CHEMISTRY OF MATERIALS
Cite This: Chem. Mater. 2018, 30, 4031−4038
Article
pubs.acs.org/cm

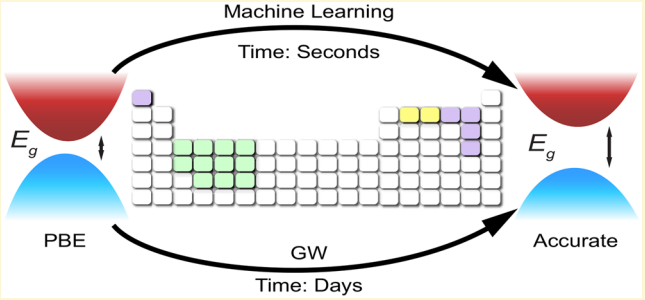
Machine-Learning-Assisted Accurate Band Gap Predictions of Functionalized MXene

Arunkumar Chitteth Rajan,[†,§] Avanish Mishra,[†,§] Swanti Satsangi,[†] Rishabh Vaish,[†] Hiroshi Mizuseki,[‡] Kwang-Ryeol Lee,[‡] and Abhishek K. Singh*[,†]

[†]Materials Research Centre, Indian Institute of Science, Bangalore 560012, India
[‡]Computational Science Research Center, Korea Institute of Science and Technology (KIST), Seoul 02792, Republic of Korea

Supporting Information

ABSTRACT: MXenes are two-dimensional (2D) transition metal carbides and nitrides, and are invariably metallic in pristine form. While spontaneous passivation of their reactive bare surfaces lends unprecedented functionalities, consequently a many-folds increase in number of possible functionalized MXene makes their characterization difficult. Here, we study the electronic properties of this vast class of materials by accurately estimating the band gaps using statistical learning. Using easily available properties of the MXene, namely, boiling and melting points, atomic radii, phases, bond lengths, *etc.*, as input features, models were developed using kernel ridge (KRR), support vector (SVR), Gaussian process (GPR), and bootstrap aggregating regression algorithms. Among these, the GPR model predicts the band gap with lowest root-mean-squared error (rmse) of 0.14 eV, within seconds. Most importantly, these models do not involve the Perdew−Burke−Ernzerhof (PBE) band gap as a feature. Our results demonstrate that machine-learning models can bypass the band gap underestimation problem of local and semilocal functionals used in density functional theory (DFT) calculations, without subsequent correction using the time-consuming GW approach.



Layered 2D materials have the potential to revolutionize modern-day energy, sensing, electronic, and optical devices. They span a larger chemical space, which is getting wider with the inclusion of newer materials on a regular basis. Among these, the most significant addition to the 2D family is an inorganic class of materials named MXene.[1−6] MXenes ($M_{n+1}X_n$; M, group IIIB to VIB; X, {C, N}; n, 1−3) are early transition metal carbides and/or nitrides and are chemically exfoliated from corresponding MAX phases.[6−8] Pristine MXenes are reactive due to existence of surface charges, which spontaneously get passivated ($M_{n+1}X_nT_2$) by functional groups (T). Combinations of transition metals, carbon/nitrogen atom, and a large number of functional groups (T: F, O, OH, *etc.*) as constituents yield tens-of-thousands of MXenes in the chemical compound space, which may have beneficial properties for optical, electronic, energy storage, and photocatalytic applications.[9−14]

Characterizing the properties of these materials by experiments or computation can be a time-consuming process. In the search of MXenes for electronic, optical, and catalytic applications, it is essential to have knowledge about their accurate band gaps, which involves complex calculations and may take several years to complete. With the advent of faster computers, density-functional-theory-based (DFT-based)[15] methods can be employed to successfully calculate band gap in a reasonable amount of time. Nonetheless, fundamental gaps computed by the local-density or generalized-gradient approximation (LDA or GGA) are underestimated. Using the many-body-perturbation-theory-based GW approach,[16] band gap underestimation can be corrected. However, such methods[16,17] are very time-consuming, and it becomes prohibitively expensive to compute accurate band gaps of these materials. Recently, statistical learning has emerged as a promising tool for predicting the structures[18−20] and properties[21−28] of various classes of materials. These methods can accurately predict the properties such as cohesive energy, lattice thermal conductivity,[24] band gaps,[22,23,25] entropy, free energy,[26] and heat capacity[25] within reasonable time.

Herein, by utilizing easily computed properties and intuitive information from chemical repositories as features, statistical learning models have been built to predict accurate band gaps of the MXene. A database consisting of 23 870 MXenes is generated, and the ground state structure calculations are performed on randomly selected 30% of the MXenes. We developed a metal−semiconductor classification model, having accuracy of 94%, which filters out finite gap MXenes. Among these, 76 MXenes are selected to build a prediction model by correlating the features to calculated GW band gaps. Feature regularization and reduction are performed by using *least absolute shrinkage and selection operator* (Lasso), which reduces

Received: February 14, 2018
Revised: May 31, 2018
Published: May 31, 2018

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

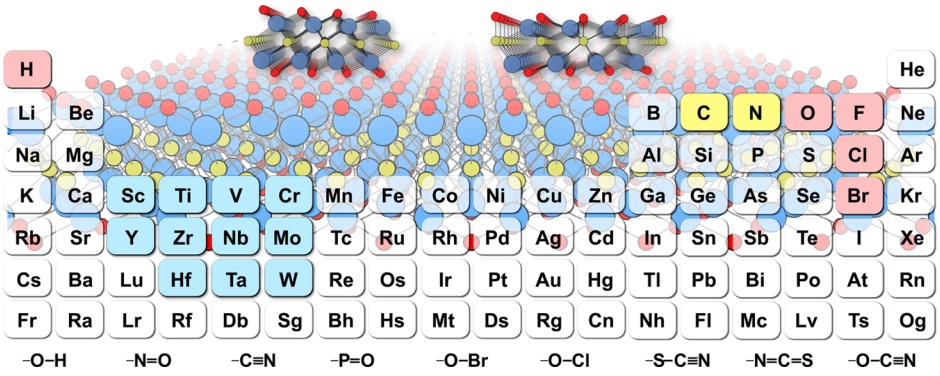


Figure 1. MXene composition. With 11 early transition metals M (blue) of IIIB to VIB groups; X, {C, N} (colored in yellow); and 14 surface functional groups T/T′ (red), a pool of 23 870 MXenes is generated. T/T′ consists of functionalization with elements {H, F, Cl, Br, O} and groups {CN, NO, PO, OH, OCl, OBr, OCN, SCN, NCS} (shown at the bottom). Two prominent phases[7] of MXene, namely, bb′ and cb, are shown at the top left and right, respectively.

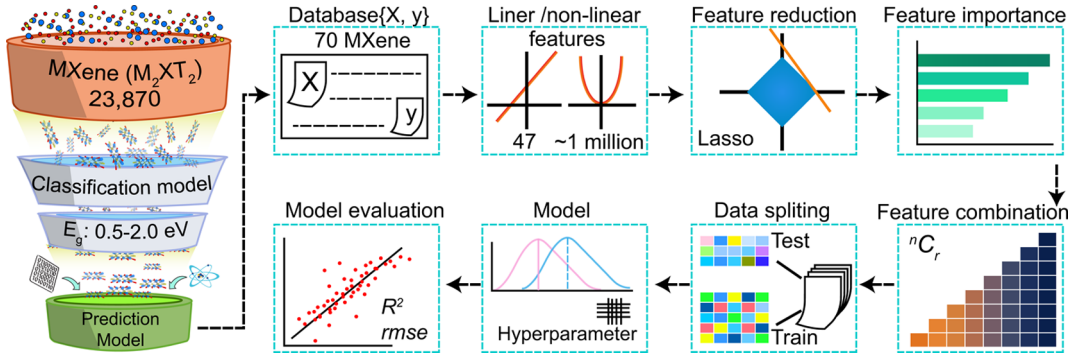


Figure 2. Work flow. Schematic of filtering semiconductors from the MXene database and feeding its subset as input to the machine-learning algorithm. The learning procedure involved training data set of {*X*, *y*} (*X*, features; and *y*, response function) and performing a highly accurate band gap prediction.

the number of features from 47 to 15. This primary features set is further extended to approximately ∼1 million nonlinear compound features by simple mathematical operations, which are then reduced to 8 by the application of Lasso. Depending upon feature-combinations, kernel ridge (KRR), support vector (SVR), Gaussian process (GPR), and bootstrap aggregating (bagging) methods resulted in optimized models yielding low root-mean-squared error (rmse) values of 0.14−0.20 eV. Among these, unprecedented low rmse values of 0.14 and 0.16 eV are achieved by the GPR and bagging models with combinations of 8 and 7 primary features, respectively. Notably, these models do not utilize PBE-level band gap and band positions as features. Hence, by applying these models, band gap with GW-level accuracy can be achieved for the entire MXene database within seconds, which otherwise would take many years.

Success of statistical learning depends upon access to high-quality data, which is essential for developing a model having excellent predictive capability. Inputs to the learning model termed as features or descriptors should be easily accessible properties, obtained by experiments and simple computations.[24,26,29−32] In general, a feature set comprises chemically important elemental and computed properties including boiling and melting points, group number, atomic radii, phases, bond lengths, *etc*. Outcome or the target property onto which the feature space is projected is obtained from state of the art experiments or theoretical calculations.

As a first step, the MXene database is constructed. The MXene structure (MM′XTT′) is composed of a unit cell having five internal layers arranged as T−M−X−M′−T′. Surfaces of MXene can either be functionalized by same T on both sides or different (T/T′) depending upon the chemical environment. In this study, the selected T/T′ are the elements {H, F, Cl, Br, O} and groups {CN, NO, PO, OH, OCl, OBr, OCN, SCN, NCS}. These T/T′ have maximum electronegativity difference with respect to M/M′. With all possible combinations of M/M′, X, and T/T′ (Figure 1), a total of 23 870 distinct MXenes are generated.[33]

Depending on the position of T/T′, the MXene can exist in two prominent structural phases, namely, bb′ and cb (discussed in the Supporting Information; Figure S1).[7] Structure optimization and band gap calculation are performed by DFT on 7200 randomly selected MXenes (∼30% of 23 870) using the Vienna *ab initio* simulation package (VASP).[34,35] Electronic exchange correlations are treated by the Perdew−Burke−Ernzerhof (PBE) parametrization of GGA, and core−valence electron interactions are represented by projector augmented wave (PAW) potentials. Plane wave basis sets with 500 eV energy cutoff are used to describe the electronic wave function. Brillouin zone is integrated using 15 × 15 × 1 Monkhorst−Pack (MP) *k*-grid. A vacuum of 30 Å (*z*-direction) is added in the unit cell to reduce any spurious interaction with its periodic images. The structure optimization is performed using conjugate gradient method until the components of the Hellman−Feynman forces on every atom are ≤0.005 eV/Å.

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

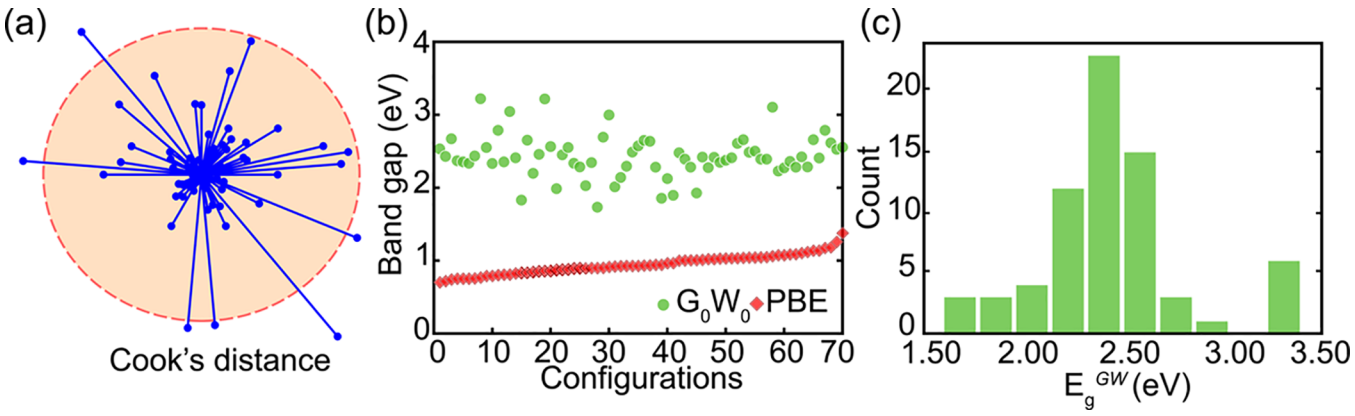


Figure 3. Semiconducting MXene. (a) Outlier analysis is performed on 76 MXenes by calculating their Cook's distance, which caused removal of 6 of them as influential outliers. Remaining 70 MXenes having (b) increasing PBE band gap (colored in red) are shown with corresponding gaps at computationally expensive G_0W_0 level (green). (c) The distribution of G_0W_0 band gaps corresponding to the data set for machine learning is also shown.

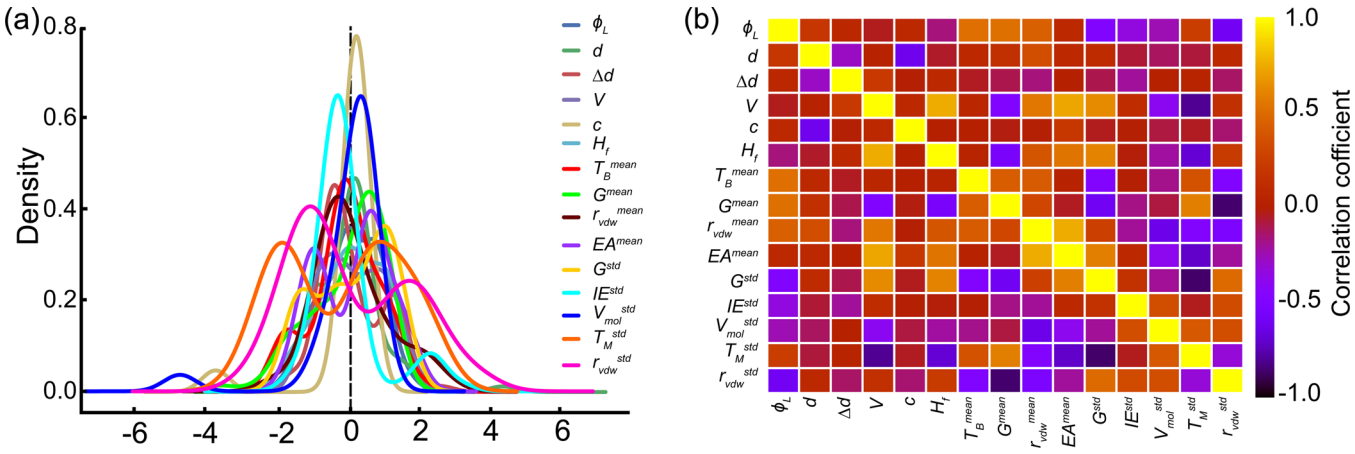


Figure 4. Density and correlation plots. (a) Density plot for the standardized data set with zero mean and unit variance in the preprocessing for machine learning. (b) Statistical heat map showing correlation of individual primary features with themselves.

Optimized 7200 MXenes are categorized into metals or semiconductors based on a classification model (see details in the Supporting Information; Figure S2). Since PBE underestimates the band gap, MXenes having PBE band gaps in a broader range (0.5−2 eV) are selected with an expectation that the corresponding GW band gap would span an adequate range, promising for electronic, optoelectronic, and photocatalytic applications. This reduces the number of MXenes further, and most of them are Sc,Y-based MXene. Calculated band gaps at the PBE level are underestimated, and finding GW-level accurate values for all the MXenes is time-consuming. Hence, for accelerated and accurate band gap prediction, a statistical learning-based approach is applied. Accordingly, 76 Sc,Y-based MXenes are randomly selected, and accurate GW band gaps for these are calculated, which are the target property for the machine-learning model.

Standard GW calculations on the selected MXenes are carried out within many-body perturbation theory using non-self-consistent GW approximation (G_0W_0)[36] as implemented in VASP. Here, G_0 is the electron's Green's function, and W_0 represents the screened Coulomb interaction. Converged quasi particle (QP) energies are calculated by varying the number of empty bands (depending on number of electrons in unit cell), frequency grids, cutoff energy, and k-grid. This included 900−1200 empty bands, 60 frequency grids, 400 eV energy cutoff, and 15 × 15 × 1 MP k-grid.

The learning model (see schematic Figure 2) is developed for the given band gap data set $\{\boldsymbol{X}, y\}$, which maps input feature set $\boldsymbol{X}$ to the targeted GW band gap, y. Here, $\boldsymbol{X} = \{x_i\}_{i=1}^{n}$ consists of n individual features (n = 47). These *primary* features are computed properties of the MXene, and mean and standard deviation (std) of elemental properties[37] (Tables S1 and S3). Due to large variation in the values of different features, data preprocessing is performed yielding standardized $\boldsymbol{X}$ with 0 mean and unit variance. Any physically meaningful data point with large residual or high leverage is a statistically influential outlier, which can adversely affect the model fitting. Hence, removal of such outliers is required (Figure S3), which is done by calculating the Cook's distance (see the Supporting Information). Figure 3a shows the Cook's distances for 76 MXenes. Data points lying out of the threshold (red circle) are removed, which reduces the data set to 70 (Figure 3b,c).

To have a faster and efficient model, a lesser number of features is preferred, and hence we search for a smaller set of relevant features. Feature reduction is performed by using Lasso,[38] which is a regression technique used for regularization and feature selection. It involves minimization of $L2$ norm (sum of square of differences) with regularization of $L1$ norm (sum of mean-absolute differences) on the coefficients (β) given by

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

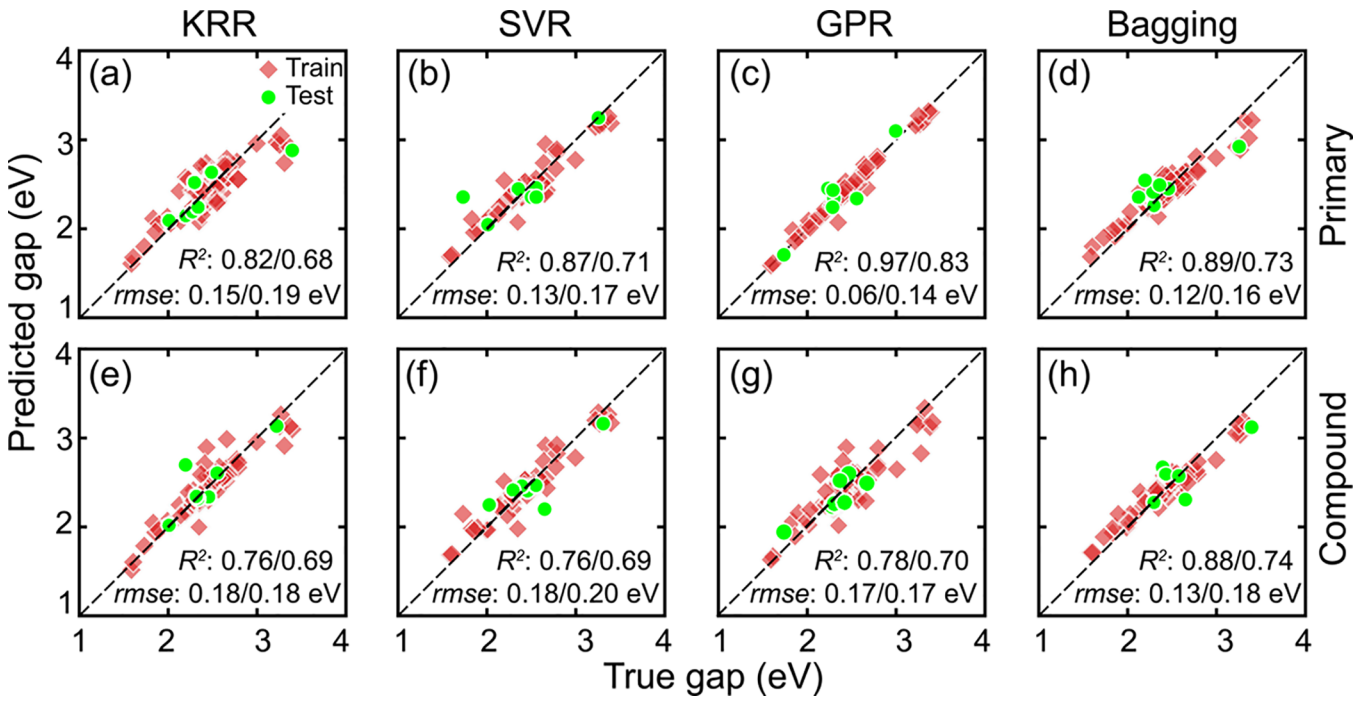


Figure 5. Band gap predictions of MXene. Scatter plots showing band gap predictions versus true (i.e., GW) gaps of important primary (top panel) and compound (bottom panel) feature-combinations. Best model predictions are shown with accuracy metrics such as R^2 and rmse (train/test), of (a, e) KRR, (b, f) SVR, (c, g) GPR, and (d, h) bagging corresponding to 90% training (colored red) and 10% testing (green) data.

$$L(\beta) = \|y - \boldsymbol{X}\beta\|_2^2 + \alpha\|\beta\|_1$$

where $\|...\|_1$ and $\|...\|_2^2$ are the $L1$ and $L2$ norms, respectively, and α controls the amount of shrinkage. By using 10-fold cross-validated Lasso, the number of primary features, having nonzero β, reduces to 15 and are shown in Figure 4a and Figure S4. Pearson's correlation coefficients $p = \mathrm{cov}(x_i, x_j)/(\sigma_{x_i}\sigma_{x_j})$ between any two features x_i and x_j are calculated to ensure absence of linear relationship, and are shown as a heat map (Figure 4b). Here, cov and $\sigma_{x_{i/j}}$ are covariance and the standard deviation of $x_{i/j}$, respectively. The value of $|p|$ ranges between 0 and 1, and we consider only one from given two features having $|p| > 0.9$. As such, $|p|$ between formation energy (H_f) and mean electron affinity ($\mathrm{EA}^{\mathrm{mean}}$) is 0.54 (Figure S5); therefore both are included in the feature set. Further, to incorporate nonlinearity in the feature space, *compound* features are generated by applying simple mathematical operations such as x_i, x_i^2, $\sqrt{x_i}$, $\sqrt[3]{x_i}$, e^{x_i}, $\ln(1 + x_i)$, and their inverses on 15 primary features. These feature sets are extended to two and three dimensions by simply multiplying them twice and thrice, respectively (see the Supporting Information), leading to ∼1 million features. These compound features are also reduced to 8 by the application of Lasso.

To build the learning model, $\{\boldsymbol{X}, y\}$ is split into the training ($\{\boldsymbol{X}, y\}_{\mathrm{train}}$) and testing ($\{\boldsymbol{X}, y\}_{\mathrm{test}}$) sets. 90% of the data set is randomly selected for training the model and the remaining for testing. Regression fit of $\boldsymbol{X}_{\mathrm{train}}$ to the corresponding y_{train} builds the model. $\boldsymbol{X}_{\mathrm{test}}$ is introduced as input to the model, which gives predicted y' as band gaps. Differences between the true (G_0W_0) y_{test} and y' are evaluated by statistical regression metrics such as rmse, mean-absolute error (mae), and the coefficient of determination (R^2) (see details in the Supporting Information). rmse and mae indicate deviation of y' from y_{test} as the measures of accuracy, and R^2 describes variability in the data sets. From now on, to avoid ambiguity, $\boldsymbol{X}$ and y correspond to the training data set unless otherwise specified.

Next, we choose appropriate methods such as the KRR and SVR to train the learning model.[39] Previous studies on band gap predictions of double perovskites[23] and several inorganic compounds[22] have used KRR and SVR methods, respectively. KRR is the kernalized version of ridge regression and uses $L2$ norm regularization, which calculates the squared error loss using all the training data. SVR uses the kernel trick and a modified loss function,[39] with a subset of the training data. KRR and SVR are applied using the radial basis function kernel (with a length scale of δ)

$$k(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2\delta^2}\right)$$

through k-fold cross-validated grid search approach. Having incorporated nonlinearity into account, these algorithms fit the model for prediction.

Here, 5-fold cross-validated KRR and SVR with grid search for hyperparameters tuning are used to predict the band gap of MXene. The features are examined as unique combination in sets of r at a time, where $r = 1, 2, \cdots, 15$. For every combination, data shuffling up to 50 trials are performed, and the models with R^2 in the range 0−1 are considered. For a given r, the best heuristic feature-combination is identified, which yields a large R^2 with minimum rmse. This helped in identifying the most significant combination with least number of features having averaged errors and R^2. As such, the optimized KRR model produces volume per atom (V), phase (c), H_f, and mean boiling point ($T_{\mathrm{B}}^{\mathrm{mean}}$) as the best features. This combination ($r = 4$) produces train/test rmse, mae, and R^2 values of 0.15/0.19 eV, 0.11/0.15 eV, and 0.82/0.68, respectively (Figure 5a). In Figure 5, scatter plots are shown for individual models with averaged regression metrics, and the predictions corresponding to the

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

training and test data sets as inputs are shown by red diamonds and green circles, respectively. Each scatter plot is the outcome of a model, which is shown with the averaged regression metrics within ±0.01 eV of the range for errors (and ±0.01 for R^2). With the same set of conditions, and mean M/M′−X bond length (d), V, c, and standard deviation of melting point (T_M^{std}) as feature-combination, SVR produces 0.13/0.17 eV, 0.11/0.15 eV, and 0.87/0.71 as rmse, mae, and R^2, respectively (Figure 5b). These models are able to achieve low errors; however, the data variability shown by these models may not be adequate, which motivates us to check other possible methods.

As the aim is to achieve better accuracy with data variability, the kernel-based nonparametric GPR method based on probability distributions is preferred for prediction with a smaller data set to KRR and SVR methods. In this method, y' is assumed to be the result of a Gaussian process (f) affected by independent additive noise (ϵ) as $y' = f + \epsilon$, where $\epsilon \sim \mathcal{N}(\epsilon|0, \nu)$ is a normal distribution with zero mean, and ν is the variance. Given data set $\{\boldsymbol{X}, y\}$ and a new input x^*, f^* can be predicted with the zero mean Gaussian predictor distribution[40]

$$p(y', f^*|\boldsymbol{X}, x^*) = \mathcal{N}(y', y^*|0_{n+1}, \boldsymbol{K})$$

where 0_{n+1} is an n+1-dimensional zero-vector, and $\boldsymbol{K}$ represents the covariance matrix. Here, $\{\boldsymbol{X}, x^*\}$ represents the joint input which corresponds to a joint output $\{y', y^*\}$. $\boldsymbol{K}_{ij} = k(x_i, x_j)$ is made from the Matern-$\frac{5}{2}$ kernel (with hyperparameters of input length scale δ and variance ν) given by

$$k(x_i, x_j) = \nu\left(1 + \frac{\sqrt{5}(x_i - x_j)}{\delta} + \frac{5(x_i - x_j)^2}{3\delta^2}\right) \exp\left(-\frac{\sqrt{5}(x_i - x_j)}{\delta}\right)$$

The GPR optimized model with 8 features, namely, d, V, c, H_f, mean of van der Waals radius (r_{vdw}^{mean}), standard deviations of the group number in periodic table (G_{std}), ionization energy (IE^{std}), and T_M^{std} resulted in train/test rmse of 0.06/0.14 eV. The corresponding mae and R^2 are found to be 0.04/0.11 eV and 0.97/0.83, respectively (see Figure 5c and Figure S6). With the lowest ever rmse of 0.14 eV and larger R^2 of 0.83 for band gap prediction, the GPR model outperforms KRR and SVR, and turned out to be the best model. Furthermore, a direct correlation exists between the hyperparameters of GPR and KRR. Hence, once the GPR model is built, one can use this correlation to get a KRR model with similar accuracy (Figure S7). Therefore, for this problem, the GPR turns out to be most efficient and accurate model.

Though the GPR comes out to be the best model, the MXene data set is also examined with the ensemble-based method. Such a model is developed by using a decision tree as the base estimator, which will predict the accurate band gaps, provided the features of all semiconducting MXene. As such, the ensemble-based bagging method is applied to the training data set for obtaining the averaged prediction metrics with reduced variance. The model prediction y' at input $\boldsymbol{X}$ by the bagging method is averaged over a collection of B bootstrap samples. Given $\boldsymbol{X}$, for each b (where, b = 1, 2, ⋯, B), the bootstrap replicates $\{\boldsymbol{X}_b, y_b\}$ fit the model and yield the prediction y_b'. The bagging estimate is defined by

$$y' = \frac{1}{B}\sum_{b=1}^{B} y_b'$$

which reduces mean-squared error considerably. The optimized model resulted with the best feature-combination of 7, namely, V, c, IE^{std}, G^{std}, mean of the group number in periodic table (G^{mean}), r_{vdw}^{mean}, and standard deviation of van der Waals radius (r_{vdw}^{std}) having the train/test rmse of 0.12/0.16 eV. Corresponding mae and R^2 are 0.09/0.13 eV and 0.89/0.73, respectively (Figure 5d). As such, rmse is reasonably low in comparison to the SVR model, and R^2 values are not better than the GPR (Table S2).

To assess the effect of compound features, models are developed using all 4 methods (Figure 5e–h). In terms of prediction accuracy, the compound features are not better than the primary features. These features resulted in similar metrics for the best KRR, SVR, GPR, and bagging models as primary features. This involved the train/test R^2 values of 0.76/0.69 for KRR and SVR both, 0.78/0.7 and 0.88/0.74 for GPR and bagging, respectively. Though the bagging model shows better R^2 values, the GPR model is the best based on its lowest rmse (Figure 5g). However, rmse from compound features (0.17 eV) is higher than rmse (0.14 eV) obtained using the primary features. In the cases of GPR and bagging, change in R^2, rmse, and mae for primary feature-combinations (r = 2, 4, 6, and 8) are shown in Figure 6. R^2 for train/test increases with respect to r, whereas rmse and mae are reduced.

For a physical intuition of the learning model, it is important to know about the feature importance of the whole data set $\boldsymbol{X}$, which is examined by comparing the bagging (Figure 7) and Relieff (Figure S7) methods. For bagging method, V, T_B^{mean}, G^{std}, and T_M^{std} are rendered as the top features of which T_B^{mean}, G^{std}, and T_M^{std} are the elemental properties. The Relieff algorithm yields three elemental (G^{std}, T_M^{std}, T_B^{mean}) and one computed property (vacuum potential: Φ_L) as the top 4 ranked features. Hence, the common features based on the importances include G^{std}, T_M^{std}, and T_B^{mean}. On the other hand, the best features can be selected based on the results of the regression algorithm. For example, 50 independent trials of the bagging regression with r = 1 are performed, which resulted in G^{std} as the best feature (70% of the trials) in comparison to G^{mean} (30%). For varying r, such trials resulted in high reproducibility of R^2 and rmse, which implies the usefulness of heuristic features resulting in meaningful predictions. For bagging with r = 7, the best heuristic combination included two computed (V and c) and 5 elemental (G^{mean}, G^{std}, IE^{std}, r_{vdw}^{mean}, and r_{vdw}^{std}) features. The 8 features listed by GPR are a combination of 4 computed (d, V, c, and H_f) and 4 elemental (r_{vdw}^{mean}, G_{std}, IE^{std}, and T_M^{std}) properties. In all the 4 regression

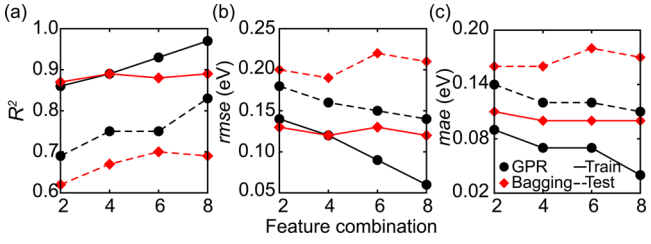


Figure 6. Variation in regression metrics of GPR (black circle) and bagging (red diamond). For r = 2, 4, 6, and 8 feature-combinations, (a) R^2, (b) rmse, and (c) mae of primary features are shown.

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

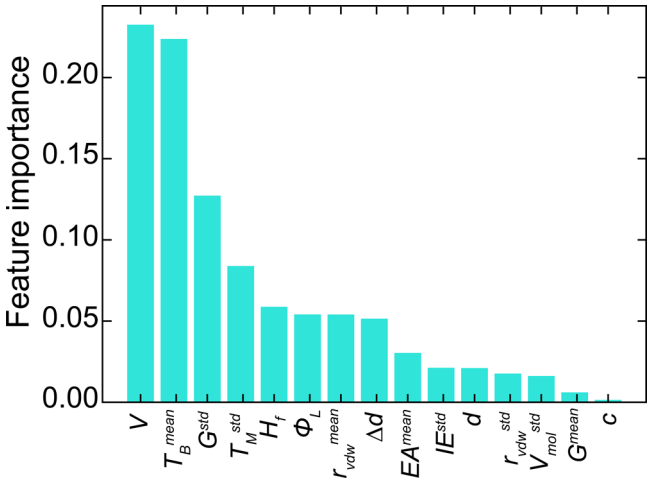


Figure 7. Primary feature importance calculated through bagging method.

methods, the best feature-combination includes one or more elemental features from G^{std}, T_M^{std}, and T_B^{mean}, which are also common important elemental features from bagging and Relieff methods. The most important aspect is absence of PBE band gap, conduction band minima (CBM), and valence band maxima (VBM) as features. Hence, computed properties in comparison to elemental are not necessarily the major ingredients for the prediction. Rather, combination with minimum number of available elemental and easily computed features predicts the accurate MXene band gaps within no time, which otherwise takes years.

Some relevant features have physical significance and a direct correlation with the GW band gap, whereas others are only statistically related (Figures S8). T_B^{mean} and T_M^{std} show a strong positive correlation with the GW band gap. A large T_B^{mean} and T_M^{std} imply stronger bonding in the elemental solids, which will result in lower formation enthalpy of pristine MXene (Figure 8). These MXenes will naturally have relatively weaker bonds and, it is easier for them to be functionalized via stronger interaction with the functional group, leading to opening of larger band gap. This intuitive picture is well captured by our model. Taking these properties as features, we can carry out high-throughput screening of the MXene.

In summary, we have developed machine-learning models for accurate band gap predictions of MXene. We generate structures of 23 870 MXenes, out of which 7200 MXenes are randomly selected to create a database containing calculated optimized structural and electronic properties. A metal−semiconductor classification model with 94% accuracy was developed. Among the semiconducting MXenes, 76 MXenes are selected to build a prediction model. Input features to the learning models are the computed properties of 76 MXenes at the DFT−PBE level and elemental information from chemical repositories, such as boiling and melting points, group number, atomic radii, phases, and bond lengths. The models are built by correlating the features to the calculated GW band gap. Among the KRR, SVR, GPR, and bagging methods, the GPR model predicted the band gaps with unprecedented low rmse of 0.14 eV. T_B^{mean} and T_M^{std} show a strong positive correlation with the GW band gap, which unveil the roles of constituent elements. Most importantly, the models do not require prior knowledge of PBE band gap, and positions of CBM and VBM as features for prediction making them more accessible to the wider community. Our model can be employed to estimate the accurate GW band gaps of the entire MXene database within minutes. Additionally, a detailed comparative study of applicability of different popular approaches applied to predict materials property would have a significant impact on the model development for high-throughput property predictions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.chemmater.8b00686.

Additional information including structures of bb′ and cb phases of MXene, energy difference between them, classification learning model, PBE and G_0W_0 band gaps for selected 76 MXenes, feature sets, outlier removal, annotated correlation plots, the GPR model predictions with confidence interval, regression metrics for primary feature-combinations and methods, equivalence of GPR and KRR, Relieff algorithm, and the MXene data set (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: abhishek@iisc.ac.in.

ORCID

Arunkumar Chitteth Rajan: 0000-0002-8255-022X
Avanish Mishra: 0000-0003-3997-0445
Abhishek K. Singh: 0000-0002-7631-6744

Author Contributions

§A.C.R. and A.M. contributed equally.

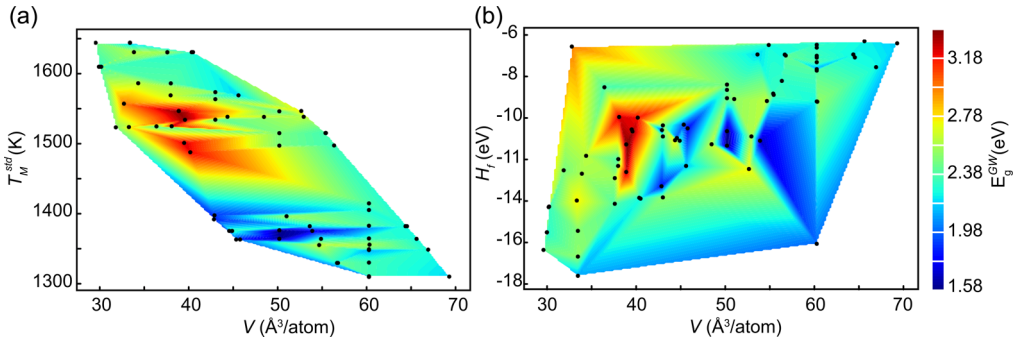


Figure 8. Variation in the GW gap as a function of (a) T_M^{std} and V, and (b) H_f and V for 70 MXenes.

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from KIST (Grant 2E26940) and thank Indian Institute of Science for providing the computing facilities of Materials Research Centre, Thematic Unit of Excellence, and Supercomputer Education and Research Centre. A.C.R. and A.M. acknowledge UGC India for Dr. D S Kothari Postdoctoral Fellowship and Senior Research Fellowship, respectively. A.K.S. and A.M. are thankful for support from DST Nano Mission.

■ REFERENCES

(1) Naguib, M.; Kurtoglu, M.; Presser, V.; Lu, J.; Niu, J.; Heon, M.; Hultman, L.; Gogotsi, Y.; Barsoum, M. W. Two-Dimensional Nanocrystals Produced by Exfoliation of Ti_3AlC_2. *Adv. Mater.* **2011**, *23*, 4248−4253.

(2) Khazaei, M.; Arai, M.; Sasaki, T.; Chung, C.-Y.; Venkataramanan, N. S.; Estili, M.; Sakka, Y.; Kawazoe, Y. Novel Electronic and Magnetic Properties of Two-Dimensional Transition Metal Carbides and Nitrides. *Adv. Funct. Mater.* **2013**, *23*, 2185−2192.

(3) Khazaei, M.; Ranjbar, A.; Arai, M.; Yunoki, S. Topological insulators in the ordered double transition metals $M_2'M''C_2$ MXenes (M'=Mo, W; M''=Ti, Zr, Hf). *Phys. Rev. B: Condens. Matter Mater. Phys.* **2016**, *94*, 125152.

(4) Si, C.; Jin, K.-H.; Zhou, J.; Sun, Z.; Liu, F. Large-Gap Quantum Spin Hall State in MXenes: *d*-Band Topological Order in a Triangular Lattice. *Nano Lett.* **2016**, *16*, 6584−6591.

(5) Sharma, G.; Naguib, M.; Feng, D.; Gogotsi, Y.; Navrotsky, A. Calorimetric Determination of Thermodynamic Stability of MAX and MXene Phases. *J. Phys. Chem. C* **2016**, *120*, 28131−28137.

(6) 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.

(7) Mishra, A.; Srivastava, P.; Carreras, A.; Tanaka, I.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Atomistic Origin of Phase Stability in Oxygen-Functionalized MXene: A Comparative Study. *J. Phys. Chem. C* **2017**, *121*, 18947−18953.

(8) Mishra, A.; Srivastava, P.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Isolation of Pristine MXene from Nb_4AlC_3 MAX Phase: A First-Principles Study. *Phys. Chem. Chem. Phys.* **2016**, *18*, 11073−11080.

(9) 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.

(10) Zhang, C. J.; Pinilla, S.; McEvoy, N.; Cullen, C. P.; Anasori, B.; Long, E.; Park, S.-H.; Seral-Ascaso, A.; Shmeliov, A.; Krishnan, D.; Morant, C.; Liu, X.; Duesberg, G. S.; Gogotsi, Y.; Nicolosi, V. Oxidation Stability of Colloidal Two-Dimensional Titanium Carbides (MXenes). *Chem. Mater.* **2017**, *29*, 4848−4856.

(11) Ran, J.; Gao, G.; Li, F.-T.; Ma, T.-Y.; Du, A.; Qiao, S.-Z. Ti_3C_2 MXene Co-Catalyst on Metal Sulfide Photo-Absorbers for Enhanced Visible-Light Photocatalytic Hydrogen Production. *Nat. Commun.* **2017**, *8*, 13907.

(12) Mashtalir, O.; Cook, K.; Mochalin, V.; Crowe, M.; Barsoum, M.; Gogotsi, Y. Dye Adsorption and Decomposition on Two-Dimensional Titanium Carbide in Aqueous Media. *J. Mater. Chem. A* **2014**, *2*, 14334−14338.

(13) Guo, Z.; Zhou, J.; Zhu, L.; Sun, Z. MXene: A Promising Photocatalyst for Water Splitting. *J. Mater. Chem. A* **2016**, *4*, 11446−11452.

(14) Lashgari, H.; Abolhassani, M.; Boochani, A.; Elahi, S.; Khodadadi, J. Electronic and Optical Properties of 2D Graphene-Like Compounds Titanium Carbides and Nitrides: DFT Calculations. *Solid State Commun.* **2014**, *195*, 61−69.

(15) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.

(16) Aryasetiawan, F.; Gunnarsson, O. The GW Method. *Rep. Prog. Phys.* **1998**, *61*, 237−312.

(17) Crowley, J. M.; Tahir-Kheli, J.; Goddard, W. A. Resolution of the Band Gap Prediction Problem for Materials Design. *J. Phys. Chem. Lett.* **2016**, *7*, 1198−1203.

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

(19) 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.

(20) 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.

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

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) Legrain, F.; Carrete, J.; van Roekeghem, A.; Curtarolo, S.; Mingo, N. How Chemical Composition Alone Can Predict Vibrational Free Energies and Entropies of Solids. *Chem. Mater.* **2017**, *29*, 6220−6227.

(27) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304−1311.

(28) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Systematically Identifying Efficient Physical Models of Materials Properties. 2017, arXiv:1710.03319. arXiv.org e-Print archive. https://arxiv.org/abs/1710.03319.

(29) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(30) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite(ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.

(31) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to Represent Crystal Structures for Machine Learning: Towards Fast Prediction of Electronic Properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 205118.

(32) Glawe, H.; Sanna, A.; Gross, E. K. U.; Marques, M. A. L. The Optimal One Dimensional Periodic Table: A Modified Pettifor Chemical Scale from Data Mining. *New J. Phys.* **2016**, *18*, 093011.

(33) aNANt: A Functional Materials Database. http://anant.mrc.iisc.ac.in/.

(34) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors using A Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038

(35) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 1758.
(36) Rinke, P.; Janotti, A.; Scheffler, M.; de Walle, C. G. V. Defect Formation Energies without the Band-Gap Problem: Combining Density-Functional Theory and the GW Approach for the Silicon Self-Interstitial. *Phys. Rev. Lett.* **2009**, *102*, 026402.
(37) Dynamic Periodic Table. https://ptable.com/ (accessed on July 10, 2017).
(38) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, 2001.
(39) Murphy, K. P. *Machine Pearning: A Probabilistic Perspective*; Massachusetts Institute of Technology Press, 2012.
(40) Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press, 2012.

DOI: 10.1021/acs.chemmater.8b00686
Chem. Mater. 2018, 30, 4031−4038