed based on space groups, classification can be done based on any other parameter. The applicability of the proposed feature blending approach can be seemingly extended to large classes of materials.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsphyschemau.1c00017.

> Additional information including features selection techniques, machine learning methods, feature selection using $P3m1$, $P\bar{3}m1$, and two/three-class combined and blended features. Performance test for two-class, mixed class, feature relevance, empirical relation and physical significance of features. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Abhishek K. Singh − *Materials Research Centre, Indian Institute of Science, Bangalore 560012, India;* orcid.org/0000-0002-7631-6744; Email: abhishek@iisc.ac.in

Authors

Swanti Satsangi − *Materials Research Centre, Indian Institute of Science, Bangalore 560012, India*

Avanish Mishra − *Materials Research Centre, Indian Institute of Science, Bangalore 560012, India;* orcid.org/0000-0003-3997-0445

Complete contact information is available at: https://pubs.acs.org/10.1021/acsphyschemau.1c00017

Author Contributions

‡S.S. and A.M. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge Indian Institute of Science for providing the computing facilities of Materials Research Centre, Institute of Eminence (IoE) MHRD grant, Thematic Unit of Excellence and, Supercomputer Education and Research Centre. A.M. acknowledges UGC India for a Senior Research Fellowship. S.S. and A.M. thank the DST Nano Mission and DST Korea for their for support.

■ REFERENCES

(1) Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Mater.* **2016**, *4*, 053208.

(2) Schleder, G. R.; Padilha, A. C.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science−a review. *Journal of Physics: Materials* **2019**, *2*, 032001.

(3) Bialon, A. F.; Hammerschmidt, T.; Drautz, R. Three-Parameter Crystal-Structure Prediction for *sp-d*-valent Compounds. *Chem. Mater.* **2016**, *28*, 2550−2556.

(4) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery using Failed Experiments. *Nature* **2016**, *533*, 73−76.

(5) Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016**, *7*, 11241.

(6) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A Universal Strategy for the Creation of Machine Learning-Based Atomistic Force Fields. *NPJ. Comput. Mater.* **2017**, *3*, 37.

(7) Li, Y.; Li, H.; Pickard, F. C.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492−4503.

(8) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.

(9) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet−A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(10) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9*, 3887.

(11) Li, X.; Zhang, Y.; Zhao, H.; Burkhart, C.; Brinson, L. C.; Chen, W. A Transfer Learning Approach for Microstructure Reconstruction and Structure-Property Predictions. *Sci. Rep.* **2018**, *8*, 1−13.

(12) Exl, L.; Fischbacher, J.; Kovacs, A.; Oezelt, H.; Gusenbauer, M.; Yokota, K.; Shoji, T.; Hrkac, G.; Schrefl, T. Magnetic Microstructure Machine Learning Analysis. *J. Phys: Mater.* **2018**, *2*, 014001.

(13) Liu, R.; Kumar, A.; Chen, Z.; Agrawal, A.; Sundararaghavan, V.; Choudhary, A. A Predictive Machine Learning Approach for Microstructure Optimization and Materials Design. *Sci. Rep.* **2015**, *5*, 11551.

(14) Khatavkar, N.; Swetlana, S.; Singh, A. K. Accelerated prediction of Vickers hardness of Co-and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning. *Acta Mater.* **2020**, *196*, 295−303.

(15) Swetlana, S.; Khatavkar, N.; Singh, A. K. Development of Vickers hardness prediction models via microstructural analysis and machine learning. *J. Mater. Sci.* **2020**, *55*, 15845−15856.

(16) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279−4283.

(17) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions using Machine Learning. *Sci. Rep.* **2013**, *3*, 2810.

(18) 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: Condens. Matter Mater. Phys.* **2016**, *93*, 115104.

(19) 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.

(20) Juneja, R.; Yumnam, G.; Satsangi, S.; Singh, A. K. Coupling the High-Throughput Property Map to Machine Learning for Predicting Lattice Thermal Conductivity. *Chem. Mater.* **2019**, *31*, 5145−5151.

(21) Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218−226.

(22) Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A Distributed Materials Properties Repository from High-Throughput Ab Initio Calculations. *Comput. Mater. Sci.* **2012**, *58*, 227−235.

(23) 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.

(24) Mishra, A.; Satsangi, S.; Rajan, A. C.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Accelerated data-driven accurate positioning of the band edges of MXenes. *J. Phys. Chem. Lett.* **2019**, *10*, 780−785.

(25) 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.

(26) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B.; Ramprasad, R.; Gubernatis, J.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(27) Juneja, R.; Singh, A. K. Guided Patchwork Kriging to Develop Highly Transferable Thermal Conductivity Prediction Models. *J. Phys.: Mater.* **2020**, *3*, 024006.

(28) Juneja, R.; Singh, A. K. Unraveling the role of bonding chemistry in connecting electronic and thermal transport by machine learning. *J. Mater. Chem. A* **2020**, *8*, 8716−8721.

(29) Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Machine-Learning Assisted Accurate Band Gap Predictions of Functionalized MXene. *Chem. Mater.* **2018**, *30*, 4031−4038.

(30) Mukherjee, M.; Satsangi, S.; Singh, A. K. A Statistical Approach for the Rapid Prediction of Electron Relaxation Time Using Elemental Representatives. *Chem. Mater.* **2020**, *32*, 6507−6514.

(31) 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.

(32) 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.

(33) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *Proc. 34th Intl. Conf. Mach. Learn.* **2017**, 1263−1272.

(34) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.

(35) 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.

(36) Haastrup, S.; Strange, M.; Pandey, M.; Deilmann, T.; Schmidt, P. S.; Hinsche, N. F.; Gjerding, M. N.; Torelli, D.; Larsen, P. M.; Riis-Jensen, A. C.; Gath, J.; Jacobsen, K. W.; Mortensen, J. J.; Olsen, T.; Thygesen, K. S. The Computational 2D Materials Database: High-Throughput Modeling and Discovery of Atomically Thin Crystals. *2D Mater.* **2018**, *5*, 042002.

(37) Gjerding, M. N.; et al. Recent progress of the computational 2D materials database (C2DB). *2D Mater.* **2021**, *8*, 044002.

(38) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 144110.

(39) Haynes, W. M. *CRC Handbook of Chemistry and Physics*, 94th ed.; CRC Press: Boca Raton, FL, 2014.

(40) Dynamic Periodic Table. https://ptable.com/ (accessed 2020/07/21).

(41) aNANt: A Functional Materials Database. http://anant.mrc.iisc.ac.in (accessed on 21/07/2020).

(42) Srivastava, P.; Mishra, A.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Mechanistic insight into the chemical exfoliation and functionalization of Ti_3C_2 MXene. *ACS Appl. Mater. Interfaces* **2016**, *8*, 24256−24264.

(43) Chandrasekaran, A.; Mishra, A.; Singh, A. K. Ferroelectricity, Antiferroelectricity, and Ultrathin 2D Electron/Hole Gas in Multifunctional Monolayer MXene. *Nano Lett.* **2017**, *17*, 3290−3296.

(44) Khazaei, M.; Mishra, A.; Venkataramanan, N. S.; Singh, A. K.; Yunoki, S. Recent advances in MXenes: From fundamentals to applications. *Curr. Opin. Solid State Mater. Sci.* **2019**, *23*, 164−178.