Application of Machine Learning in Developing Quantitative Structure−Property Relationship for Electronic Properties of Polyaromatic Compounds

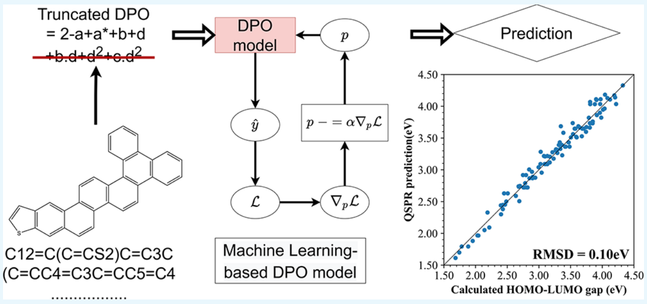
Tuan H. Nguyen, Lam H. Nguyen, and Thanh N. Truong*

Cite This: *ACS Omega* 2022, 7, 22879−22888

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: The degree of π orbital overlap (DPO) model has been demonstrated to be an excellent quantitative structure−property relationship (QSPR) that can map two-dimensional structural information of polycyclic aromatic hydrocarbons (PAHs) and thienoacenes to their electronic properties, namely, band gaps, electron affinities, and ionization potentials. However, the model suffers from significant limitations that narrow its applications due to inefficient manual procedures in parameter optimization and descriptor formulation. In this work, we developed a machine learning (ML)-based method for efficiently optimizing DPO parameters and proposed a truncated DPO descriptor, which is simple enough that can be automatically extracted from simplified molecular-input line-entry system strings of PAHs and thienoacenes. Compared with the result from our previous studies, the ML-based methodology can optimize DPO parameters with four times fewer data, while it can achieve the same level of accuracy in predictions of the mentioned electronic properties to within 0.1 eV. The truncated DPO model also has similar accuracy to the full DPO model. Consequently, the ML-based DPO approach coupled with the truncated DPO model enables new possibilities for developing automatic pipelines for high-throughput screening and investigating new QSPR for new chemical classes.



1. INTRODUCTION

Applications of machine learning techniques to research in chemical and related fields have been gaining attention recently. Some novel applications include generating drug candidates,[1] investigating chemical phenomena,[2] and assisting theoretical calculation.[3−8] One of the most prominent tasks for applying machine learning is physical or chemical property predictions. In this direction, quantitative structure−activity relationship (QSAR) specializes in the prediction of the biological activity of compounds from their structural information.[9−12] Similarly, materials scientists employ a similar technique called quantitative structure−property relationship (QSPR) for predicting various properties of materials from their 2D or 3D structural data.[13−17] In most QSPR and QSAR methodologies, predicting tasks heavily depend on a set of descriptors,[9,13] which serve as numerical representations of structural information of molecules. Typically, these descriptors are numerical objects obtained by transforming raw molecular data by some predefined procedure. These descriptors can be as simple as a list of molecule's compositions[5,6] or as complex as matrices,[18−20] fingerprints,[15,16,21] etc.[8,14,22−26]

Recently, representation learning[27] has been introduced to harness deep learning concepts for actively refining the process of molecular representation extraction through learning. For instance, neural fingerprints[28] are modeled after the extended connectivity fingerprints (ECFPs).[21] The ECFPs are designed to solely represent each molecular fragment as uniquely as possible. On the other hand, neural fingerprints replace some operations in ECFP generation with learnable modules that actively configured themselves during training to optimally represent molecular fragments for a certain task. As such, fingerprints of fragments are similar if the learned model can recognize the similarity between them. Hence, the representation has more expressiveness and interpretability. This approach was recently featured in multiple publications and yielded excellent results.[29−32]

The degree of π orbital overlap (DPO) model is a quantitative structure−property relationship (QSPR) model for predicting electronic properties of polyaromatic hydrocarbon (PAH) and thienoacene compounds.[33,34] The DPO

Received: April 28, 2022
Accepted: May 25, 2022
Published: June 17, 2022

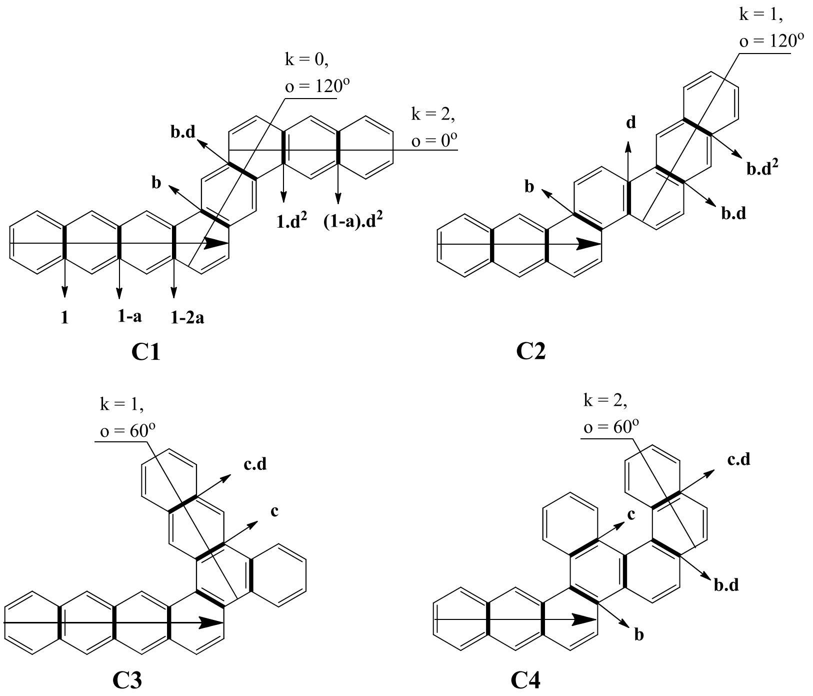


Figure 1. Examples of assigning DPO values to different fused bonds. The overlayer order k and orientation o are also given for some segments.

model relies on a representation or descriptor called the degree of π orbital overlap (DPO). Based on a quantum mechanical physical model, this descriptor represents a PAH or thienoacene molecule by a polynomial of six nonzero parameters, each associating the contribution of a topological trait of a polyaromatic hydrocarbon (PAH) compound to its final electronic properties. This has been proven to be rather accurate as it can accurately predict electronic properties, namely, electron affinities, ionization potentials, and band gaps to within 0.1 eV.[33,34] The original parameter optimization schedule is based on the prerecognition that each of them associates with a structural feature, which suggests a specific data set for optimizing it by trial and error. The described optimization procedure gives rise to several setbacks. First, it requires a large number of data points to calibrate the model, yet the final set of parameters may not be optimal globally. Second, the application of the DPO model to other chemical classes is not trivial. Finally, to date, determination of the DPO value of each molecule is done manually, and thus, it is unrealistic to use for high-throughput screening of massive databases.

In this work, we implement and assess the performance of a learning procedure for developing a DPO model for a class of molecules. The goal of this new procedure is to provide an automated pipeline for parameter optimization in order to remove the number of setbacks of our previous works as mentioned above. This is motivated by recognizing that DPO descriptors can be treated as a learnable representation, which can be done by applying various learning principles and techniques such as empirical risk minimization, gradient descent, and backpropagation. Furthermore, toward the goal of creating an automatic pipeline for the DPO model, we devised a truncated DPO model. The analysis of results in the optimization of the DPO descriptors reveals that several terms in the DPO descriptor can be neglected with little consequence, thus suggesting a truncated DPO model. With the truncated DPO model, we demonstrate an automated pipeline that uses simplified molecular-input line-entry system (SMILES) strings as input for molecular representations that can be employed for high-throughput screening. Assessments for the accuracy and applicability of the truncated DPO model in predicting the electronic properties of PAHs and thienoacenes are then provided.

2. METHODS

2.1. DPO Models and Descriptors. *2.1.1. DPO Models.*

A DPO model is a physics-based QSPR model for predicting electronic properties of polyaromatic compounds such as PAHs and thienoacenes. It is based on a simple particle in the 2D box quantum mechanical model to connect structural information to its physical properties related to its energy levels such as its electron affinity (EA), ionization potential (IP), and highest occupied molecular orbital (HOMO)−lowest unoccupied molecular orbital (LUMO) gap. Technically, input structural information can be represented by 2D structures of polyaromatic compounds. The procedure of the model can be described briefly as follows. First, from a polyaromatic compound's structure, a polynomial of six preoptimized nonzero parameters is evaluated to obtain the structure's DPO value. This step is done manually by following a set of structural descriptive rules. The DPO value is subsequently used in QSPR linear equations to obtain the predicted properties. Three distinct linear equations of the form $y = wx + w_b$, in which x is the DPO value, y is the property, and w and w_b are weights, correspond to three modeled electronic properties mentioned above. In this work, the term DPO descriptor is referred to the polynomial, as it reflects the topological structure value of the polyaromatic compound. The parameters that associate with these DPO descriptors are called DPO parameters.

2.1.2. Rules for Determining DPO Descriptors. A brief explanation of the DPO rules is presented here to introduce some important features of the rules. Complete descriptions of the DPO rules with examples are given in our previous

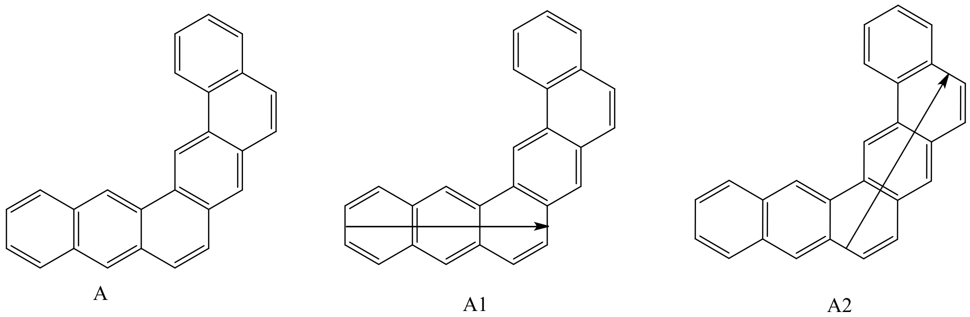


Figure 2. Illustrations of computing truncated DPO values for the case of multiple longest segments. Molecule A has the two longest segments shown as arrows in A1 and A2. The truncated DPO of A is the average of DPO values when A1 and A2 are chosen as the reference segment, i.e., $\mathrm{DPO_A} = \frac{\mathrm{DPO_{A1}} + \mathrm{DPO_{A2}}}{2}$.

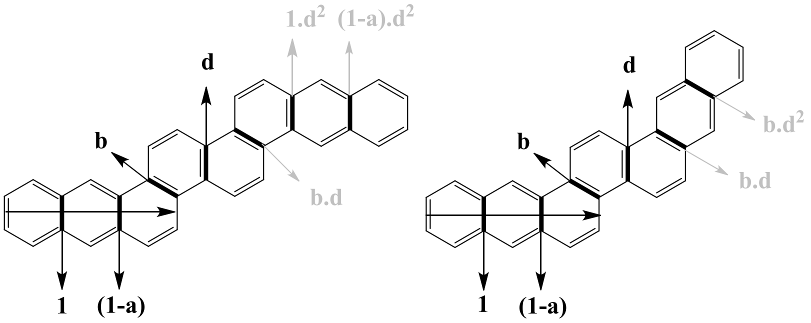


Figure 3. Illustration on the assignment of DPO values to each fused bond; terms in gray are neglected in the truncated DPO model.

works.[33,34] In this study, we explain how the DPO descriptor captures topological structural information and how the truncated DPO descriptor arises.

The goal of the DPO rule is to assign each fused bond in the polyaromatic compound a component in the DPO polynomial suggested by the use of an effective 2D particle-in-a-box quantum mechanical model. Fused bonds of the same segment are assigned similarly depending on the topological position of the segment in the molecule. The topological position can be defined with respect to the reference segment, which can be uniquely determined via a set of rules. Thus, the reference segment needs to be determined first and is generally the longest segment. With reference to this segment, the topological position of any segment is specified by the orientation and the overlayer order. In terms of orientations of different segments in a PAH, a segment can be parallel, forming an angle of 120°, or forming an angle of 60° to the reference segment. For each of those segments, the parameters a, b, and c are used in assigning value to its fused bonds. The overlayer order of a segment represents its distance to the reference segment orthogonally. As such, the farther the segment is above or below the reference segment, the larger the order is. The sum of values assigned to fused bonds of such segments is scaled by a factor of d raised to the power of the overlayer order k. In summary, supposing o and k are respectively the relative orientation and the overlayer order of a segment, the equation below is the contribution of a given segment to the overall DPO value and describes how fused bonds in that segment are assigned value.

$$g_{\text{segment}}(o, k) = \begin{cases} [1 + (1 - a) + (1 - 2a) + ...]d^{k}, \ o = 0° \\ [b + bd^{1} + bd^{2} + ...]d^{k}, \ o = 120° \\ [c + cd^{1} + cd^{2} + ...]d^{k-1}, \ o = 60° \end{cases} \tag{1}$$

Figure 1 illustrates examples for assigning DPO values to fused bonds using the above equation. For compound C1, the reference segment is marked with an arrow across. Its fused bonds are assigned as the first case of the equation with $k = 0$. The segment that forms with the reference segment an angle of 120° is assigned to the second case with $k = 0$. Finally, the uppermost segment that is parallel with the reference segment is two orders above the reference; therefore, the first case of the equation is used with $k = 2$. For compound C2, consider the upper rightmost segment that forms with the reference segment (marked with an arrow) an angle of 120°. Since it stems from the parallel segment that is one order above the reference segment, the second case of the equation is used with $k = 1$ assigned to this segment. For similar cases where there are segments that form an angle of 60° with the reference one, refer to compounds C3 and C4.

Lastly, the parameters a^* and d^* are used whenever thiophene rings are in the segment. The DPO value of a structure is the sum of DPO values of all segments with the assigned terms to all of its fused bonds.

2.1.3. Truncated DPO Descriptors.

In the DPO model, the contribution of each overlayer segment is scaled by d^k. Previous studies showed that the d parameter is between 0.2 and 0.3. This suggests that segments that are far from the reference segment can be assumed to have a negligible contribution and thus can be dropped from the total DPO

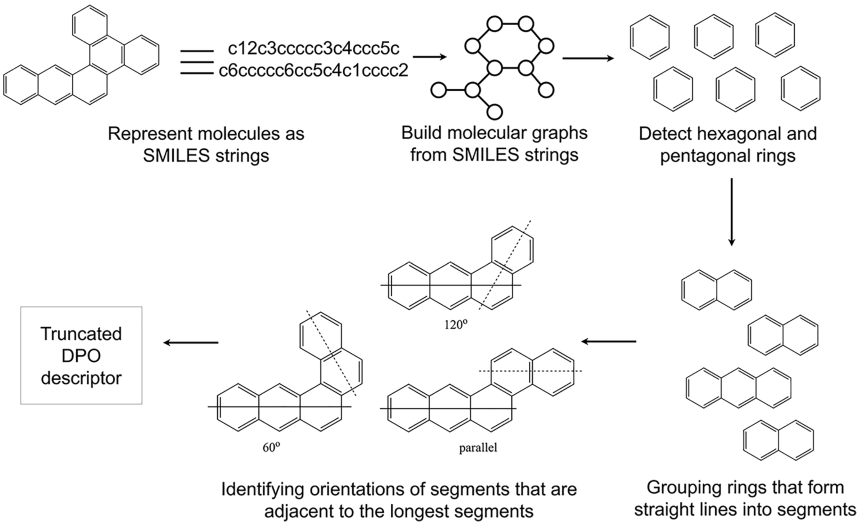

c12c3ccccc3c4ccc5c
c6ccccc6cc5c4c1cccc2
Represent molecules as SMILES strings
Build molecular graphs from SMILES strings
Detect hexagonal and pentagonal rings
Grouping rings that form straight lines into segments
120°
60°
parallel
Identifying orientations of segments that are adjacent to the longest segments
Truncated DPO descriptor

Figure 4. Flow chart illustrating the process of extracting truncated DPO descriptors from SMILES strings.

descriptor. The truncated DPO model presented here omits the contributions of several types of overlayer segments.

The truncated DPO inherits most of the original DPO assignment rules two new simplifications. First, if multiple segments have the same longest length, then there is no need to go through an elaborate process to determine the unique reference segment. Instead, the final DPO value is the average of all DPO values where each one of such segments is treated as the reference segment. Figure 2 illustrates this rule.

Second, only two types of overlayer segments are considered. One is the adjacent segments that have one ring in common with the reference segment. The other is segments that have one ring in common with the adjacent segments. Contributions from other overlayer segments are ignored. For instance, Figure 3 shows two different structures that have the same value of truncated DPO.

2.2. Generating Truncated DPO from SMILES Strings. In order to employ machine learning techniques and further expand the applicability of the DPO model to different chemical classes, we need to provide a means to extract structural information from polyaromatic molecules represented by the simplified molecular-input line-entry system (SMILES).[35] A SMILES string provides information regarding atoms and bonds of a molecule that can be used to create graph data structures. From there, recognizing groups of atoms that form rings can be done by employing a graph traversal algorithm. Rings that are in a straight line can be easily recognized and grouped into segments. Subsequently, the longest segments along with their adjacent segments and topology can be determined easily for the truncated DPO. The process is schematically illustrated in Figure 4.

2.3. Optimization of the DPO Model. In this work, instead of manually optimizing six DPO parameters for PAHs and thienoacenes one at a time as in our previous works, a machine learning technique is developed. The optimization scheme is based on the empirical risk minimization (ERM) principle.[36] The empirical risk refers to the error of the model's predictions for all training samples and is assessed by the so-called loss function, which is the mean square error (MSE) for regressing models. According to the ERM paradigm, the set of optimal parameters of the model is obtained by minimizing the value of the empirical risk and can be solved by the gradient descent algorithm. The first-order derivative of the loss function with respect to parameters is obtained by working backward from the loss function to the parameters using chain rules. Hence, this technique of finding derivatives is named backpropagation.[37] The optimization procedure is given step by step below.

In step 1, set $t = 0$. Initiate six nonzero DPO parameters $p^{[0]} = a^{[0]}, b^{[0]}, c^{[0]}, d^{[0]}, a*^{[0]}$, and $d*^{[0]}$ as zeros. We adopt the convention that the number inside the square bracket of the superscript of a parameter denotes the number of time steps or iterations, while the subscript denotes the general index of a molecule. Also, p is used to collectively denote six nonzero parameters a, b, c, d, $a*$, and $d*$.

In step 2, derive analytically the gradient of the DPO polynomial g_i with respect to the parameter p for each molecule in the training set.

$$\nabla_p g_i^{[t]} = \frac{\partial}{\partial p} g_i(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}, a^{*[t]}, d^{*[t]}) \tag{2}$$

In step 3, compute the DPO value $\mathrm{x}_i^{[t]}$ and its derivative with respect to p $\nabla_\mathrm{p}\mathrm{x}_i^{[t]}$ numerically using the set of parameters in the current time step. Note that we denote x_i as a numerical DPO value, while g_i denotes its polynomial.

$$x_i^{[t]} = g_i(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}, a^{*[t]}, d^{*[t]}) \tag{3}$$

$$\nabla_p x_i^{[t]} = \nabla_p g_i(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}, a^{*[t]}, d^{*[t]}) \tag{4}$$

In step 4, apply the least-square algorithm to determine $w^{[t]}$ and $\mathrm{w}_b^{[t]}$ from DPO value x_i's and the true value of the modeled property y_i's.

In step 5, compute the prediction $\hat{y}^{[t]}$ by plugging $x_\mathrm{i}^{[t]}$ into the linear equation (eq 5).

$$\hat{y}_i^{[t]} = w^{[t]} x_i^{[t]} + w_b^{[t]} \tag{5}$$

In step 6, compute the mean square error (MSE) loss function (eq 6) for all molecules in the training set.

$$\mathcal{L}^{[t]}(\hat{y}^{[t]}, y^{[t]}) = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i^{[t]} - y_i^{[t]})^2 \tag{6}$$

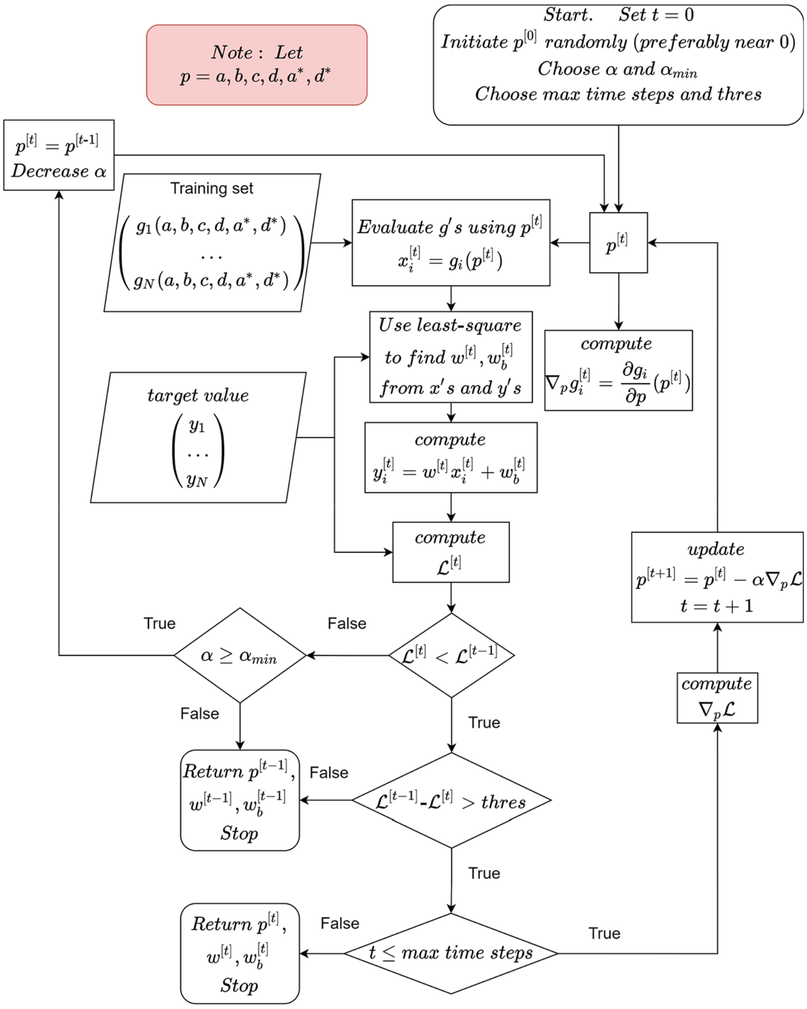

Note : Let
$p = a, b, c, d, a^*, d^*$
Start. Set $t = 0$
Initiate $p^{[0]}$ randomly (preferably near 0)
Choose α and α_{min}
Choose max time steps and thres
$p^{[t]} = p^{[t-1]}$
Decrease α
Training set
$g_1(a, b, c, d, a^*, d^*)$
...
$g_N(a, b, c, d, a^*, d^*)$
Evaluate g's using $p^{[t]}$
$x_i^{[t]} = g_i(p^{[t]})$
$p^{[t]}$
Use least-square
to find $w^{[t]}, w_b^{[t]}$
from x's and y's
compute
$\nabla_p g_i^{[t]} = \frac{\partial g_i}{\partial p}(p^{[t]})$
target value
y_1
...
y_N
compute
$y_i^{[t]} = w^{[t]} x_i^{[t]} + w_b^{[t]}$
compute
$\mathcal{L}^{[t]}$
update
$p^{[t+1]} = p^{[t]} - \alpha \nabla_p \mathcal{L}$
$t = t + 1$
True
False
$\alpha \geq \alpha_{min}$
$\mathcal{L}^{[t]} < \mathcal{L}^{[t-1]}$
False
True
compute
$\nabla_p \mathcal{L}$
Return $p^{[t-1]}$,
$w^{[t-1]}, w_b^{[t-1]}$
Stop
False
$\mathcal{L}^{[t-1]}$-$\mathcal{L}^{[t]} > thres$
True
Return $p^{[t]}$,
$w^{[t]}, w_b^{[t]}$
Stop
False
$t \leq max\ time\ steps$
True

Figure 5. Flow chart illustrating the ML-based DPO parameter optimization procedure.

In step 7, compute the gradient of the loss function with respect to each parameter by working backward using the chain rule:

$$\nabla_p \mathcal{L}^{[t]} = \frac{\partial \mathcal{L}^{[t]}}{\partial p^{[t]}} = \sum_{i=1}^{N} \frac{\partial \mathcal{L}^{[t]}}{\partial \hat{y}_i^{[t]}} \frac{\partial \hat{y}^{[t]}}{\partial g_i^{[t]}} \frac{\partial g_i^{[t]}}{\partial p^{[t]}} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i^{[t]} - y_i^{[t]}) w^{[t]} \nabla_p x_i^{[t]} \tag{7}$$

In step 8, update these parameters according to equations (eq 8).

$$p^{[t+1]} = p^{[t]} - \alpha \nabla_p \mathcal{L}^{[t]} \tag{8}$$

α is the adjustable learning rate that is 0.1 by default.

In step 9, increment $t = t + 1$ and repeat steps 3 through 6 to obtain a new loss value with the value of new parameters. If predefined conditions are not satisfied, then continue steps 7 through 9. Else, if predefined conditions are satisfied, then stop and return to the model with the optimized parameters.

Many conditions can be set in step 9 to mitigate training hardships and make the training process more automatic. For instance, to make the algorithm practical, a maximum number of cycles and the convergent threshold were introduced. The algorithm is succinctly illustrated in the form of a flow chart as in Figure 5.

The DPO model described above is called the machine learning (ML)-based DPO model. This model is implemented in the Python programming language. The model is built with various libraries. Numpy[38] is employed to rapidly carry out vector computation. Pandas[39] is employed for working with data sets. Sklearn[40] is used to rapidly deploy linear regression models. Sympy[41] is used for symbolic differentiation and rapid evaluation of symbolic expressions.

2.4. Data.

This work mainly concerns the electronic properties of polyaromatic compounds, namely, PAHs and thienoacenes. The PAH data is reused from our previous work on the DPO model for PAH.[34] The data sets consist of a training set containing PAH molecules of 3−6 rings and a test set consisting of PAH molecules of 7−8 rings. Similarly, taken from our previous work on the DPO model for thienoacene molecules[33] are two pairs of training and testing data for two classes of thienoacenes containing either one or two thiophene rings. Cumulatively, there are 248 data points, of which 132 of them are training instances and the remaining 116 data points are for testing.

The current methodology enables us to investigate the convergence and stability of the parameter optimization procedure via machine learning framework as functions of the data training size. To make the optimization procedure more robust, for each run, all 248 data points are freshly split into 132 data points for training and 116 data points for testing in a random and stratified manner. This is done by first binning all data points according to their bandgap values. The bin width is chosen to be 0.5 eV. The range of bandgap values starts from 1.5 eV and ends at 5.0 eV. Data points are then drawn from each bin to assemble the training set. The number of data points to be sampled from each bin is the size of the training set scaled by the ratio of the bin size and the total number of data points. Additional data points may be sampled randomly regardless of bins to meet the specific sampling size. This sampling method is referred to as stratified sampling throughout this work.

3. RESULTS AND DISCUSSION

3.1. Convergence Rate of the ML-Based DPO Models.

To examine the convergence rate of the ML-based DPO models, we fitted the DPO models against the bandgap property of samples in a full-size training set with various learning rate values. The mean square error (MSE) of the model for the HOMO−LUMO gap is monitored at each time step. The model is considered converged if the difference between its MSE before and after an update is less than 10^{-5} eV^2. The maximum number of time steps is 150.

Plots of the root-mean-square deviation (RMSD, square root of MSE) of models for the bandgap property as a function of the number of time steps are shown in Figure 6. Models trained with learning rate values of 0.01, 0.05, and 1.5 fail to reach convergence within the allotted number of time steps. With learning rates of 0.01 and 0.05, the model's error descends too slowly, and thus, it is unable to reach the converged error of 0.1 eV within 160 time steps. On the other hand, the error at a learning rate of 1.5 oscillates around 0.20 and 0.25 eV and is unable to converge. Models trained at learning rate values of 0.50, 1.00, and 1.25 take 43, 24, and 126 time steps to converge, respectively, while it is barely converged for models at a learning rate of 0.10. Note that at a learning rate of 1.25, the error exhibits some oscillation and does not monotonically decrease as observed for those between 0.1 and 1.0. These results suggest that the learning rate values should range between 0.10 and 1.0. For this study, a learning rate of 1.0 was used.

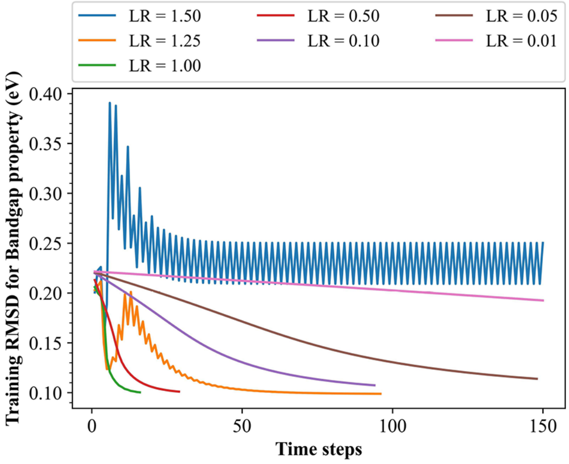


Figure 6. Plots of RMSD for the band gap of ML-based DPO models trained at different learning rates versus the number of time steps.

3.2. Training and Testing ML-Based DPO Models.

Here, we present the results of the ML-based full and truncated DPO models trained with full-size (132 data points) training sets and their accuracy assessed with the full-size test set (116 data points). In order to compare them with our previous works,[33,34] DPO parameters are fitted against the same density functional theory (DFT) calculation HOMO−LUMO gap values. These parameters are then used to obtained QSPR linear equations for predicting band gaps, EA, and IP properties.

The converged values for DPO parameters and parameters of the linear equation fitted against HOMO−LUMO gap values are presented in Table 1 along with those from our previous studies. The machine learning-based optimized parameters are rather close to those obtained using the manual optimizing process mentioned. The converged parameters for both the full and truncated DPO models are also quite close together.

Figures 7 and 8 show the linear correlations between the optimized DPO values and the quantum mechanical DFT-calculated values of the HOMO−LUMO gap, EA, and IP. Excellent linear correlations between the values of structural DPO and the physical electronic properties further confirm the physics behind the QSPR model. Figure 9 plots correlations between predictions from both ML-based full and truncated DPO models and the DFT-calculated values of electronic properties of molecules in the test set. Moreover, we repeat this experiment 20 times, each with freshly generated training data sets, and then average them over the optimized parameters and the root-mean-square deviation (RMSD) values. It is found that both the full and truncated DPO models converged to the average, and standard deviation values of the RMSD of all 20 runs are 0.10 ± 0.01, 0.07 ± 0.01, and 0.06 ± 0.00 eV for the HOMO−LUMO gap, EA, and IP, respectively. They are the same magnitude of errors compared to our previous studies and are within the accuracy range of the DFT level of theory, which was used to generate the data.

Two general problems in machine learning that models often encounter are underfitting, which is identified by the low training accuracy, and overfitting, which is recognized by a high training accuracy but low test accuracy. The results presented in Figure 9 and the average results of multiple test runs indicate that the ML-based DPO models do not suffer from either the overfitting or underfitting case. The former is no surprise since the DPO model has been based on a quantum

Table 1. Values of DPO Parameters and Linear Regression Weights for the ML-Based Full and Truncated DPO Models and Published Values

	DPO parameters						parameters of the linear equation	
	a	*b*	*c*	*d*	*a**	*d**	w_b	*w*
full DPO	0.08	−0.12	0.32	0.30	0.40	0.16	4.99	−0.77
truncated DPO	0.07	−0.13	0.36	0.28	0.41	0.16	4.99	−0.76
refs 33 and 34	0.05	−0.25	0.33	0.33	0.50	0.15	4.68	−0.65

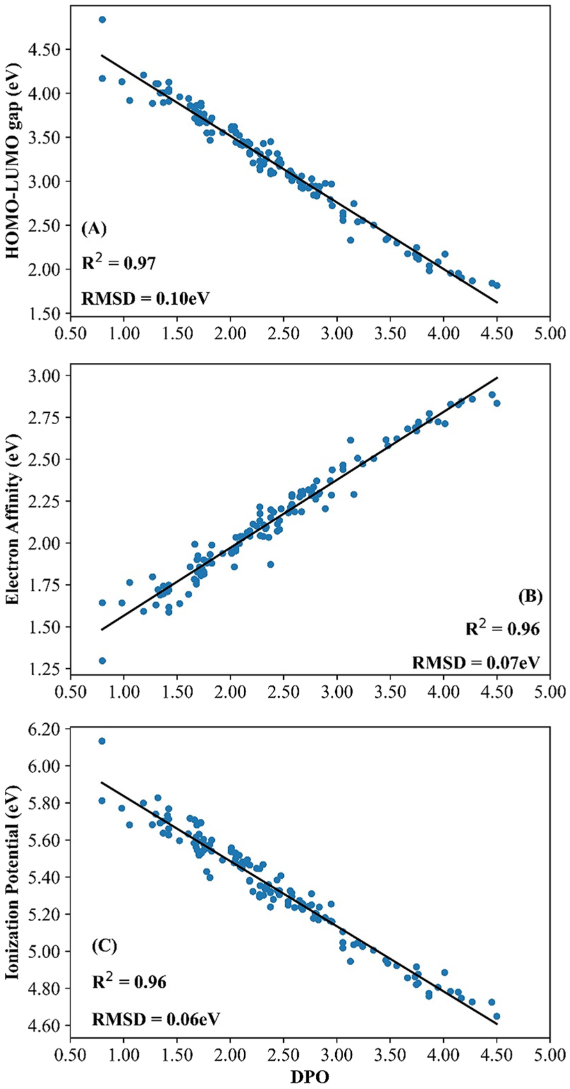


Figure 7. Plots of linear correlations between the optimized full DPO values and the DFT-calculated properties of (A) HOMO−LUMO gap, (B) electron affinity, and (C) ionization potential of molecules from the training set.

mechanical model that connects structural variables to their physical properties.[33,34] The latter problem is entirely overcome thanks to the use of the stratified data splitting approach.

The above results justify the present truncation to simplify the calculation of the total DPO value for a given molecule. Furthermore, the ML-based approach provides equally accurate results to the previous methodology in deriving QSPR relationships but is more robust and can be automated.

Next, we investigated the stability of the ML-based models, particularly the robustness of the ML-based methodology in optimizing the DPO model, namely, the effect of the training set size on the accuracy of the models is examined. To this end, using the stratified data splitting approach, different training sets with sizes of 13, 26, 40, 53, 66, 79, 106, and 132 data points were constructed. The performances of those models for the task of predicting values of the HOMO−LUMO gap, EA, and IP were then assessed using the full-size test set of 116 data points. Similar to the above, for each training data size, 20 experiments were done with different random data splits (data ensembles from 132 data points), and the results were then averaged. The plots of average RMSDs for both the truncated and full DPO models as a function of training set sizes are presented in Figure 10. Figure 11 presents the plots of the variation of DPO parameters versus the sizes of the training set.

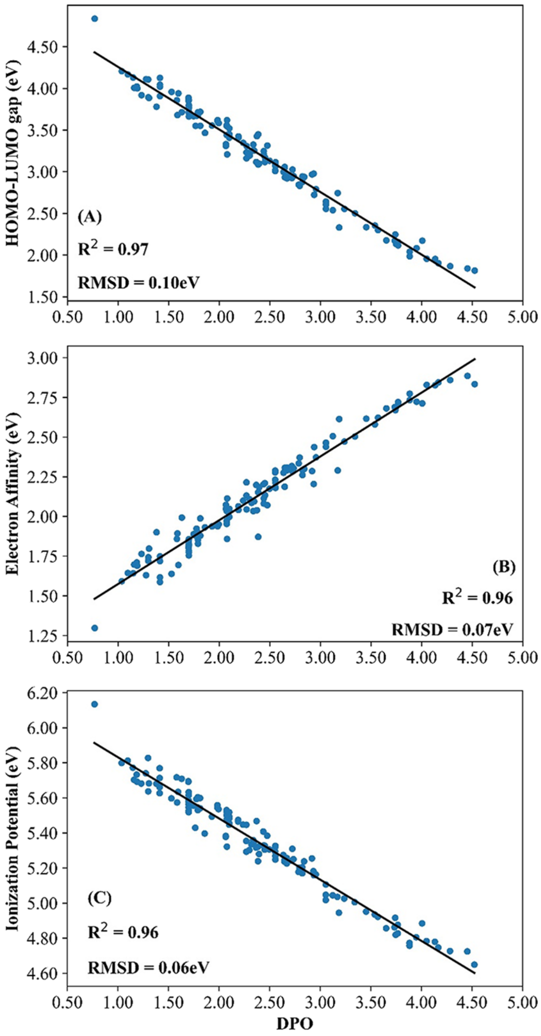


Figure 8. Plots of linear correlations between the optimized truncated DPO values and the DFT-calculated properties of (A) HOMO−LUMO gap, (B) electron affinity, and (C) ionization potential of molecules from the training set.

Figure 10 shows that the RMSDs for both the full and the truncated DPO models decrease rapidly as the training set size increases. In particular, both ML-based DPO models converge to an RMSD value between 0.10 and 0.11 eV for the band gap and 0.06 and 0.07 eV for both the EA and IP by about 53 training data points, which is less than half of the full-size training set. Even with a smaller data set of 26 training samples, both models already achieve a satisfactory RMSD of around 0.12−0.13 eV for the band gap and 0.07−0.08 eV for the EA and IP. Similarly, as shown in Figure 10, both models' parameters nearly converge to their respective optimal values with only 26 training points.

The result indicates that the machine learning methodology presented in this study is much more robust and efficient in optimizing the DPO parameters as compared to our previous manual approach. Furthermore, the fact that nearly identical results were achieved for both the full and truncated DPO

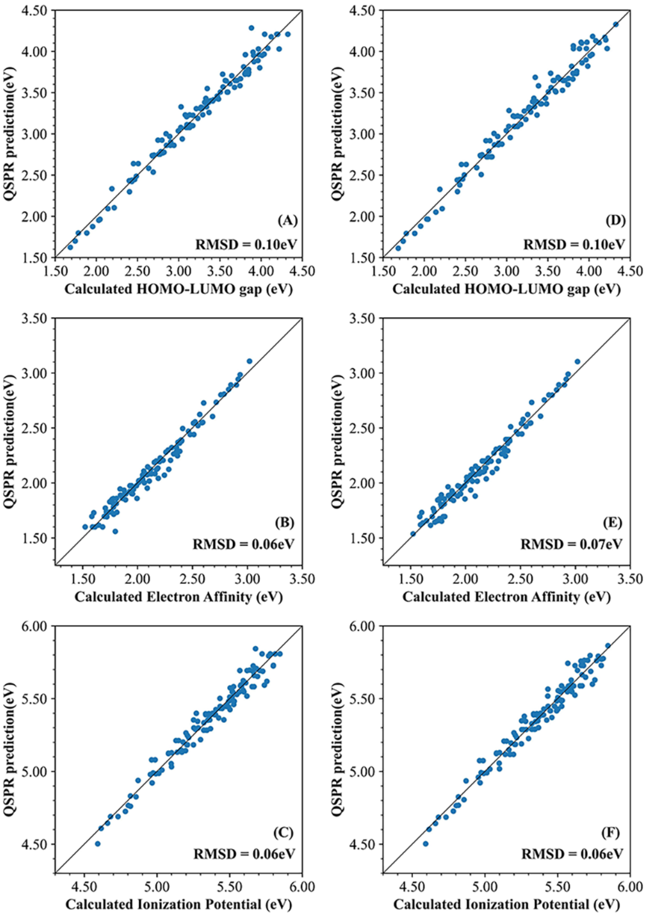


Figure 9. Plots of QSPR-predicted values versus DFT-calculated electronic properties of the HOMO−LUMO gap, electron affinity, and ionization potential, respectively from the top to bottom for compounds in the test set. Plots (A)−(C) (left side) are from the full DPO model, and plots (D)−(F) (right side) are from the truncated DPO model.

models justify the truncation approximations used and the applicability of the truncated DPO model in automated pipelines for high-throughput applications. In addition, the success of the truncated DPO model also indicates that for electronic properties of PAH or thienoacene, only the longest segment and its nearest neighbor segments are important. Contributions from other segments are negligible.

3.3. How to Use the ML-Based DPO Models. To encourage applications of the ML-based DPO model presented here, we provided a set of tools that users can easily modify to apply to other physical properties of aromatic hydrocarbons or other classes of molecules. In particular, code for the ML-based DPO model can be found in the GDDPO.py script. Data splitting, model training, and testing can be done in a couple of lines using the all-in-one Python Object Trainer from the trainer.py script. The data set consists of either manually formulated DPO polynomials or SMILES strings representing molecules and their "true" values of physical properties such as the band gap, EA, or IP. Data and codes presented in this work are available via GitHub at https://github.com/Tuan-H-Nguyen/Machine-Learning-Degree-Pi-Orbital.

In addition, it is also worth pointing out that models that have a similar structure to the DPO model but with different rules for mapping compounds to polynomial descriptors can also be used with this script. Note that users can customize for different descriptors, which requires a list of symbols used to denote parameters to be provided to either the model or the Trainer object. In this case, the object serves as an optimizer for seeking optimal values of those designated parameters using a given set of data.

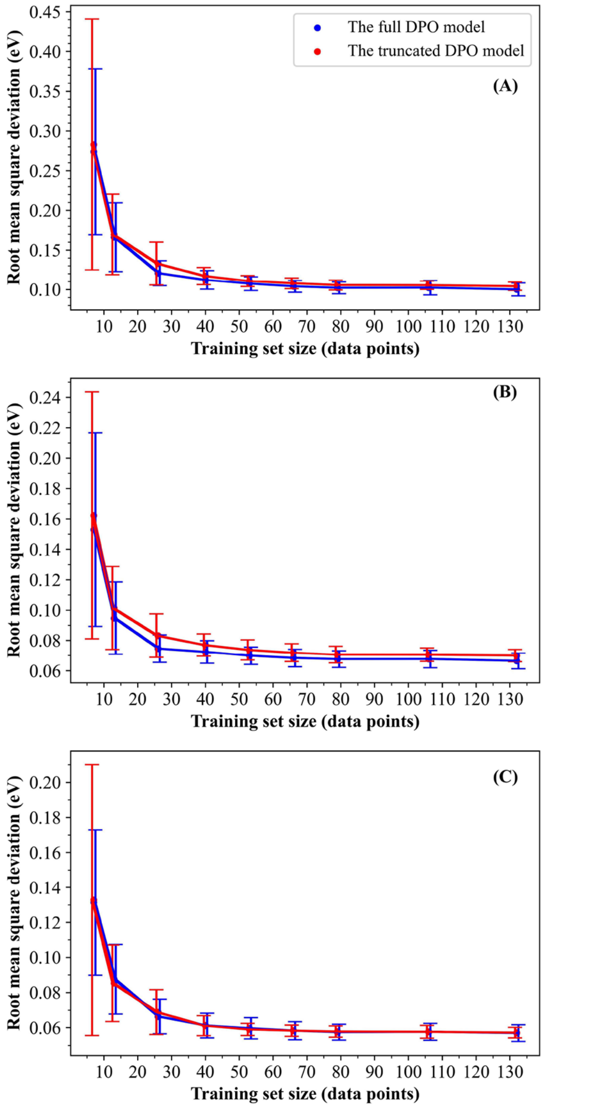


Figure 10. Plots of RMSDs and their standard deviations (as error bars) of the full and truncated DPO models of data from the test set versus the sizes of the training set. (A) HOMO−LUMO gap, (B) electron affinity, and (C) ionization potential.

4. CONCLUSIONS

In this study, we present a machine learning approach to optimize the QSPR-DPO model for mapping structural information of PAHs and thienoacenes to their electronic properties. Furthermore, to expand the possibility of employing the DPO model to other chemical applications, a truncated DPO model that can be easily implemented in an automated pipeline is proposed. While the former provides the original model with a deep learning-inspired learning mechanism, the latter introduces some new rules that facilitate the extraction of DPO descriptors from linear molecular SMILES strings. Systematic assessments on the performance and accuracy of both the ML-based methodology and the truncated DPO

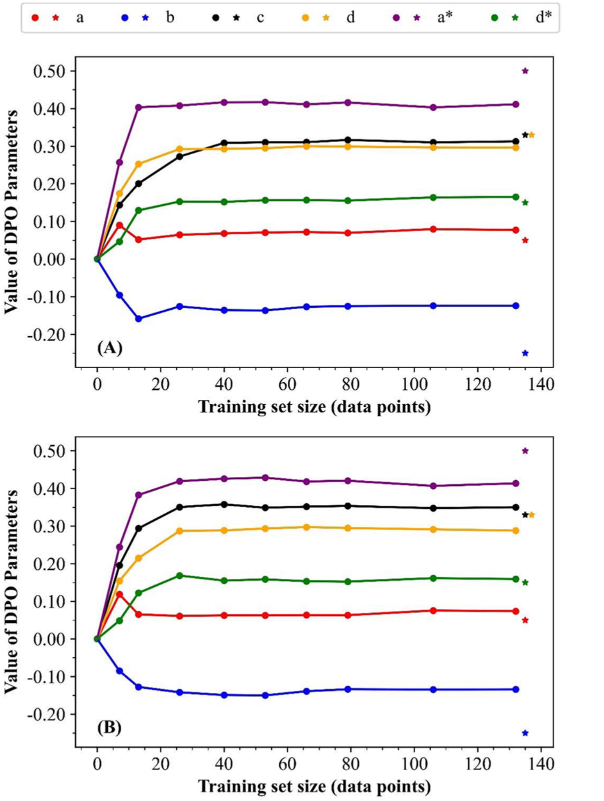


Figure 11. Plots of DPO parameter values versus the training set size for (A) full DPO model and (B) truncated DPO model. The stars are values from our previous studies.[33,34]

model in comparison with previous works were done using the same data set of 248 PAHs and thienoacenes.

The results indicate that the ML-based methodology can optimize DPO parameters to a reasonable accuracy of around 0.12 eV in the band gap with as little as 26 data points, which is much smaller than the 132 data points used in our previous works. The methodology rapidly converges as the number of training samples increases. The ML-based methodology also shows the same level of accuracy in generating QSPR relationships for predicting the band gap, ionization potential, and electron affinity electronic properties to within 0.1 eV as compared to our previous works and within the accuracy of the quantum chemistry method used to generate the data. Comparison between results from the full and truncated DPO models shows that the truncated DPO model can achieve the same level of accuracy as the full model. This confirms the validity of the truncated DPO model. Consequently, the ML-based methodology combined with the truncated DPO model enables an automated DPO model-based pipeline that takes in SMILES strings and returns predictions on electronic properties to be implemented, thus expanding the ease of use and applicability of the DPO model to different chemical classes as well as its applicability in high-throughput screening for materials design.

■ AUTHOR INFORMATION

Corresponding Author

Thanh N. Truong − *Department of Chemistry, University of Utah, Salt Lake City, Utah 84112, United States;* orcid.org/0000-0003-1832-1526;
Email: Thanh.Truong@utah.edu

Authors

Tuan H. Nguyen − *Institute for Computational Science and Technology, Ho Chi Minh City 700000, Vietnam*
Lam H. Nguyen − *Institute for Computational Science and Technology, Ho Chi Minh City 700000, Vietnam;* orcid.org/0000-0003-3347-4379

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.2c02650

Funding

This work is supported in part by funding from The Office of Science and Technology, Ho Chi Minh City, Vietnam via the Institute for Computational Science and Technology at Ho Chi Minh City with contract number 01/2021/HĐ-KHCNTT.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

QSAR, quantitative structure−activity relationships
QSPR, quantitative structure−property relationships
SMILES, simplified molecular-input line-entry system
ECFP, extended connectivity fingerprints
DPO, degree of π orbital overlap
HOMO, highest occupied molecular orbital
LUMO, lowest unoccupied molecular orbital
IP, ionization potential
EA, electron affinity
RMSD, root-mean-square deviation

■ REFERENCES

(1) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(2) Alonso, M.; Herradón, B. Neural Networks as a Tool To Classify Compounds According to Aromaticity Criteria. *Chem. − Eur. J.* **2007**, *13*, 3913−3923.

(3) Nandi, A.; Qu, C.; Houston, P. L.; Conte, R.; Bowman, J. M. Δ-machine learning for potential energy surfaces: A PIP approach to bring a DFT-based PES to CCSD(T) level of theory. *J. Chem. Phys.* **2021**, *154*, No. 051102.

(4) Ghosh, S. K.; Rano, M.; Ghosh, D. Configuration interaction trained by neural networks: Application to model polyaromatic hydrocarbons. *J. Chem. Phys.* **2021**, *154*, No. 094117.

(5) Balabin, R. M.; Lomakina, E. I. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.* **2009**, *131*, No. 074104.

(6) Balabin, R. M.; Lomakina, E. I. Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *PCCP* **2011**, *13*, 11710−11718.

(7) Glick, Z. L.; Koutsoukas, A.; Cheney, D. L.; Sherrill, C. D. Cartesian message passing neural networks for directional properties: Fast and transferable atomic multipoles. *J. Chem. Phys.* **2021**, *154*, 224103.

(8) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(9) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.;

Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.

(10) Zhou, P.; Liu, Q.; Wu, T.; Miao, Q.; Shang, S.; Wang, H.; Chen, Z.; Wang, S.; Wang, H. Systematic Comparison and Comprehensive Evaluation of 80 Amino Acid Descriptors in Peptide QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2021**, *61*, 1718−1731.

(11) Mamada, H.; Nomura, Y.; Uesawa, Y. Prediction Model of Clearance by a Novel Quantitative Structure−Activity Relationship Approach, Combination DeepSnap-Deep Learning and Conventional Machine Learning. *ACS Omega* **2021**, *6*, 23570−23577.

(12) Matsumoto, K.; Miyao, T.; Funatsu, K. Ranking-Oriented Quantitative Structure−Activity Relationship Modeling Combined with Assay-Wise Data Integration. *ACS Omega* **2021**, *6*, 11964−11973.

(13) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure−Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.

(14) Chaudhari, P.; Ade, N.; Pérez, L. M.; Kolis, S.; Mashuga, C. V. Quantitative Structure-Property Relationship (QSPR) models for Minimum Ignition Energy (MIE) prediction of combustible dusts using machine learning. *Powder Technol.* **2020**, *372*, 227−234.

(15) Atahan-Evrenk, S.; Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *J. Phys. Chem. A* **2019**, *123*, 7855−7863.

(16) Sato, T.; Honma, T.; Yokoyama, S. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening. *J. Chem. Inf. Comput. Sci.* **2010**, *50*, 170−185.

(17) Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sánchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **2011**, *4*, 4849−4861.

(18) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.

(19) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.

(20) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **2019**, *6*, 1801367.

(21) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Comput. Sci.* **2010**, *50*, 742−754.

(22) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(23) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(24) Fernandez, M.; Trefiak, N. R.; Woo, T. K. Atomic Property Weighted Radial Distribution Functions Descriptors of Metal−Organic Frameworks for the Prediction of Gas Uptake Capacity. *J. Phys. Chem. C* **2013**, *117*, 14095−14105.

(25) Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO2 Capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056−3060.

(26) Liu, A. L.; Venkatesh, R.; McBride, M.; Reichmanis, E.; Meredith, J. C.; Grover, M. A. Small Data Machine Learning: Classification and Prediction of Poly(ethylene terephthalate) Stabilizers Using Molecular Descriptors. *ACS Appl. Polym. Mater.* **2020**, *2*, 5592−5601.

(27) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *J. Med. Chem.* **2020**, *63*, 8705−8722.

(28) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292* **2015**.

(29) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212* **2017**.

(30) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(31) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

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

(33) Nguyen, L. H.; Nguyen, T. H.; Truong, T. N. Quantum Mechanical-Based Quantitative Structure−Property Relationships for Electronic Properties of Two Large Classes of Organic Semiconductor Materials: Polycyclic Aromatic Hydrocarbons and Thienoacenes. *ACS Omega* **2019**, *4*, 7516−7523.

(34) Nguyen, L. H.; Truong, T. N. Quantitative Structure−Property Relationships for the Electronic Properties of Polycyclic Aromatic Hydrocarbons. *ACS Omega* **2018**, *3*, 8913−8922.

(35) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(36) Shalev-Shwartz, S.; Ben-David, S., *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press: 2014, DOI: 10.1017/CBO9781107298019.

(37) Goodfellow, I.; Bengio, Y.; Courville, A., *Deep Learning*. The MIT Press: 2016.

(38) Walt, S. V. D.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22−30.

(39) McKinney, W. In *Data structures for statistical computing in python*, 2010; Austin, TX: pp. 51−56.

(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. mach. Learn. Res.* **2011**, *12*, 2825−2830.

(41) Meurer, A.; Smith, C. P.; Paprocki, M.; Čertı̌k, O.; Kirpichev, S. B.; Rocklin, M.; Kumar, A.; Ivanov, S.; Moore, J. K.; Singh, S.; Rathnayake, T.; Vig, S.; Granger, B. E.; Muller, R. P.; Bonazzi, F.; Gupta, H.; Vats, S.; Johansson, F.; Pedregosa, F.; Curry, M. J.; Terrel, A. R.; Roučka, Š.; Saboo, A.; Fernando, I.; Kulal, S.; Cimrman, R.; Scopatz, A. SymPy: symbolic computing in Python. *PeerJ Computer Science* **2017**, *3*, No. e103.