ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Machine Learning-Based Quantitative Structure−Property Relationships for the Electronic Properties of Cyano Polycyclic Aromatic Hydrocarbons

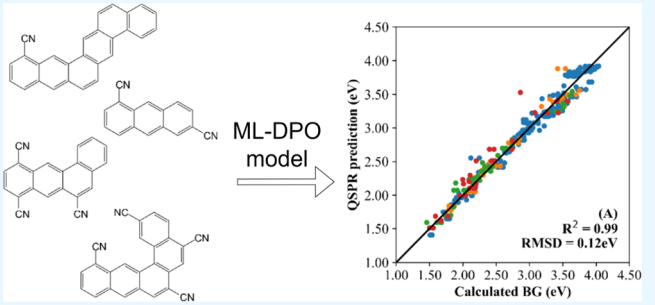
Tuan H. Nguyen, Khang M. Le, Lam H. Nguyen, and Thanh N. Truong*

Cite This: *ACS Omega* 2023, 8, 464−472

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: In this study, quantitative structure−property relationships (QSPR) based on a machine learning (ML) methodology and the truncated degree of π-orbital overlap (DPO) to predict the electronic properties, namely, the bandgaps, electron affinities, and ionization potentials of the cyano polycyclic aromatic hydrocarbon (CN-PAH) chemical class were developed. The level of theory B3LYP/6-31+G(d) of density functional theory (DFT) was used to calculate a total of 926 data points for the development of the QSPR model. To include the substituents effects, a new descriptor was added to the DPO model. Consequently, the new ML-DPO model yields excellent linear correlations to predict the desired electronic properties with high accuracy to within 0.2 eV for all multi-CN-substituted PAHs and 0.1 eV for the mono-CN-substituted PAH subclass.



I. INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) have been used as the core framework for many types of organic semiconductor materials such as transistor materials in a thin-film form with highly performant field-effect mobility.[1] However, one of the largest challenges in using PAHs is their rather selective solubility.[2] Moreover, PAHs are usually unstable under photo-oxidation when exposed to light and air.[1] To improve the usage of PAHs in such applications, functional groups such as cyano or nitrile (−CN) groups are introduced to the PAH framework.[3]

The substitution of −CN groups on the PAH framework improves not only the stability under oxidizing agents but also the increase of thermal stability and intermolecular interaction for thin-film fabrication.[3] In addition, the nitrile groups can withdraw electrons from the rings making backbone PAH more positive in the molecular surface electrostatic potential, thus improving the control of the crystallization process.[3,4] Furthermore, nitrile groups are also good electron acceptors and thus are often used for n-type semiconductors,[2,5−7] and in the production of fused-ring electron acceptors (FREAs) and other acceptor components in high-performance organic solar cells.[5] For this reason, one focus in synthetic studies is to find cost-effective methodologies for attaching the nitrile group at different sites on PAH molecules.[8] Consequently, any assistance in this endeavor such as computer-aided material designs of cyano-substituted PAH materials using quantitative structure−property relationship models (QSPR models) for specific electronic physical properties would be of great interest.

Our recent development of the QSPR model, known as the degree of π-orbital overlap (DPO) model for predicting the bandgaps, electron affinities (EA), and ionization potentials (IP) of polycyclic aromatic hydrocarbon (PAH) molecules, has shown to be accurate with errors to within 0.2 eV for two chemical classes, namely, PAH and thienoacenes.[9,10] This model is based on the quantum mechanical particle-in-the-2D-box model for describing orbital energy levels in PAH molecules. The DPO model has six nonzero parameters representing different topological features of PAH and thienoacene molecules. Optimizing these parameters with a training set of these molecules leads to QSPR for predicting the bandgaps, IPs, and EAs of all molecules in that chemical class. Traditionally, these parameters are optimized stepwise and manually, namely, determining one parameter at a time using a small subset of molecular data which has specific structural features for that parameter. This approach has proven unsuccessful for this cyano-substituted PAH chemical class. To illustrate this point, plots of the bandgaps, EAs, and

Received: August 11, 2022
Accepted: November 15, 2022
Published: December 17, 2022

IPs of cyano-substituted PAHs vs DPO using the previous methodology, which only depends on the chemical feature of the PAH core, are quite scattered, as shown in Figure 1. In particular, the larger number of CN substitutions yield values farther away from the linear trend of the data. This suggests the need for a new descriptor depending on the number of CN groups. Such scattering data also suggests that the task of optimizing a DPO model is very difficult if it is still possible for the present chemical class.

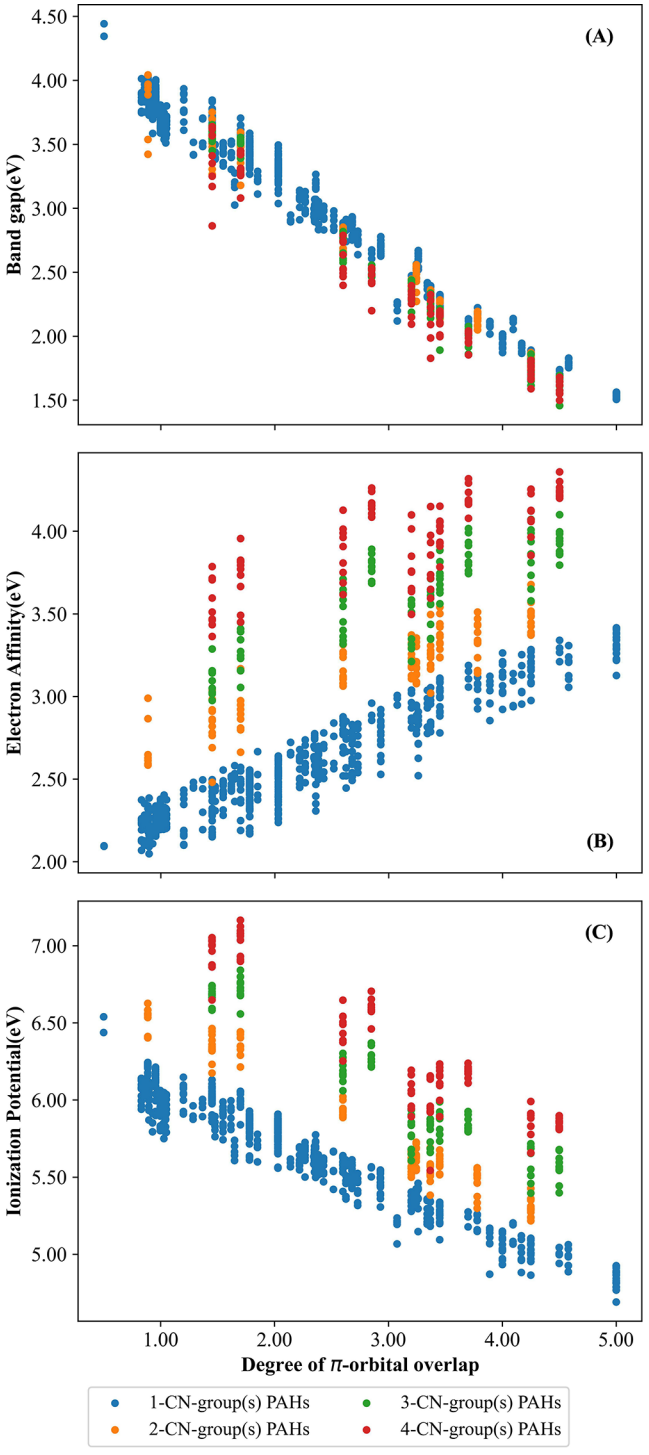


Figure 1. Plots of the electronic properties of CN-PAHs vs DPO values of the PAH framework for (A) bandgap, (B) electron affinity, and (C) ionization potential.

Recently, we introduced an application of a machine learning methodology for automatically optimizing DPO parameters and the truncated DPO model that can simplify the determination of DPO values for a given PAH molecule.[11] Both of these advances enable an automated pipeline for extracting structural features from SMILES 1D representations of molecules, assigning DPO values, and optimizing DPO parameters, all of which are needed for this study. In this study, in addition to developing a QSPR model for this cyano-substituted PAH chemical class, we also examine the applicability of the ML-based truncated DPO model to this complicated case. This study also introduced a necessary improvement to the DPO model for substituent effects for applications to more general chemical classes.

II. COMPUTATIONAL DETAILS

The data set for this study consists of a total of 926 molecules generated by attaching 1 to 4 cyano groups to different sites of 85 PAHs ranging from 3 to 10 benzene rings, as shown in Figure 2. The geometries of these molecules are fully

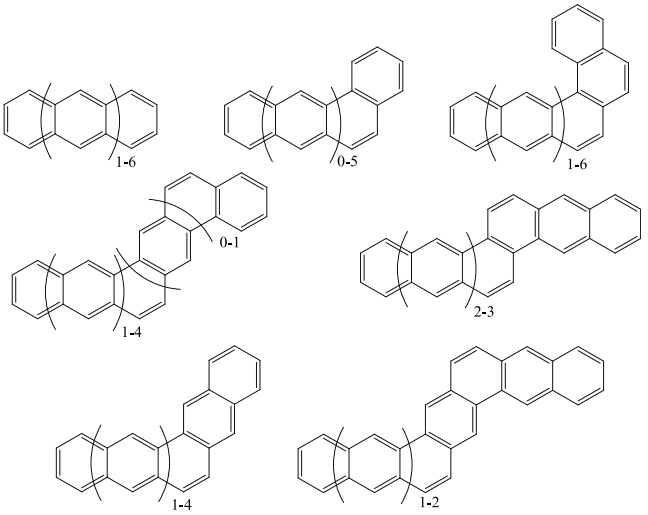


Figure 2. PAH molecules used to construct the data set.

optimized at the B3LYP/6-31+G(d) level of theory with the energy convergence criteria of 10^{-6} a.u. using the GAUSSIAN16 package.[12] For isolated molecules, according to the Koopman theorem, the difference between the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energy levels can be proven to be the difference between the ionization potential and electron affinity, that is, the experimental bandgap.[10] However, for condensed phased systems such as molecules in solution or crystal structures, the HOMO-LUMO difference can only be considered as the first-order approximation to the experimental bandgaps[1,13] and optical gap[8,14] since it does not include condensed phase effects. Such practice is common in quantum chemistry studies using isolated molecules as first-order models for condensed phase systems. We have discussed this matter in detail in our previous studies.[9,10] For consistency, the bandgap term is used in this study.

For optimization of the DPO model using the ML-based method, this data set is divided into two main subsets: a training set and a test set. In particular, the training set with

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

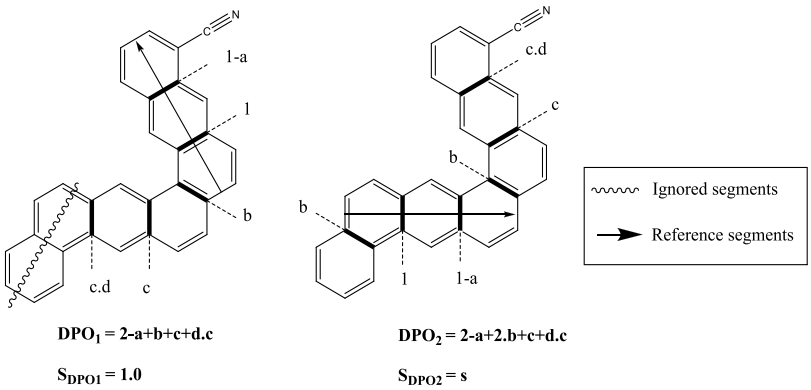


Figure 3. Illustration of how to calculate the truncated DPO descriptor value with cyano substitution when more than one PAH segment can be used as the reference segment, as indicated by the arrows. For the left figure, the CN group is on the reference segment, and thus $S_{DPO} = 1$. Assignments of the *a*, *b*, *c*, and *d* parameters are described in detail in refs 9 and 10. For the right figure, CN is not on the reference segment; thus, $S_{DPO} = s$. Finally, the DPO and S_{DPO} values are the averages from the two cases: DPO = $(DPO_1 + DPO_2)/2$ and $S_{DPO} = (S_{DPO1} + S_{DPO2})/2 = (1 + s)/2$.

556 molecules approximates 60% of the total, and the test set includes 370 molecules. A random yet stratified splitting procedure described in our previous work was used to construct the training set to preserve the distribution of bandgap values of the original data set.[11] To achieve this, the total data set is divided into bins with a bandgap width of 0.5 eV. The training set is obtained by randomly assembling from roughly 60% of every bin. The test set is simply the remaining data.

III. RESULTS AND DISCUSSION

III.I. Substituent Corrections for the DPO Model. The truncated DPO model employed in this study has been described in our previous study.[11] We only briefly mentioned it here to provide a base for introducing corrections for the substituent effects. The DPO descriptor value for a PAH molecule is a polynomial that has four nonzero *a*−*d*. The DPO value results from summing all of the parametrized values assigned to all CC fused bonds (bonds between two benzene rings) in a PAH molecule according to a set of rules, as illustrated in Figure 3. This rule starts by identifying the unique reference segment, which is the longest segment. Fused bonds on a segment are assigned parameters according to this segment's location and the orientation relative to the reference segment. In general, parameters *a*, *b*, and *c* are used to assign fused bonds in segments that form an angle of 0, 120, or 60° with the reference segment, respectively, while the *d* parameter describes the effects of the distance from a given segment to the reference one. The novelty of the truncated DPO is that only segments close to the reference segment are considered. Furthermore, in cases where there are more than one segment that can be chosen as the reference segment, rather than using an elaborate scheme for determining a unique reference segment as in the original work, the truncated DPO value is determined as the average of DPO values calculated when each of those is considered as the reference segment. An example of the calculation of truncated DPO is given in Figure 3.

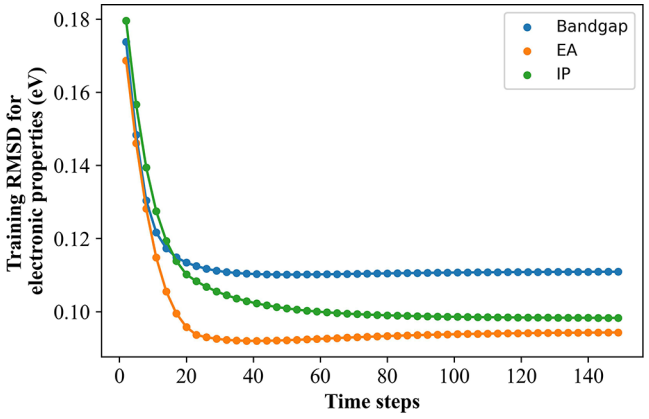


Figure 4. Plot of the training RMSDs of the ML-DPO model for three electronic properties as functions of the time steps.

As mentioned earlier, the DPO model is based on the quantum mechanical 2D particle-in-a-box system, so the above four parameters can be thought of as the effective size of the box for a given PAH. It is known that substituents placed on different edge sites of PAH have the effects of donating or withdrawing electrons from the rings and thus affecting the number of electrons placed in the rings. To account for these effects, we introduce a new substituent descriptor S_{DPO}. S_{DPO} is calculated by summing the number of cyano groups on the reference segment ($n_{CN\in ref.seg.}$) with the number of cyano groups on other segments ($n_{CN\notin ref.seg.}$) scaled by a new parameter *s* as described in eq 1 below

$$S_{DPO} = n_{CN\in ref.\ seg.} + s \times n_{CN\notin ref.\ seg.} \quad (1)$$

Table 1. Optimized Values of the DPO and S_{DPO} Descriptors for the CN-Substituted PAH Class, Along with Those from Previous Works

parameters	*a*	*b*	*c*	*d*	*s*
in this work	0.03 ± 0.00	0.19 ± 0.01	0.42 ± 0.02	0.36 ± 0.01	0.68 ± 0.01
ML-based optimization[11]	0.07	0.13	0.36	0.28	
manual optimization[9,10]	0.05	0.25	0.33	0.33	

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

Table 2. QSPR Equations for Bandgap, Electron Affinity, and Ionization Potential (all in eV)

electronic properties	QSPR equations for CN-PAH	universal QSPR equations[9,10]
bandgap	$y = 4.57-0.59 \times \mathrm{DPO}-0.09 \times S_{\mathrm{DPO}}$	$y = 4.68-0.65 \times \mathrm{DPO}$
electron affinity	$y = 1.66 + 0.23 \times \mathrm{DPO} + 0.42 \times S_{\mathrm{DPO}}$	$y = 1.36 + 0.35 \times \mathrm{DPO}$
ionization potential	$y = 6.22-0.34 \times \mathrm{DPO} + 0.33 \times S_{\mathrm{DPO}}$	$y = 6.04-0.30 \times \mathrm{DPO}$

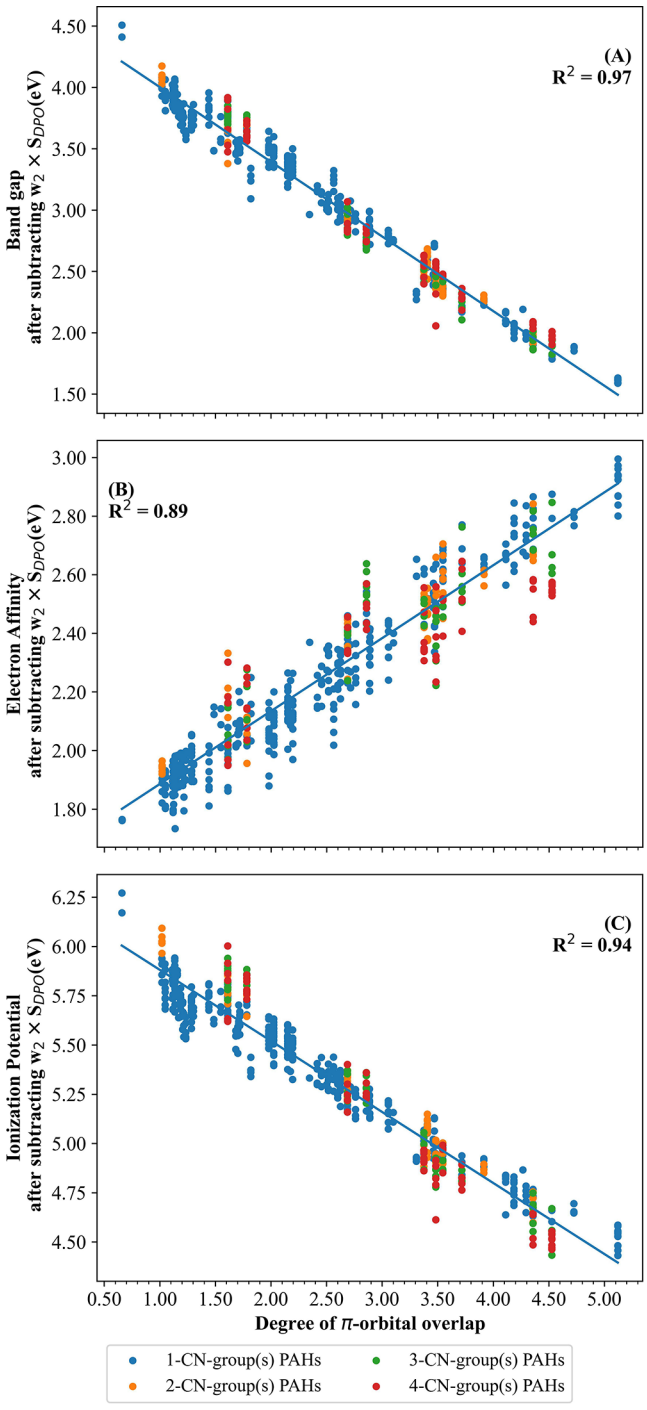


Figure 5. Linear correlations between DPO values and the (A) bandgap, (B) EA, and (C) IP without contribution from S_{DPO}.

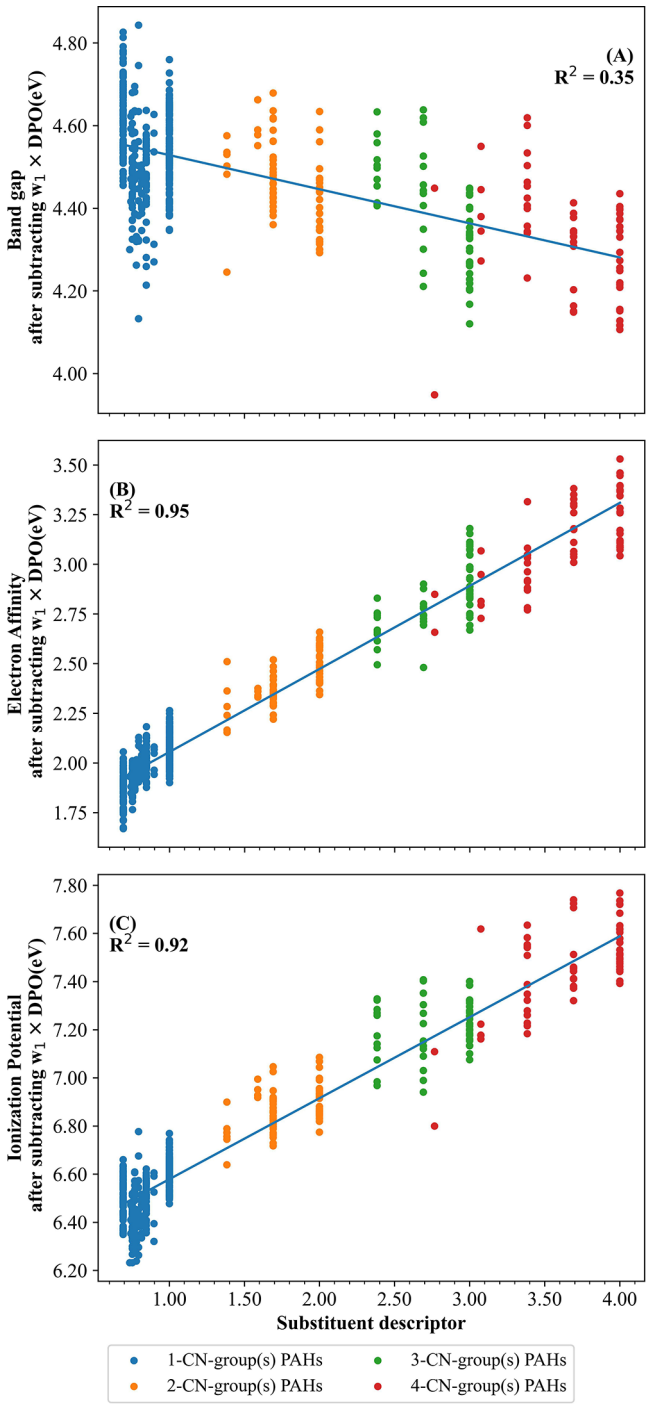


Figure 6. Linear correlations between S_{DPO} values and the (A) bandgap, (B) EA, and (C) IP without contribution from DPO.

Since the calculation of S_{DPO} depends on the reference segment, a similar averaging rule mentioned above for the truncated DPO descriptor is also used when a molecule has several segments that have the same length. Assignments and calculations of the S_{DPO} value for a cyano-substituted PAH are also illustrated in Figure 3. Finally, QSPRs for the bandgap, EA, or IP electronic properties are the bivariate linear equation as follows

$$y = w_{\mathrm{b}} + w_1 \mathrm{DPO} + w_2 S_{\mathrm{DPO}} \quad (2)$$

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

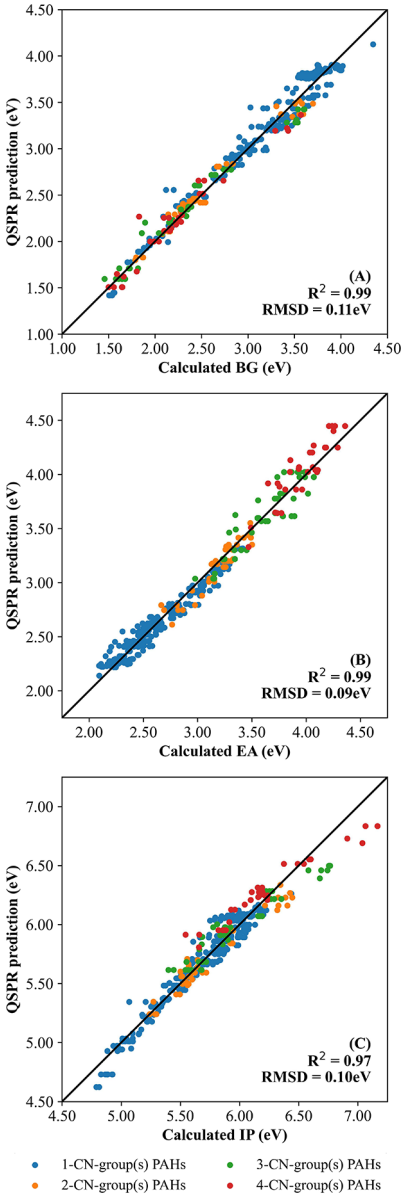


Figure 7. Plots of the QSPR predicted versus DFT explicitly calculated electronic properties for cyano-PAH molecules in the test set. (A) Bandgaps, (B) electron affinities, and (C) ionization potentials.

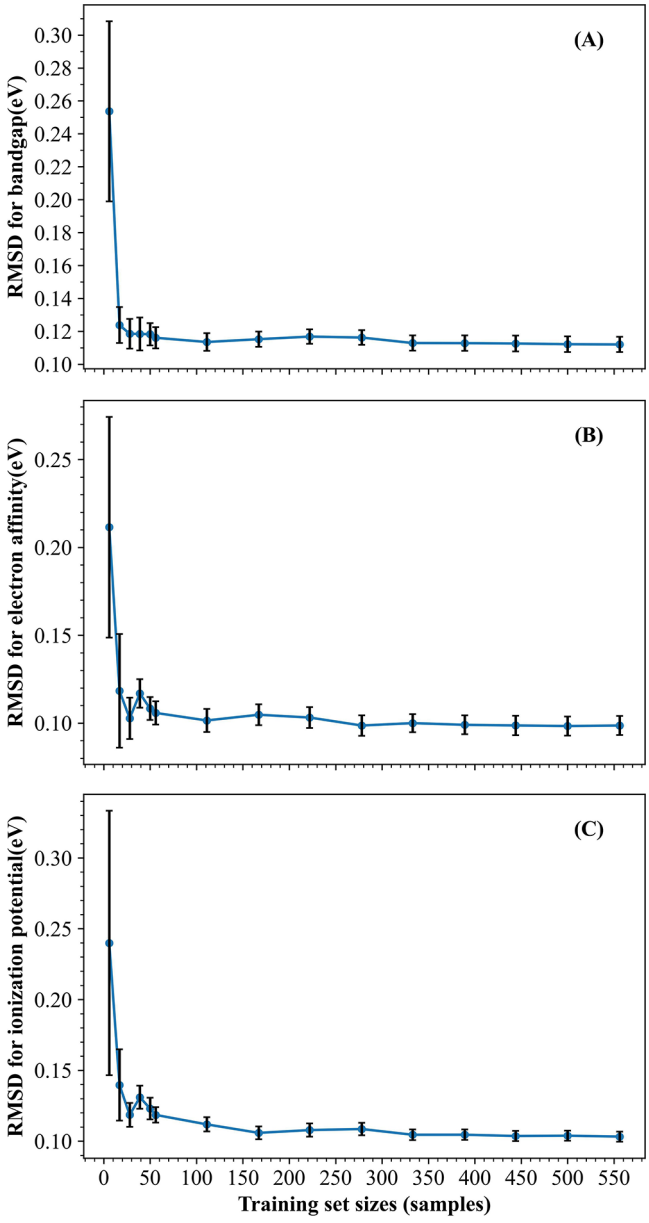


Figure 8. Plots of RMSDs and standard deviations of the truncated DPO model as functions of the training set size. (A) Bandgap, (B) electron affinity, and (C) ionization potential.

where w_b, w_1, and w_2 are parameters that can also be determined by the ML optimization procedure.

III.II. Parameters Optimizing with the ML-Based DPO Model. In this study, the truncated DPO model is optimized with the ML-based method,[11] which is an iterative process that has been proposed previously for optimizing the DPO model's parameters. To concisely include the new S_{DPO} descriptor in this process, let $X_i^{[t]}$ be a vector that is composed of both the DPO value and the S_{DPO} value of the i-th compound as below:

$$X_i^{[t]} = \begin{bmatrix} \mathrm{DPO}_i(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}) \\ S_{\mathrm{DPO},i}(s^{[t]}) \end{bmatrix} \quad (3)$$

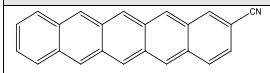
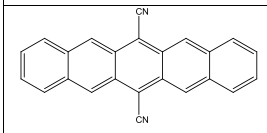
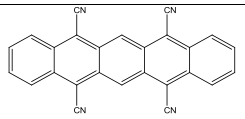
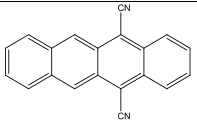
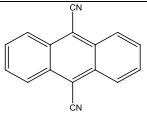
where the superscript t denotes the tth iteration. The ML-based optimization process consists of a number of steps.

Table 3. Root Mean Square Deviation (RMSD) of Each Type of CN-PAHs of the Test Set

RMSD (eV)	PAHs attached with 1 −CN group	PAHs attached with 2 −CN groups	PAHs attached with 3 −CN groups	PAHs attached with 4 −CN groups	all CN-PAHs
bandgap	0.11 ± 0.005	0.11 ± 0.01	0.12 ± 0.01	0.14 ± 0.02	0.12 ± 0.006
EA	0.08 ± 0.003	0.10 ± 0.01	0.12 ± 0.01	0.15 ± 0.02	0.10 ± 0.003
IP	0.09 ± 0.004	0.09 ± 0.01	0.13 ± 0.01	0.14 ± 0.01	0.10 ± 0.004

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

Table 4. Bandgap Comparison of Several Structures of CN-PAHs from the Experiment, the ML-QSPR Model, and the DFT Values[a,b]

structure	ref	band gap values (eV)		
		QSPR prediction	DFT calculation	experiment optical gap
	1	2.29	2.13	1.99[a]
	8	2.19	1.92	1.62[b]
	8	2.00	1.91	1.60[b]
	16	2.71	2.39	1.97[b]
	17	3.25	3.06	2.85[a]

[a]Experimental data in solution. [b]Experimental data on a thin film.

Details of these steps were described in our previous study;[11] thus, only modifications are discussed here.

Step 1: Initialize all parameters $p^{[0]} = a^{[0]}, b^{[0]}, c^{[0]}, d^{[0]}, s^{[0]}$ to zeros. As before, p collectively denotes all parameters a, b, c, d, and s.

Step 2: Calculate $X_i^{[t]}$ of PAH molecules (values for DPO and S_{DPO}) and its gradients with respect to all parameters. This gradient can be written in terms of gradients of both descriptors as in eq 4.

$$\nabla_p X_i^{[t]} = \begin{bmatrix} \frac{\partial \mathrm{DPO}_i}{\partial p}(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}) \\ \frac{\partial S_{\mathrm{DPO},i}}{\partial p}(s^{[t]}) \end{bmatrix} = \begin{bmatrix} \nabla_p \mathrm{DPO}_i(a^{[t]}, b^{[t]}, c^{[t]}, d^{[t]}) \\ \nabla_p S_{\mathrm{DPO},i}(s^{[t]}) \end{bmatrix} \quad (4)$$

Step 3: Using the least-squares method to determine the linear eq 2, which is recasted into a vector form in eq 5 below, for predictions of bandgap, EA, or IP.

$$\hat{y}_i^{[t]} = W^{[t]\dagger} X_i^{[t]} + w_b^{[t]} \quad (5)$$

where $W^{[t]^\dagger} = [w_1 w_2]$ with w_1, w_2 from eq 2. Note that in this step and subsequent steps, rather than solely using values of the bandgap property as in our previous work, all three electronic properties are used. Our preliminary works indicate that fitting the model's parameters to all three properties leads to a better model prediction overall.

Step 4: For a given physical property to be fitted, calculate the mean square error (MSE) loss function $L^{[t]}$ from the predicted values using the linear eq 5 and corresponding DFT values.

Step 5: Compute the gradient of the MSE loss with respect to the set of parameters as follows

$$\nabla_p \mathcal{L}^{[t]} = \frac{2}{N} \sum_{i=1}^{N} (\hat{y}_i^{[t]} - y_i^{[t]}) W^{[t]\dagger} \nabla_p X_i^{[t]} \quad (6)$$

Step 6: Update all parameters with the gradient descent technique; then, perform steps 2−4 to obtain a new loss value and then test for convergence. If not, carry on with steps 5 and 6.

The effect of the learning rate in training the ML-DPO model has been surveyed in our previous work.[11] It was found that the learning rate value of 1.0 is the optimal value in terms of both the rate of convergence and stability; thus, this value is also used in this study. The RMSD of the model on the training set is plotted as the function of the number of time steps (iterations) in Figure 4 to demonstrate the convergence of the training process of the modified ML-DPO model described above.

Average values and standard deviations obtained from 10 training sessions on different training sets for parameters s and a−d and QSPR linear equations are listed in Tables 1 and 2, respectively. Interestingly, the magnitudes of the DPO parameters, as well as the intercept and the coefficient of the DPO descriptor in the linear equations, resemble those of the

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

earlier models.[9,10] The small differences in parameters $a-d$ from different works using different data sets indicate that these parameters were able to account for the PAH core. This suggests that the additional S_{DPO} descriptor and its QSPR parameters are able to account for most of the cyano substitution effects.

Using the partial residue plots,[15] which are projections of the overall function on each individual independent variable, we can examine the dependence of the electronic properties on each descriptor. In particular, Figures 5 and 6 show projections of the three electronic properties on the DPO and S_{DPO} descriptors, separately and respectively. *R*-square correlation values for all three electronic properties, namely, bandgap, IP, and EA, with DPO values above 0.87, as shown in Figure 5, indicate that the main DPO descriptor is still able to represent the PAH core. Correlations for the substituent descriptor S_{DPO}, as shown in Figure 6, are good for IP and EA with *R*-square values above 0.9; however, the bandgap is not as good. Note that bandgaps of CN-PAHs linearly correlate with the DPO values even without the S_{DPO} descriptor, as suggested in Figure 1; thus, less correlation with the S_{DPO} descriptor is expected. In fact, comparing Figures 5 and 6, one can see that the new S_{DPO} descriptor was able to separate out the effect of cyano substitutions from the core PAH and linearize it with the number of substitutions encoded in the SDPO descriptor as seen by the separations in the data for the 1, 2, 3, and 4 substituted CN groups while maintaining the linear correlations.

III.III. Accuracy of the Substituent-Corrected DPO Model. The accuracy of the optimized truncated DPO model is assessed by plotting its predicted electronic properties vs the corresponding DFT values, as shown in Figure 7. The experiment is repeated 10 times with different random data splits for assembling the test set. Average calculated RMSD values from these experiments for 10 different test sets are presented in Table 3. The results suggest that the model achieves good accuracies on all electronic properties. Overall, the model achieves errors of around 0.1 eV, which is within the uncertain range of the quantum mechanical DFT methodology.[10] More specifically, the model achieves the best error for singly substituted PAH molecules. The errors are slightly higher for a larger number of CN substituents.

To assess the robustness of the model, different training set sizes are used to optimize the model parameters and then calculate the RMSD errors. Note that RMSD errors of those trained models are assessed on a fixed-size test set of 370 sample data, which is the same for all experiments here. For each training set size, the experiment is repeated 10 times, and the average result of 10 runs is reported. The plots of average values and standard deviations of RMSD for all three electronic properties as functions of the training set size are given in Figure 8. As the size of the training set increases, the model's performances improve dramatically as it converges rather quickly on all electronic properties. Consequently, errors of the bandgap and EA, IP properties converge to around 0.12 and 0.10−0.11eV, respectively, with only 50 training data points, roughly 10% of the training set. This finding is consistent with our previous study[11] for PAH and thienoacene chemical classes.

Table 4 lists bandgaps of several CN-PAHs whose experimental data are also available for comparison with those from the ML-QSPR model and DFT calculations presented here. First, notice that the predicted bandgaps from the ML-QSPR model are consistently higher than those of DFT calculated values by about 0.1−0.3 eV. Since these molecules have only one PAH segment, which is the reference segment, and thus all CN substitutions for molecules in Table 4 represent only one type of substitution, whereas the model also considers all types of CN substitutions. This result suggests that the ML-QSPR model may overestimate the effects of CN substitutions on the reference segment relative to substitutions on other segments. Also, both the DFT and calculated bandgap and ML-QSPR predictions for isolated molecules are consistently larger compared to experimental data, which are in solution or on a thin film. This suggests that the condensed phase effects would lower the bandgaps for molecules in this chemical class, though experimental values from dilute organic solution are close to those from our model compared to those measured in our thin-film form. Since solvent effects on the bandgaps for the dilute solution are expected to be smaller, this also suggests that the level of DFT theory used in this study is reasonably accurate.

From the optimal corrected DPO model, the effects of cyano substituents on the electronic properties of PAH molecules can be extracted. Such knowledge would be useful for designing organic semiconductors.

First, the model suggests that attaching cyano groups to the longer segments of the PAH molecule induces more changes in magnitude to its electronic properties than attaching these groups to shorter segments. This can be realized from the optimal value of parameter $s = 0.68$, which suggests that attaching a cyano to one of the longest segments increases S_{DPO} more than attaching to other segments, according to eq 1.

Second, the attachment of cyano groups to the PAH molecule increases both its EA and IP properties. Also, since both the HOMO and LUMO levels are shifted down by roughly equal magnitudes, the bandgap is roughly unchanged upon CN substitutions. On the contrary, adding aromatic rings to a PAH molecule shifts these frontier orbitals in opposite directions, thus changing its bandgaps significantly. These trends are visible in Figure 1 and are quantitatively confirmed by equations listed in Table 2.

III.IV. How to Use the Substituent-Corrected DPO Model. Note: the running code, along with the guide and all data, is available for public use at https://github.com/Tuan-H-Nguyen/Corrected-ML-DPO-for-CN-PAH.

Note that the current trained model is only for the cyano-substituted PAH chemical class. The algorithm reads the SMILES string of any CN-PAH molecule as input data and then calculates the truncated DPO and S_{DPO} values. These values are used with the optimized QSPR equations given in Table 2 to yield predicted values for its bandgap, electron affinity, and ionization potential. Such a SMILES-to-properties pipeline can be used in high-throughput screening for determining materials with desired properties.

The general DPO model, however, can be applied to similar chemical systems. Our first study[10] presented the general DPO framework for the basic PAH core. The application of the DPO model to the thienoacene chemical class[9] illustrated how the DPO model treats aromatic heteroatomic rings fused to PAH rings. This study, combined with the ML-DPO methodology,[11] illustrates how the model can be used for more general PAH-based chemical classes.

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

IV. CONCLUSIONS

In this work, quantitative structure−property relationships based on the machine learning-based degree of the π-orbital overlap (ML-DPO) model are determined for predicting the electronic properties of the cyano polycyclic aromatic hydrocarbon (CN-PAH) chemical class. To describe the substituent effects, a new descriptor S_{DPO} is introduced to the DPO model. This modified model is then assessed with training sets and test sets randomly formed from a data set composed of over 900 CN-PAH molecules computed with the DFT level of theory.

It is found that the errors of the models are all less than 0.20 eV with average errors of around 0.12 and 0.10 eV for bandgap and EA, IP, respectively. Furthermore, the model converges and can achieve this level of error with only 50 training data points, which is quite small in comparison to the total data set. The results suggest that the truncated ML-DPO model is robust for the broad general PAH-based chemical class beyond the PAH and thienoacene chemical classes. This study opens new potential for applications of the DPO model for QSPRs for other physical and chemical properties not limited to those considered here. Furthermore, recent developments in deep learning make it possible to predict molecular properties without employing expert-derived chemical features, and it is the subject of our forthcoming study.

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; Faculty of Chemical Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City 7000000, Vietnam*

Khang M. Le − *Faculty of Chemistry, VNUHCM-University of Science, Ho Chi Minh City 700000, Vietnam*

Lam H. Nguyen − *Institute for Computational Science and Technology, Ho Chi Minh City 700000, Vietnam; Faculty of Chemistry, VNUHCM-University of Science, 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.2c05159

Author Contributions

All calculations are conducted and written in this manuscript. All authors read and approved the final manuscript.

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: **42/2020/HĐ-QPTKHCN**.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the Institute for Computational Science and Technology and the University of Utah Center for High-Performance Computing for computing resources.

ABBREVIATIONS

HOMO, highest occupied molecular orbital
LUMO, lowest unoccupied molecular orbital
EA, electron affinity
IP, ionization potential
RMSD, root mean square deviation
ML, machine learning
DPO, degree of π-orbital overlap
PAH, polycyclic aromatic hydrocarbon
CN-PAH, cyano-substituted polycyclic aromatic hydrocarbon

REFERENCES

(1) Okamoto, T.; Senatore, M. L.; Ling, M. M.; Mallik, A. B.; Tang, M. L.; Bao, Z. Synthesis, Characterization, and Field-Effect Transistor Performance of Pentacene Derivatives. *Adv. Mater.* **2007**, *19*, 3381−3384.

(2) Xiong, Y.; Qiao, X.; Li, H. Nitrile-substituted thienyl and phenyl units as building blocks for high performance n-type polymer semiconductors. *Polym. Chem.* **2015**, *6*, 6579−6584.

(3) Wan, Y.; Zhang, Z.; Liu, H.; Zhou, J.; Liu, L.; Deng, J.; Ma, Y. Molecular design and crystallization process control for thin sheet-shaped organic semiconductor crystals with two-dimensional packing. *J. Mater. Chem. C* **2021**, *10*, 2556−2561.

(4) Cho, N. S.; Hwang, D.-H.; Lee, J.-I.; Jung, B.-J.; Shim, H.-K. Synthesis and color tuning of new fluorene-based copolymers. *Macromolecules* **2002**, *35*, 1224−1228.

(5) Lim, Y.-F.; Shu, Y.; Parkin, S. R.; Anthony, J. E.; Malliaras, G. G. Soluble n-type pentacene derivatives as novel acceptors for organic solar cells. *J. Mater. Chem.* **2009**, *19*, 3049−3056.

(6) Yao, C.; Yang, Y.; Li, L.; Bo, M.; Zhang, J.; Peng, C.; Huang, Z.; Wang, J. Elucidating the Key Role of the Cyano (− C≡N) Group to Construct Environmentally Friendly Fused-Ring Electron Acceptors. *J. Phys. Chem. C* **2020**, *124*, 23059−23068.

(7) Wagner, J.; Crocomo, P. Z.; Kochman, M. A.; Kubas, A.; Data, P.; Lindner, M. Modular, n-Doped Concave PAHs for High-Performance OLEDs with Tunable Emission Mechanisms. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202202232.

(8) Glöcklhofer, F.; Petritz, A.; Karner, E.; Bojdys, M. J.; Stadlober, B.; Fröhlich, J.; Unterlass, M. M. Dicyano-and tetracyanopentacene: foundation of an intriguing new class of easy-to-synthesize organic semiconductors. *J. Mater. Chem. C* **2017**, *5*, 2603−2610.

(9) 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.

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

(11) Nguyen, T. H.; Nguyen, L. H.; Truong, T. N. Application of Machine Learning in Developing Quantitative Structure−Property Relationship for Electronic Properties of Polyaromatic Compounds. *ACS Omega* **2022**, *7*, 22879−22888.

(12) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.;

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472

Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J.*Gaussian 16*, Rev. C.01; Wallingford, CT, 2016.

(13) Jezowski, S. R.; Baer, R.; Monaco, S.; Mora-Perez, C. A.; Schatschneider, B. Unlocking the electronic genome of halogenobenzenes. *Phys. Chem. Chem. Phys.* **2017**, *19*, 4093−4103.

(14) Baerends, E. J.; Gritsenko, O. V.; van Meer, R. The Kohn−Sham gap, the fundamental gap and the optical gap: the physical meaning of occupied and virtual Kohn−Sham orbital energies. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16408−16425.

(15) Larsen, W. A.; McCleary, S. J. The Use of Partial Residual Plots in Regression Analysis. *Technometrics* **1972**, *14*, 781−790.

https://doi.org/10.1021/acsomega.2c05159
ACS Omega 2023, 8, 464−472