THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

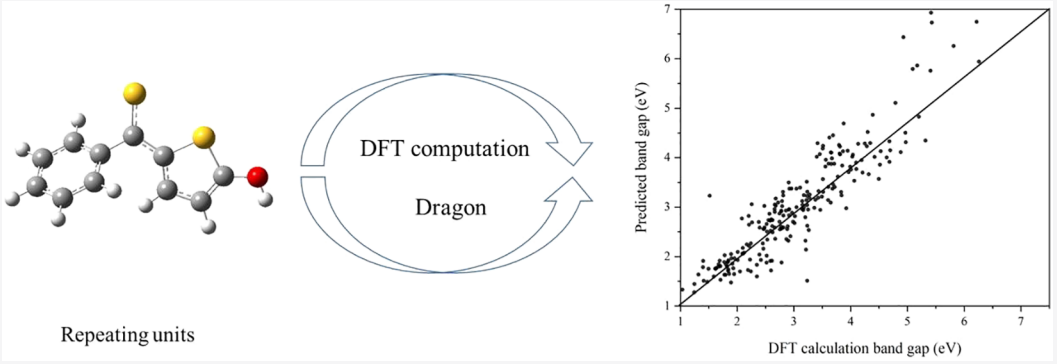
Machine Learning Aided Design of Polymer with Targeted Band Gap Based on DFT Computation

Pengcheng Xu, Tian Lu, Lifei Ju, Lumin Tian, Minjie Li, and Wencong Lu*

Cite This: *J. Phys. Chem. B* 2021, 125, 601−611

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Polymer band gap is one of the most important properties associated with electric conductivity. In this work, the machine learning model called support vector regression (SVR) was developed to predict the polymer band gap, where the training data of the polymer band gap were obtained from DFT computation while the descriptors were generated from Dragon. After feature selection with the maximum relevance minimum redundancy, the SVR model using 16 key features as inputs gave the optimal performance for predicting polymer band gaps. The determination coefficient (R^2) of the SVR model between the DFT computations and SVR predictions of polymer band gaps reached as high as 0.824 for the leave-one-out cross-validation and 0.925 for the independent test. Besides, the 16 key features were explored through correlation analysis and sensitivity analysis. The available model can be used to screen out the polymers with targeted band gaps before experiments, which is very helpful for rapid design of new polymers.

1. INTRODUCTION

Polymer has been widely permeated into the branches of material science due to the unique properties from all aspects.[1,2] A surge in the application of polymer materials has been witnessed in the past decades, ranging from daily products[3] to biomedical materials[4−6] to military applications[7] to aircraft fields,[8] etc. In recent years, considerable attention has been attracted to the polymer electrical properties[9−11] because of the great potential in electroluminescence materials,[12] polymer solar cells,[13] and organic thin film transistors.[14] Band gap, one of the most important electrical properties, is regarded as a crucial screening parameter in rational design of functional materials. The determination of polymer band gap by using traditional experimental ways such as cyclic voltammetry and UV−visible spectroscopy[15] is generally limited and time-consuming in the search for a new polymer with targeted band gap. Thus, more efforts have been targeted on the DFT computation to obtain the band gap before experiments.[16−21] However, it is still tough to obtain band gaps of huge candidate polymers via DFT computation, which requires a powerful computing platform and professional knowledge to get band gap data with high fidelity. Consequently, how to obtain the polymer band gap data in a convenient, fast, and accurate way remains a challenge.

With the presence of the Materials Genome Initiative (MGI), the increasing availability of big data has brought about the fast-growing development in the field of material informatics.[22,23] Meanwhile, the applications of the machine learning algorithm in material science have caught remarkable focus during the past few years.[24−27] On the basis of robust material properties data, the correlations between target properties and crucial features can be mapped through machine learning. Accordingly, the development of material design will be accelerated at a rapid rate.[28] In recent studies, much research has been focused on the prediction of polymer properties with machine learning algorithms, including band

Received: September 23, 2020
Revised: December 23, 2020
Published: January 7, 2021

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

gap, thermal conductivity, and other properties.[29−32] G. Pilania et al.[33] demonstrated a multifidelity co-kriging statistical learning framework combined with variable quantum mechanical calculations of band gap to generate a machine-learned model for accurate predictions of the band gap at the highest fidelity level. Arun Mannodi-Kanakkithodi et al.[34] utilized a generated data set of the electronic and dielectric properties of polymers to test different kinds of regression algorithms and explored several possibilities for the hyperparameter optimization to increase the prediction accuracy of the model. Stephen Wu et al.[35] used transfer learning to build a model for polymer thermal conductivity prediction and employed the Bayesian algorithm to design molecular structures. After molecular screening by the model and synthetic accessibility (SA), three polymers with high thermal conductivity were selected and synthesized through experiments. These related works indicate that machine learning will assist experimental researchers in the prediction or structure design for polymers with target properties.

However, the disadvantages of the models mentioned above are setting the polymer fingerprint as descriptors, which only considers the composition of the polymer and ignores the structural information. It might be feasible only on the condition that the polymerization degree of the polymer has little effect on the band gap value.[36] To make full use of the structural information on the monomeric unit building block, we used Dragon software to obtain thousands of feature data including both compositional and structural information. Dragon, developed by the Italian Kode team, is the most inclusive descriptor software, which can calculate a total of 5720 molecular descriptors in 30 types, including the most component and structural parameters with mature theoretical system foundation.[37,38]

In this work, both DFT computation and machine learning algorithms were applied to predict the band gap values. DFT methods with different fidelity were taken into consideration for the appropriate band gap values. The most suitable DFT method with best performance was selected by comparing the calculated and experimental values of the target. Then, the DFT method was used to get the band gap data of the training set for machine learning. Afterward, the machine learning model was constructed to predict band gap values based on the descriptors generated from Dragon and the DFT calculated band gaps. Therefore, the band gap of any polymer could be predicted via machine learning instead of experiment or sophisticated DFT computation.

2. MATERIALS AND METHODS

2.1. Data Preparation. In total, 284 four-block polymers have been collected from the published papers.[39] The structures were optimized at the level of B3LYP/6-31G(d,p)[40] before the descriptor generation, followed by the frequency analysis that was used to certificate the minima energy at the same level. The band gap values are simulated with the same basis set, where various functions have been considered to guarantee the rationality. Then, the entire 5270 descriptors are generated via Dragon 7; after removing the features with a Pearson correlation coefficient greater than 0.95 and a standard deviation less than 0.0001 among the features, 1093 features finally remained. As for a more rational data splitting, sphere-exclusion[41] is adopted to divide the data set at 4:1 ratio with 228 training samples and 56 testing samples. The training set was used for the feature selection and model construction, while the test set was used to validate the model generalization ability.

2.2. Maximum Relevance Minimum Redundancy. Feature selection is aiming to find a subset of the relevant features from original features with the accordance to defined criteria, which is frequently restrained by the redundancy of feature pairs.[42] The mRMR method was applied to search for the optimal subset with maximum relevance as well as minimum redundancy, and the detail of algorithm for mRMR is shown as follows.[43−45]

The maximum relevance reflects the correlated ability of features to the target property. A feature subset S containing n features is found to maximize the relevance between all features and target property in S. And minimum redundancy takes the minimum similarity among all features into consideration, which makes every feature in S the least similar to each other. The formulas are as follows

$$\max D(S, T) \quad D = \frac{1}{n} \sum_{x_i \in S} I(x_i, T)$$

$$\min R(S) \quad R = \frac{1}{n^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

where n is the number of features, S is the feature subset, T is the target property, and x_i and x_j are the ith and jth features. $I(x_i, T)$ is the mutual information function of x_i and T, D is the relevance index between x_i and T, R is the redundancy index between x_i and x_j, and $I(x, y)$ is determined by probabilistic density functions $p(x)$.

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \, dx \, dy$$

Then, the strategy of incremental search is used to find approximately optimal features. Assuming X and S_m are the original feature set and the subset, x_i belongs to the different set of X and S_m, $\max \Phi(D, R)$ is the feature evaluation standard, in which Φ is the division of $D(S, T)$ and $R(S)$, and the evaluation conditional function of mRMR is

$$\mathrm{mRMR}_m(x_i) = \max_{x_i \in X - S_m} \left[\frac{I(x_i, T)}{\frac{1}{n} \sum_{x_i \in X - S_m} I(x_i, x_j)} \right]$$

After all of the features are ranked according to the mRMR evaluation results, the most appropriate feature subset is extracted from the initial features to form the input feature vector of the SVM model.

2.3. Support Vector Machine. Support vector machine (SVM) has gained lots of popularity since it was introduced formally in 1995.[46] It has been proven to be very effective not only in classifications but also in regression problems.[47−49] The general idea of SVM is that it maps the input vector into high dimensional space and finds the most optimal hyperplane as the criterion to classify the samples.

In classification, the support vector machine algorithm is named the support vector classifier (SVC) as well. A segmentation plane will be established to make samples of different types furthest from each other; the plane is called the maximal margin hyperplane. SVC is to get the classification line with the maximal margin hyperplane. Given the input matrix of feature x and target y, the function of the hyperplane in the sample space could be formed as

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

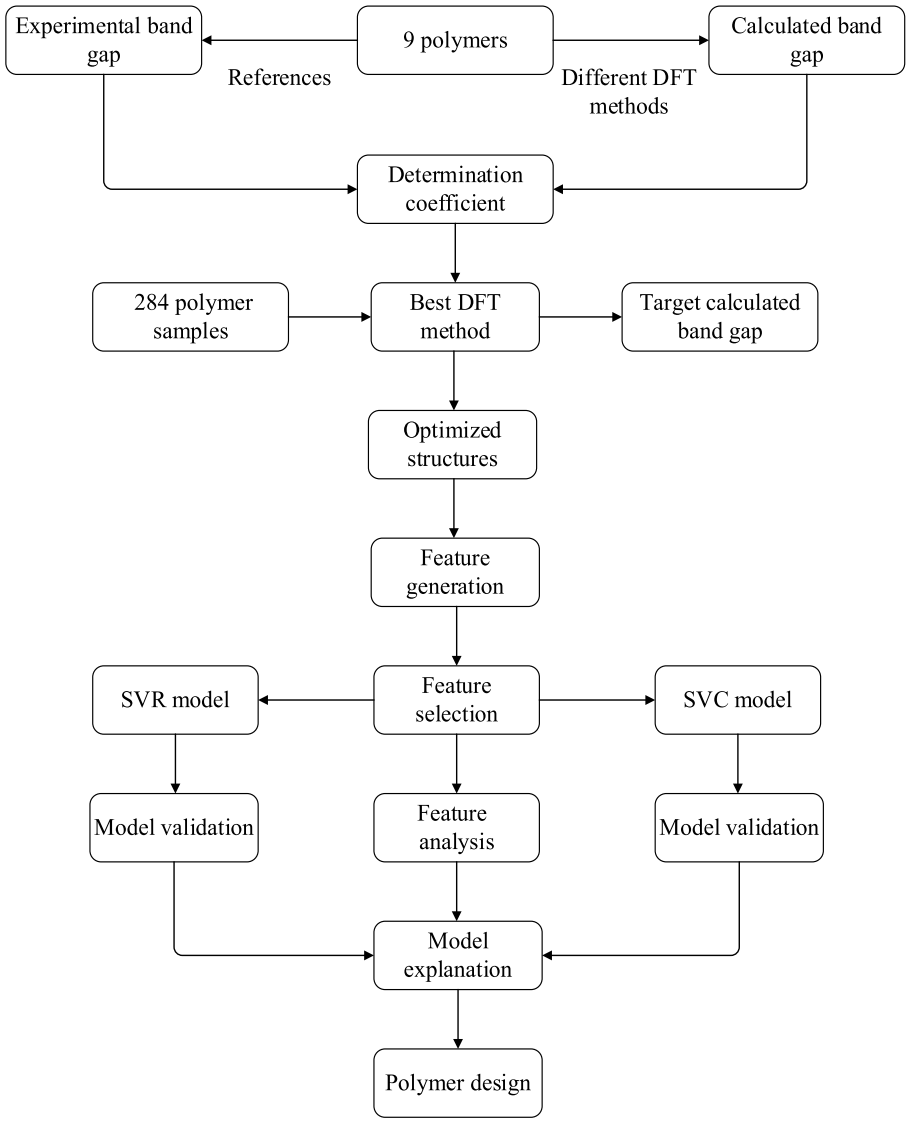


Figure 1. Workflow of polymer data mining in this work.

$$w^T\mathbf{x} + b = 0$$

where w is the normal vector that determines the direction of the hyperplane, $\mathbf{x}$ is the feature matrix, and b is the displacement parameter that determines the distance between the hyperplane and the origin. It is clear that the hyperplane could be confirmed by $\mathbf{w}$ and b. It is assumed that the hyperplane can classify the training samples correctly as the function shown below

$$\begin{cases} w^T x_i + b \geq +1, y_i = +1 \\ w^T x_i + b \leq -1, y_i = -1 \end{cases}$$

where x_i and y_i belong to the training set. The training samples closest to the hyperplane make the above equation equal, which are called support vectors. The sum of the distance between different support vectors and the hyperplane is called the margin γ:

$$\gamma = \frac{2}{\|w\|}$$

The hyperplane should maximize the margin of different samples, and the function can be viewed to find the W and b that maximize γ under the constraint condition shown as follows

$$\max_{W,b} \frac{2}{\|w\|}$$

$$s.t.\ y_i(w^T x_i + b) - 1 \geq 0,\ i = 1, 2, \ldots, n$$

Since the objective function is quadratic and the constraint is linear, the equations above could be regarded as the convex function optimization, which can be solved through the Lagrange multiplier. The formula is shown as follows

$$L(w, b, \alpha) = \frac{1}{2}(W^T W) - \sum_{i=1}^{n} \alpha_i[y_i(w^T x_i + b) - 1] = 0$$

where α_i is a Lagrange coefficient. To obtain the minimum value of the Lagrange function, partial derivatives of the function to α_i, W, and b, respectively, are set to 0.

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

$$\frac{\partial L}{\partial W} = 0 \rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow \alpha_i[y_i(w^T x_i + b) - 1] = 0$$

The three functions above with the original constraint condition will transfer the issue into a dual problem for convex quadratic programming.

$$\begin{cases} \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j) \\ s.t.\ \alpha_i \geq 0,\ i = 1, 2, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

This is a quadratic function extremum problem with inequality constraint, where a unique optimum solution exists. w^* and b^* in the classification line can be obtained by obtaining the optimum solution α_i^*, and the SVC prediction result can be obtained according to the following formula.

$$f(x) = sgn(w^{*T} x + b)$$

In regression, the support vector machine algorithm is also called support vector regression (SVR). The SVR algorithm weighs empirical and structural risks. Different from traditional regression algorithms, SVR considers the errors of limited sample data in the model. The insensitive channels ε are used to handle the problem. Specifically, the error is ignored when the predicted value $\hat{y}$ meets $|y_i - \hat{y}_i| \leq \varepsilon$; otherwise, the error is $|y_i - \hat{y}_i| - \varepsilon$. The deviation is concerned only when it is greater than ε in the empirical risk calculation. There are countless solutions to the regression model that satisfies the conditions above, which forms a "pipe". Similar to the constraint conditions of SVC, SVR takes the minimum value of $\|w\|$ as the standard to select the optimum solution of the "pipe" to improve the prediction accuracy of the model.

3. RESULTS AND DISCUSSION

3.1. Workflow. The workflow of the whole process is shown in Figure 1. In total, 293 polymer samples were collected from the published papers, consisting of the following building blocks: CH_2, NH, CO, C_6H_4, C_4H_2S, CS, and O. These polymer samples were divided into two parts, 9 samples with determined experimental band gap and 284 samples without. After structure optimization, various DFT methods were considered to calculate the band gap of 9 samples. According to the determination coefficient (R^2) of the DFT calculated and experimental values, the optimal DFT method was used to archive the band gap of the 284 four-block (the number of monomeric unit building blocks is four) polymer samples, while the features were generated from Dragon. In order to reduce the input information, the maximum relevance minimum redundancy (mRMR) algorithm was employed to select crucial features correlated with the band gap most. Afterward, the regression model was constructed with the support vector regression (SVR) algorithm. In addition, a SVC model was also built according to the characteristics of the data to further explore the patterns between selected features and polymer band gap. The performances of the models were evaluated based on prediction errors of the leave-one-out cross-validation and independent test. The constructed models were explained in feature correlation analysis and sensitivity analysis. At last, some polymer samples with targeted band gap were designed for experimental reference.

3.2. DFT Method Exploration. In order to calculate the band gap of 284 samples without experimental target, various DFT methods including common functions and customized combinations with different exchange and correlation functions were taken into consideration based on the 9 samples after structure optimization with B3LYP/6-31G(d,p). As shown in Figure 2, the combination methods have better performance

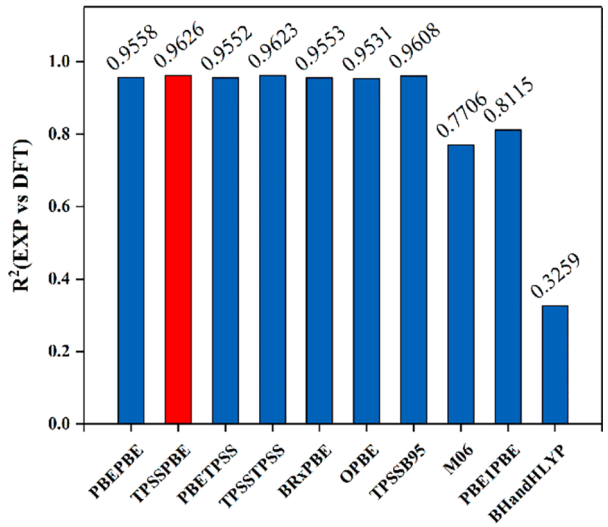


Figure 2. DFT method exploration by the R^2 of experimental values versus DFT calculations.

than the commonly used DFT methods with R^2 values fluctuating around 0.95. Besides, the combinations containing TPSS and PBE have a notable contribution to R^2, which might indicate that, compared to the usual DFT methods, the hybridized functional could be more suitable for the calculation of polymer band gap. According to the highest R^2 value of 0.9626, TPSSPBE was finally adopted to obtain band bap values of the 284 samples. The computing results can be found in the Supporting Information. It should be noted that the DFT methods for polymer band gap calculation in different papers vary a lot due to the complexity of the polymer systems. The optimal DFT method usually requires exploration in combination with specific data. De Oliveira et al.[50] explored the DFT method for the energy gap of cyano-bithiophene derivatives, and both SVWN/6-31G* and BLYP/6-31G* showed reasonable agreement with the experimental data. Ling et al.[51] found the functionals O3LYP and B3LYP compared well with the experimental band gap value of alternating triphenylamine−fluorene copolymers. Arı et al.[16] made a comparison of DFT functionals for the prediction of conjugated band gap, and B3PW91 was viewed to be the best method to obtain accurate band gap values.

3.3. Feature Selection. After data preprocessing, there are 1093 descriptors in total for feature selection. In this work, the machine learning results of training data set via maximum

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

relevance minimum redundancy (mRMR)-support vector machine (SVM) were employed to screen out the subset of features for modeling. mRMR-SVM can be used to find and rank the optimal subset of features correlated with the target most. In order to evaluate the feature subset, the correlation coefficient (R) between the DFT calculated and model predicted band gap of the leave-one-out cross-validation (LOOCV) was employed as the measure of goodness-of-fit.

All of the considered features could be sorted according to the correlation with the target property by mRMR-SVM. Figure 3 shows the trend of correlation coefficient with the

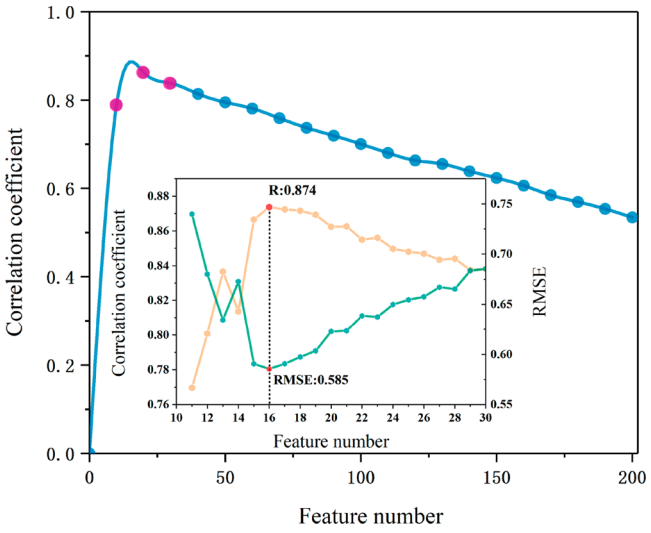


Figure 3. Feature selection by mRMR.

feature numbers. The top 200 features were considered to be selected at an interval of 10 to construct the SVM model, under the evaluation of R in the LOOCV. It can be found in Figure 3 that R raises with the increasing feature number and then gradually decreases after reaching the peak. The most suitable feature numbers may be around the peak, so more detailed calculations were made during this interval. The trends of correlation coefficient (R) and root-mean-square error (RMSE) with the number of selected features ranging from 10 to 30 are represented in the inset of Figure 3. It could be found that the trend of R is just opposite to RMSE. As shown in the figure, the SVM model with the first 16 features has the best performance with the highest R as well as the lowest RMSE. The selected features and the corresponding meanings are listed in Table 1. These 16 selected features include the compositional information (nO and nTA), topological and geometrical structural information (HATS8u, VE1sign_RG, E3v, SRW05, SM14_AEA(bo), GATS7i, DISPp, F03[C-S], and CATS2D_00_DD), burden eigenvalues (SpMin5_Bh(m)), VSA-based descriptors (P_VSA_ppp_D, P_VSA_LogP_4, and P_VSA_MR_2), and edge adjacency indices (SpDiam_EA(dm)).

3.4. Feature Correlation Analysis. In order to further explore the linear correlation between the band gap and the selected features, the Pearson correlation matrix plot was introduced. As shown in Figure 4, the gradation of element color in the matrix plot from yellow to black maps the numeric increase of the Pearson correlation coefficient ranging from −1 to +1, indicating the negative linear correlation to the positive. We may find that F03[C-S] and DISPp have a strongly

Table 1. 16 Selected Features by Maximum Relevance Minimum Redundancy-Support Vector Regression

selected features	meaning
nO	number of O atoms
HATS8u	leverage-weighted topological structure autocorrelation of path length 8
VE1sign_RG	coefficient sum of the last eigenvector from the reciprocal squared geometrical matrix
P_VSA_ppp_D	sum of the VSA (van der Waals surface area) contributions of all of the atoms assigned to PPP (potential pharmacophore point) atom types D(O in −OH, N in NH or NH_2)
nTA	number of total atoms
E3v	third component accessibility directional WHIM (weighted holistic invariant molecular descriptors) index/weighted by van der Waals volume
SRW05	the "walks" of fifth order vertex or self-returning walk count of fifth order
SM14_AEA(bo)	spectral moment of order 14 from augmented edge adjacency matrix weighted by bond order
P_VSA_LogP_4	sum of the VSA (van der Waals surface area) contributions of all of the atoms assigned to logP of bin 4
SpDiam_EA(dm)	spectral diameter from edge adjacency mat, weighted by dipole moment
P_VSA_MR_2	sum of the VSA (van der Waals surface area) contributions of all of the atoms assigned to molar refractivity of bin 2
GATS7i	Geary autocorrelation of path length 7 weighted by ionization potential
F03[C-S]	Number of C-S pair at topological distance of 3
CATS2D_00_DD	number of D-type-atom pair (O in −OH, N in NH or NH_2) at a topological distance of zero
DISPp	the distance between the geometric center and the polar center
SpMin5_Bh(m)	it is the fifth of the first eight smallest negative eigenvalues of Burden matrix weighted by mass

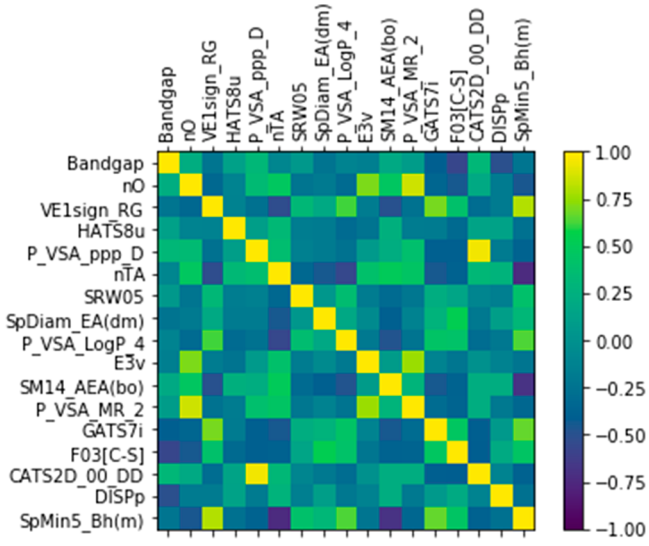


Figure 4. Correlation analysis matrix plot of the 16 selected features and band gap.

negative correlation with band gap. F03[C-S] stands for the number of C-S pairs at a topological distance of 3 in the monomeric unit. To design the polymer with a higher band gap, it is feasible to appropriately decrease the block number of CS and C_4H_2S or reduce the carbon-containing blocks in the meta-position of the two groups. DISPp stands for the distance between the geometric center and the polar center, which may reveal the relationship between the band gap and geometric structure as well as the block property. The band gap could be

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

kept in control by adjusting the symmetry and polarity of the building blocks in the monomer. Combined with the data set, some patterns are concluded. The symmetry of the monomer may have a great contribution to the band gap. $CH_2-CH_2-CH_2-CH_2$, whose geometric center coincides with the polar center, has the highest band gap with 8.823 eV. In asymmetric monomers, the presence of CS and CO will lead to the deviation of the polarity center and further reduce the band gap, which may be reflected in $CH_2-CO-CH_2-CS$ with a band gap of 2.174 eV. The replacement of CS or CO with O could effectively reduce this deviation, and CH_2-O-CH_2-CS has a higher band gap with 4.348 eV. For monomers with ring blocks like C_6H_4 and C_4H_2S, C_4H_2S may contribute better to the band gap increase because the conjugacy of the monomer could limit the band gap. These patterns could indicate a direction for the experimenters to design polymers with desired band gaps.

In addition to the correlation between the target and selected features, the linear dependence among features is also shown in the plot. We can find that there is a strong positive correlation between CATS2D_00_DD and P_VSA_ppp_D, which could be explained, as these features are calculated on the foundation of a D-type atom pair (O in −OH, N in NH or NH_2), resulting in the same information on linear independence. Similarly, the strong negative correlation between SpMin5_Bh(m) and SpDiam_EA(dm) exists because they are both obtained based on the adjacency matrix of monomers. However, it is surprising that SpMin5_Bh(m) has a strongly negative correlation with nTA, even though there are different types in Dragon. SpMin5_Bh(m) and nTA stand for the fifth of the first eight smallest negative eigenvalues of Burden matrix weighted by mass and total atom number, respectively. This strong correlation indicates that the dispersion of the eigenvalues of the adjacency matrix may increase with the total atom number.

The Pearson correlation matrix plot could reflect the linear correlation between the features and the target, while mRMR uses mutual information between features and target to sort the importance of features, reflecting the nonlinear correlation. There is no strong agreement between the trends in Pearson correlation and mutual information, but it can be noticed that features having strong Pearson correlation with the band gap are ranked relatively lower in mRMR and features ranking high have weak Pearson correlation with the target. This may indicate that the nonlinear relationship between the features and band gap plays a crucial part in the model prediction.

3.5. SVR Model for Predicting the Band Gap of a Polymer. Based on the results of feature selection, 16 selected features were adopted to construct the model for predicting the band gap of a polymer. In order to select an appropriate modeling algorithm, three machine learning methods including partial least squares (PLS), multiple linear regression (MLR), and support vector regression (SVR) were considered to evaluate different machine learning algorithms with R^2 and RMSE based on the LOOCV. Since SVR is a kernel-based regression, the proper kernel function is the key part for the SVR algorithm. The performance of the SVR model with different kernel functions was also taken into account. As shown in Table 2, SVR with the Gaussian kernel function was the most appropriate algorithm for model construction. In the Gaussian kernel function, the three parameters, ε-insensitive loss function, capacity parameter C, and gamma value, have a significant influence on the performance of the SVM model. In order to further optimize the regression model with the most generalization ability, these parameters were optimized by conducting the tree of Parzen estimator (TPE) algorithm[52] with python and evaluated by the RMSE of LOOCV results. The key codes could be found in scikit-learn. It was found in Figure 5a that the lowest RMSE was 0.485 when the optimal C, ε, and the gamma of the radial basis function were 18.896, 0.098, and 0.026, respectively.

Table 2. Determination Coefficients (R^2) and Root-Mean-Square Error (RMSE) of the Polymer Band Gap in LOOCV of PLS, MLR, and SVR with Different Kernel Functions

regression algorithm	R^2	RMSE
PLS	0.491	0.698
MLR	0.503	0.695
SVR-linear kernel	0.424	0.715
SVR-Gaussian kernel	0.824	0.485
SVR-polynomial kernel	0.762	0.569

In order to guarantee the diversity of the model after hyperparameter optimization, both LOOCV and 5-fold cross-validation of the training data set were carried out to evaluate the performance of the SVM model obtained. Parts b and c of Figure 5 show the plots of the predicted values versus the DFT calculation values of the band gap for polymer based on the LOOCV and 5-fold cross-validation of the training data set, respectively. There was little difference between the LOOCV and 5-fold cross-validation results, with R^2 of 0.824 and 0.802 and RMSE of 0.485 and 0.500, respectively.

It is reported that the R^2 of the independent test validation was 0.865 by using SVR combined with boosting.[34] However, they set "fingerprint" $\mathbf{M}_{III}$, a $7 \times 7 \times 7$ matrix composed of triplet blocks, as features, for which it is difficult to build the model with three-dimensional features. Our prediction results showed that R^2 of the independent test was 0.925 (shown in Figure 5d), which is 6% higher than the reference. Besides, we compared the predictions of nine samples with experimental values as well as DFT results. As shown in Figure 5e and Table 3, it could be found that the mean absolute error (MAE) between DFT calculated and experimental values is much lower than the others, demonstrating that the adopted DFT method is reliable and the errors of DFT calculations are also lower than the uncertainty of machine learning. The R^2 of DFT vs MODEL just drops slightly from 0.962 to 0.902 compared to that of EXP vs DFT, while the MAE drops by more than 6 times, from 0.260 to 1.743, which might be the cause of sample distribution. It could be noticed in Figure 5b that the sample size with band gap values higher than 5 eV is quite limited. The lack of data with high band gap values may cause the situation where the machine learning algorithm could not capture the information between selected features and band gap, indirectly leading to the large error of the model in the prediction of high band gap samples. The goodness of model predictions with DFT calculation as well as with actual experiments indicates that the model can instantly predict the properties with reasonable accuracies. Our model may also be applied to the prediction for arbitrary long polymers. To validate the feasibility of the conjecture, four-block (not contained in the data set) and even five-block and six-block polymer samples, 24 in total and 8 for each, were randomly designed to test the model predictions. After comparing the model predictions with the DFT calculations in Figure 5f and Table 4, it was found that increase of the block number of the polymer chain length

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

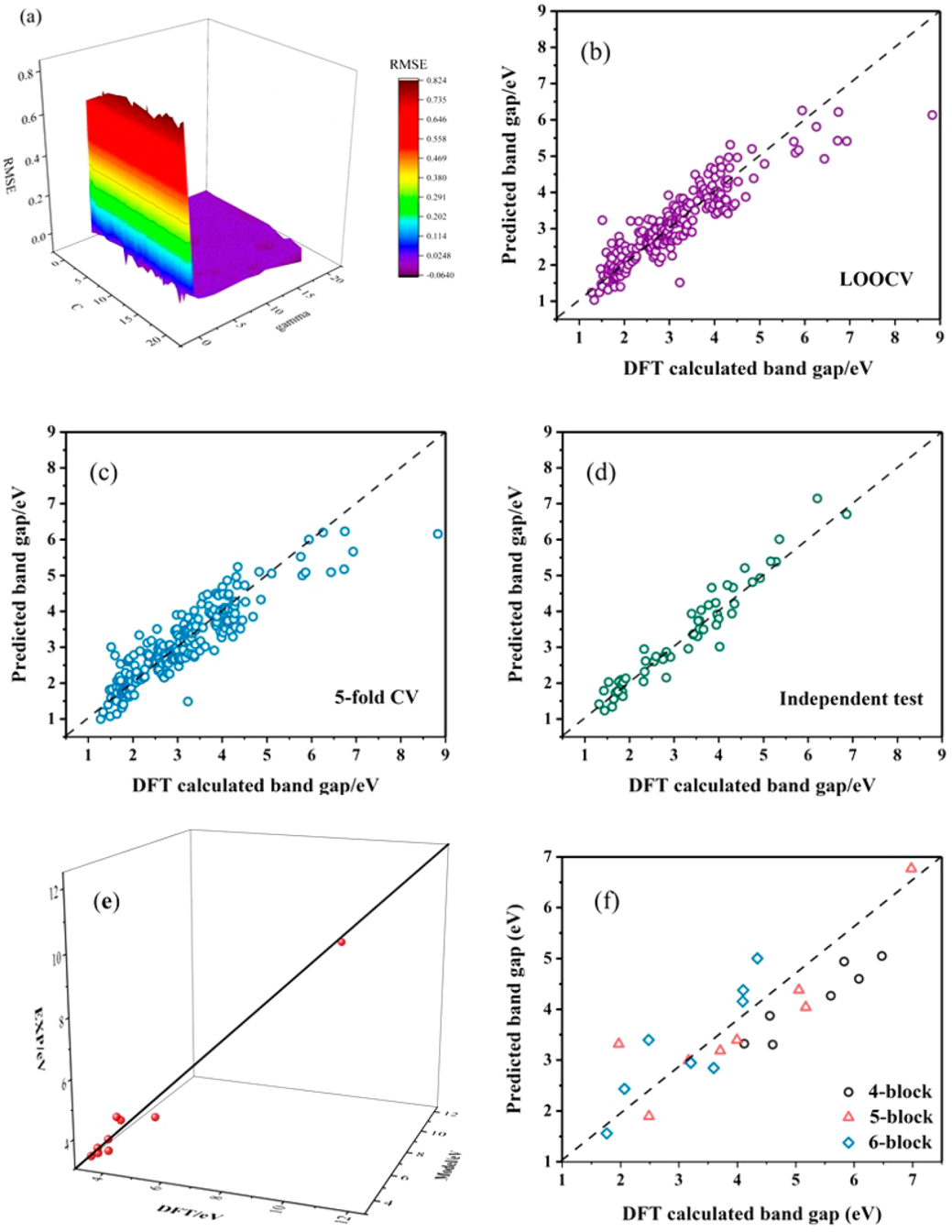


Figure 5. (a) RMSE of LOOCV vs gamma, ε, and C by using SVR with Gaussian kernel. (b) Predicted band gap vs DFT calculated band gap based on LOOCV. (c) Predicted band gap vs DFT calculated band gap based on 5-fold cross-validation. (d) Predicted band gap vs DFT calculated band gap based on independent test. (e) Predicted band gap vs DFT calculated band gap vs experimental band gap. (f) Predicted band gap vs DFT calculated band gap of designed polymer samples with different block numbers.

Table 3. Mean Absolute Error (MAE) and Determination Coefficient (R^2) of the Comparison between Experimental, DFT calculated, and Model Predicted Polymer Band Gaps

	EXP vs DFT	DFT vs MODEL	EXP vs MODEL
MAE	0.260	1.743	1.713
R^2	0.962	0.902	0.949

Table 4. Root-Mean-Square Error (RMSE) and Determination Coefficient (R^2) of the Designed Sample

	four-block	five-block	six-block
RMSE	1.164	0.760	0.520
R^2	0.850	0.692	0.605

limited the model prediction accuracy. According to the RMSE and R^2 of the designed sample, the constructed model can be not only applied to four-block samples but also applied to five-block samples. However, the R^2 of the six-block samples is lower than 0.65, which means the prediction of six-block polymer samples may have relatively big error.

3.6. SVC Model for Classifying the Polymer Band Gap. A material with a band gap higher or equal 2.3 eV is called a wide band gap semiconductor material.[53] Devices

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

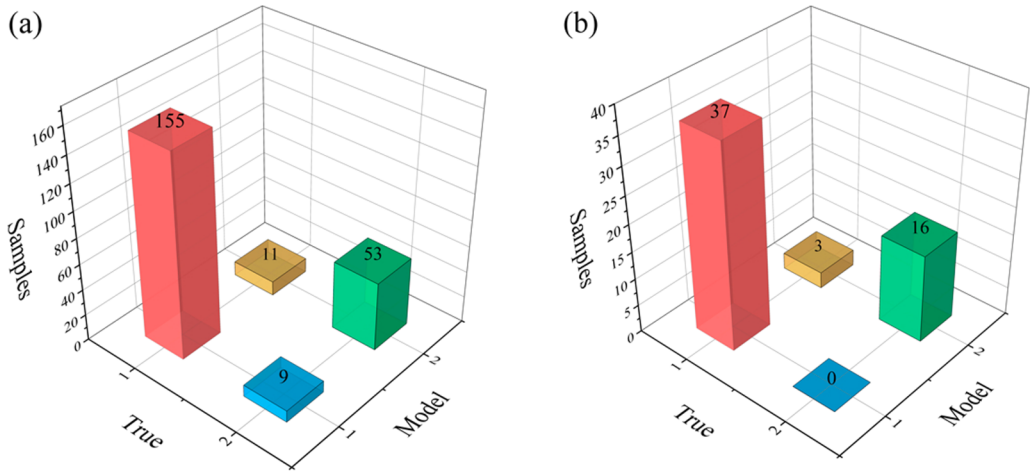


Figure 6. Confusion matrix for classification of band gap based on (a) LOOCV and (b) independent test.

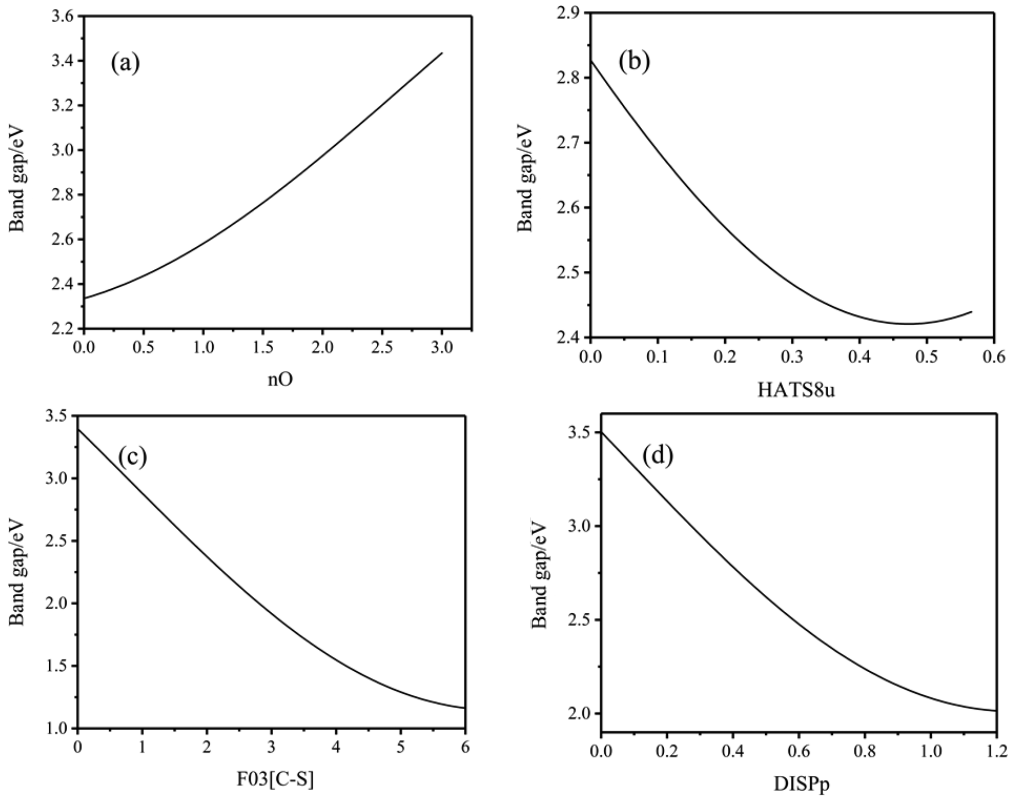


Figure 7. Sensitivity analysis of the (a) nO, (b) HATS8u, (c) F03[C-S], and (d) DISPp.

made up of this material could keep operating under extreme conditions. According to the standard above, polymer samples with a band gap higher or equal to 2.3 eV are defined as "positive" samples and those with a band gap less than 2.3 eV are defined as "negative" samples. The total data set includes 204 "positive" samples and 80 "negative" samples. The training set contains 228 samples, including 164 "positive" samples and 64 "negative" samples, and the test set contains 56 samples, including 40 "positive" samples and 16 "negative" samples.

The SVC model was constructed by using 16 selected features and a kernel function with the parameter $C = 8$ and gamma = 0.5. A resubstitution test is an examination for the self-consistency of a prediction method, which was performed in the current study. The rate of correct prediction of the resubstitution test for the specificity is 98.68%. Although the results of the resubstitution test were good, it is insufficient for evaluating the prediction method because the SVC classifier developed may be overfitting. Therefore, the LOOCV test and independent test were employed to validate the generalization and reliability of the classifier in this application. Shown in Figure 6, the total rates of correct prediction of the LOOCV test and independent test are 91.23 and 94.64%, respectively. More importantly, it is of great significance to guarantee the correlation between selected features and band gap when trying to figure out the hiding information in the data by feature analysis.

Fortunately, the good performances of both the SVR and SVC models could ensure this high correlation.

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

3.7. Model Explanation. *3.7.1. Sensitivity Analysis.* Sensitivity analysis (SA) is the study of how the change in the model can be quantitatively or qualitatively apportioned to the target property. The main purpose of sensitivity analysis is to assess whether the obtained results under the given conditions are sufficiently reliable when other conditions are not completely satisfied, which has been widely applied in many fields of data mining.[54] A good model should require an evaluation of confidence such as assessment of the uncertainties that related to the process and outcome of the model. Fortunately, SA offers a valid method for characterizing the uncertainty associated with the model. It can be used to examine the trend of a target variable depending on one of the features while the other features are kept constant. Different from Pearson correlation analysis, SA is the result of machine learning from the perspective of model, while the former is the result of data analysis from the perspective of data. Figure 7 illustrates the sensitivity analysis for the top 4 selected features, nO, HATS8u, F03[C-S], and DISPp. The results of SA and Pearson correlation are consistent with those of F03[C-S] and DISPp, both of which have a negative impact on the band gap. The nO has a positive impact on the band gap, which means the band gap increases as the number of O atoms increases in the repeating units when the values of other features are fixed. It is shown in the figure that HATS8u seems to have a weak negative linear correlation with the polymer band gap. However, it can be seen in the sensitivity analysis that the band gap decreases with the increase of HATS8u and then gradually increases after reaching the minimum point. HATS8u is a kind of GETAWY (GEomertry, Topology, and Atom-Wights AssemblY) descriptor based on the molecular influence matrix and the topological information by molecular graph.[55] HATS8u means the leverage-weighted topological structure autocorrelation of path length 8 with the mathematical formula shown below

$$\mathrm{HATS8u} = \sum_{i=1}^{\mathrm{nTA}-1} \sum_{j>i} h_{ii} h_{jj} \delta(8;\, d_{ij})$$

where nTA is the number of molecule atoms; h_{ii} and h_{jj} are the leverages of the two considered atoms; d_{ij} is the topological distance between atoms i and j; $\delta(8; d_{ij})$ is a Dirac-delta function ($\delta = 1$ if $d_{ij} = 8$ and zero otherwise). Through the analysis of the data set, it was found that the repeating unit samples with HATS8u equal to 0 tend to be made up of small blocks. For example, $CH_2-CH_2-CH_2-CH_2$ has no atom pairs with the path length of 8, resulting in the zero value of the HATS8u. As the value of HATS8u continues to increase, it is reflected in the repeating unit that the number of ring structural blocks could exist and keep increasing. The conjugation effect brought by the ring structure would lead to a decrease in the band gap. However, the steric effect might occur and restrain the electron transport when the repeating unit contains too many ring structural blocks. $C_6H_4-C_4H_2S-C_4H_2S-C_4H_2S$ shows a higher HATS8u value than $C_6H_4-CS-C_4H_2S-C_4H_2S$, which are 0.367 and 0.319, respectively, while the band gap of the former is lower than that of the latter (1.73 and 2.32 eV).

3.8. Model Application. With the aim of helping experimental scientists to utilize the model in designing new polymer structures with desired band gap, an online server was developed to predict the polymer band gap value using the constructed SVR model. In the process of applying the model, the user needs to input the values of 16 selected features and click the "Predict" button to obtain the band gap, which is very helpful for the experimenters to design a new polymer with a targeted band gap. After designing the composition of the monomer, the model could be used to estimate the band gap of the polymer. The online web server to share the available model for the prediction of polymer band gap is accessible at http://luktian.cn/polymer2019/.

Besides, several polymer samples consisting of different block numbers were designed to meet the required band gap of 5.0 eV. All of these samples were designed based on the patents described in the feature correlation analysis. The band gap values were validated by both DFT calculation and model prediction. It is expected that the designed samples would be instructive for the experiments. The repeating units of these samples were shown in Table 5.

Table 5. Predicted and DFT Calculated Band Gap Values of Designed Polymers

repeating units	model predicted band gap	DFT calculated band gap
$O-CH_2-CH_2-C_6H_4$	4.675 eV	5.049 eV
$CH_2-O-CH_2-C_6H_4-OH$	5.055 eV	4.380 eV
$O-CH_2-CH_2-C_6H_4-NH$	5.172 eV	4.041 eV
$CH_2-O-CH_2-CH_2-CH_2-C_6H_4$	4.343 eV	5.001 eV

4. CONCLUSION

In this work, we constructed the SVR and SVC models for predicting band gaps of polymers based on DFT computation and the descriptors generated from Dragon. The SVR model with Gaussian kernel function, $C = 18.896$, and gamma = 0.098 appeared to perform the best with R^2 and RMSE of 0.824 and 0.485 in LOOCV, respectively. It was found that the model was successful in the prediction of polymer band gap in a fast and convenient way, which was shared through an online web server. The model available can be used to screen out the polymers with targeted band gaps before experiments, which is very helpful for rapid design of new polymers. Besides, we have figured out some pattern between the selected features and polymer band gap through correlation analysis and sensitivity analysis to get a better understanding the relationship between the features and the target, which could be applied to design polymers with targeted band gap.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c08674.

Tables of the repeating units with experimental and DFT calculated band gaps (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Wencong Lu − *Materials Genome Institute, Shanghai University, and Shanghai Materials Genome Institute, Shanghai 200444, China; Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China;* orcid.org/0000-0001-5361-6122

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611

Authors

Pengcheng Xu − *Materials Genome Institute, Shanghai University, and Shanghai Materials Genome Institute, Shanghai 200444, China*

Tian Lu − *Materials Genome Institute, Shanghai University, and Shanghai Materials Genome Institute, Shanghai 200444, China*

Lifei Ju − *Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*

Lumin Tian − *Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*

Minjie Li − *Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c08674

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the National Key Research and Development Program of China (2016YFB0700504), Science and Technology Commission of Shanghai Municipality (18520723500).

■ REFERENCES

(1) Wang, Y.; Feng, L.; Wang, S. Conjugated Polymer Nanoparticles for Imaging, Cell Activity Regulation, and Therapy. *Adv. Funct. Mater.* **2019**, *29*, 1806818.

(2) White, B. T.; Long, T. E. Advances in Polymeric Materials for Electromechanical Devices. *Macromol. Rapid Commun.* **2019**, *40*, 1800521.

(3) Ma, Z.; Chen, P.; Cheng, W.; Yan, K.; Pan, L.; Shi, Y.; Yu, G. Highly Sensitive, Printable Nanostructured Conductive Polymer Wireless Sensor for Food Spoilage Detection. *Nano Lett.* **2018**, *18*, 4570−4575.

(4) Dickinson, E. Biopolymer-based particles as stabilizing agents for emulsions and foams. *Food Hydrocolloids* **2017**, *68*, 219−231.

(5) Benbettaieb, N.; Tanner, C.; Cayot, P.; Karbowiak, T.; Debeaufort, F. Impact of functional properties and release kinetics on antioxidant activity of biopolymer active films and coatings. *Food Chem.* **2018**, *242*, 369−377.

(6) Palza, H.; Zapata, P. A.; Angulo-Pineda, C. Electroactive Smart Polymers for Biomedical Applications. *Materials* **2019**, *12*, 277.

(7) Hill, I. R.; Andrukaitis, E. E. Lithium-ion polymer cells for military applications. *J. Power Sources* **2004**, *129*, 20−28.

(8) Katunin, A.; Krukiewicz, K.; Catalanotti, G. Modeling and synthesis of all-polymeric conducting composite material for aircraft lightning strike protection applications. *Mater. Today: Proc.* **2017**, *4*, 8010−8015.

(9) Cao, Q.; Tian, X.; You, H. Electrohydrodynamics in nano-channels coated by mixed polymer brushes: effects of electric field strength and solvent quality. *Modell. Simul. Mater. Sci. Eng.* **2018**, *26*, 035003.

(10) Yamane, T.; Todoroki, A.; Fujita, H.; Kawashima, A.; Sekine, N. Electric current distribution of carbon fiber reinforced polymer beam: analysis and experimental measurements. *Adv. Compos. Mater.* **2016**, *25*, 497−513.

(11) Budkov, Y. A.; Kolesnikov, A. L.; Kiselev, M. G. Communication: Polarizable polymer chain under external electric field in a dilute polymer solution. *J. Chem. Phys.* **2015**, *143*, 201102.

(12) Jung, J.-H.; Lee, D.-M.; Kim, J.-H.; Yu, C.-J. Circularly polarized electroluminescence by controlling the emission zone in a twisted mesogenic conjugate polymer. *J. Mater. Chem. C* **2018**, *6*, 726−730.

(13) Hou, W.; Xiao, Y.; Han, G.; Lin, J.-Y. The Applications of Polymers in Solar Cells: A Review. *Polymers* **2019**, *11*, 143.

(14) Mosciatti, T.; Del Rosso, M. G.; Herder, M.; Frisch, J.; Koch, N.; Hecht, S.; Orgiu, E.; Samori, P. Light-Modulation of the Charge Injection in a Polymer Thin-Film Transistor by Functionalizing the Electrodes with Bistable Photochromic Self-Assembled Monolayers. *Adv. Mater.* **2016**, *28*, 6606−11.

(15) Co, T. T.; Tran, T. Q.; Le, H. V.; Ho, V. A. P.; Tran, L. D. Band Gap, Molecular Energy and Electrochromic Characterization of Electrosynthesized Hydroxymethyl 3,4-Ethylenedioxythiophene. *J. Electron. Mater.* **2017**, *46*, 1669−1673.

(16) Arı, H.; Büyükmumcu, Z. Comparison of DFT functionals for prediction of band gap of conjugated polymers and effect of HF exchange term percentage and basis set on the performance. *Comput. Mater. Sci.* **2017**, *138*, 70−76.

(17) Jorgensen, P. B.; Mesta, M.; Shil, S.; Garcia Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **2018**, *148*, 241735.

(18) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613−23.

(19) Peng, S. P.; Zhao, Y. Convolutional Neural Networks for the Design and Analysis of Non-Fullerene Acceptors. *J. Chem. Inf. Model.* **2019**, *59*, 4993−5001.

(20) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495−6502.

(21) St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for highthroughput polymer screening. *J. Chem. Phys.* **2019**, *150*, 234111.

(22) Meredig, B. Industrial materials informatics: Analyzing large-scale data to solve applied problems in R&D, manufacturing, and supply chain. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 159−166.

(23) Ward, L.; Aykol, M.; Blaiszik, B.; Foster, I.; Meredig, B.; Saal, J.; Suram, S. Strategies for accelerating the adoption of materials informatics. *MRS Bull.* **2018**, *43*, 683−689.

(24) Balachandran, P. V.; Xue, D.; Theiler, J.; Hogden, J.; Lookman, T. Adaptive Strategies for Materials Design using Uncertainties. *Sci. Rep.* **2016**, *6*, 19660.

(25) Lu, W.; Xiao, R.; Yang, J.; Li, H.; Zhang, W. Data mining-aided materials discovery and optimization. *J. Materiomics* **2017**, *3*, 191−201.

(26) Hu, B.; Lu, K.; Zhang, Q.; Ji, X.; Lu, W. Data mining assisted materials design of layered double hydroxide with desired specific surface area. *Comput. Mater. Sci.* **2017**, *136*, 29−35.

(27) Shi, L.; Chang, D.; Ji, X.; Lu, W. Using Data Mining to Search for Perovskite Materials with Higher Specific Surface Area. *J. Chem. Inf. Model.* **2018**, *58*, 2420−2427.

(28) 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.

(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) 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.

(31) 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.

(32) 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.

(33) 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.
(34) Mannodi-Kanakkithodi, A.; Pilania, G.; Ramprasad, R. Critical assessment of regression-based machine learning methods for polymer dielectrics. *Comput. Mater. Sci.* **2016**, *125*, 123−135.
(35) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **2019**, *5* (1), 66.
(36) Yang, B.; Liu, X.-D.; Xu, H.; Zheng, Y.; Lu, P.; Yu, J.-S.; Ma, Y.-G.; FENG, J.-K. Effect of Connecting Form of Repeat Unit in Polyphenylene Derivatives Conjugated Polymers on Band Gap. *Wuli Huaxue Xuebao* **2006**, *22*, 962−966.
(37) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon software: An easy approach to molecular descriptor calculations. *MATCH* **2006**, *56*, 237−248.
(38) Fan, T.; Sun, G.; Zhao, L.; Cui, X.; Zhong, R. QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds. *Int. J. Mol. Sci.* **2018**, *19*, 3015.
(39) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
(40) Yahyaoui, M.; Bouchama, A.; Anak, B.; Chiter, C.; Djedouani, A.; Rabilloud, F. Synthesis, molecular structure analyses and DFT studies on new asymmetrical azines based Schiff bases. *J. Mol. Struct.* **2019**, *1177*, 69−77.
(41) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−8.
(42) Urbanowicz, R. J.; Meeker, M.; La Cava, W.; Olson, R. S.; Moore, J. H. Relief-based feature selection: Introduction and review. *J. Biomed. Inf.* **2018**, *85*, 189−203.
(43) Vinh, L. T.; Lee, S.; Park, Y.-T.; d'Auriol, B. J. A novel feature selection method based on normalized mutual information. *Appl. Intell.* **2012**, *37*, 100−120.
(44) Ramírez-Gallego, S.; Lastra, I.; Martínez-Rego, D.; Bolón-Canedo, V.; Benítez, J. M.; Herrera, F.; Alonso-Betanzos, A. Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data. *Int. J. Intell.t Syst.* **2017**, *32*, 134−152.
(45) Sakar, C. O.; Kursun, O.; Gurgen, F. A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy−Maximum Relevance filter method. *Expert Systems with Applications* **2012**, *39*, 3432−3437.
(46) Cortes, C.; Vapnik, V. Support-vector networks. *Machine Leaming* **1995**, *20*, 273−297.
(47) Tomar, D.; Agarwal, S. Twin Support Vector Machine: A review from 2007 to 2014. *Egypt. Inform. J.* **2015**, *16*, 55−69.
(48) Ding, S.; Zhang, N.; Zhang, X.; Wu, F. Twin support vector machine: theory, algorithm and applications. *Neural Comput. Appl.* **2017**, *28*, 3119−3130.
(49) Okwuashi, O.; Ndehedehe, C. Tide modelling using support vector machine regression. *J. Spat. Sci.* **2016**, 1−18.
(50) De Oliveira, M. A.; Dos Santos, H. l. F.; De Almeida, W. B. Structure and electronic properties of cyanothiophene derivatives: A theoretical ab initio and DFT study. *Int. J. Quantum Chem.* **2002**, *90*, 603−610.
(51) Ling, L.; Lagowski, J. B. DFT study of electronic band structure of alternating triphenylamine-fluorene copolymers. *Polymer* **2013**, *54*, 2535−2543.
(52) Ghanbari-Adivi, F.; Mosleh, M. Text emotion detection in social networks using a novel ensemble classifier based on Parzen Tree Estimator (TPE). *Neural Comput. Appl.* **2019**, *31*, 8971−8983.
(53) Newhouse, P. F.; Hersh, P. A.; Zakutayev, A.; Richard, A.; Platt, H. A. S.; Keszler, D. A.; Tate, J. Thin film preparation and characterization of wide band gap Cu3TaQ4 (Q = S or Se) p-type semiconductors. *Thin Solid Films* **2009**, *517*, 2473−2476.
(54) Yun, W.; Lu, Z.; Jiang, X. An efficient sampling approach for variance-based sensitivity analysis based on the law of total variance in the successive intervals without overlapping. *Mech. Syst. Signal Pr.* **2018**, *106*, 495−510.
(55) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.

https://dx.doi.org/10.1021/acs.jpcb.0c08674
J. Phys. Chem. B 2021, 125, 601−611