ACS OMEGA

Open Access

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

Article

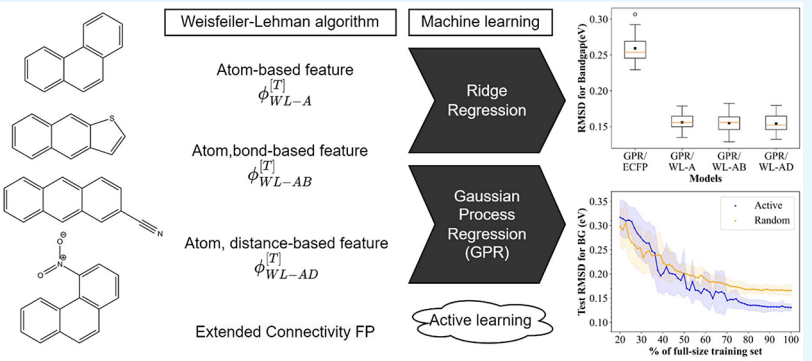
Atom-Based Machine Learning Model for Quantitative Property−Structure Relationship of Electronic Properties of Fusenes and Substituted Fusenes

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

Cite This: *ACS Omega* 2023, 8, 38441−38451

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: This study presents the development of machine-learning-based quantitative structure−property relationship (QSPR) models for predicting electron affinity, ionization potential, and band gap of fusenes from different chemical classes. Three variants of the atom-based Weisfeiler−Lehman (WL) graph kernel method and the machine learning model Gaussian process regressor (GPR) were used. The data pool comprises polycyclic aromatic hydrocarbons (PAHs), thienoacenes, cyano-substituted PAHs, and nitro-substituted PAHs computed with density functional theory (DFT) at the B3LYP-D3/6-31+G(d) level of theory. The results demonstrate that the GPR/WL kernel methods can accurately predict the electronic properties of PAHs and their derivatives with root-mean-square deviations of 0.15 eV. Additionally, we also demonstrate the effectiveness of the active learning protocol for the GPR/WL kernel methods pipeline, particularly for data sets with greater diversity. The interpretation of the model for contributions of individual atoms to the predicted electronic properties provides reasons for the success of our previous degree of π-orbital overlap model.

1. INTRODUCTION

The vast chemical space makes it difficult to screen potential compounds using experiments or ab initio methods for specific purposes. To overcome this challenge, data-driven models are being used as low-cost tools to narrow down the search space.[1] One such practice is the quantitative structure−properties relationship (QSPR),[2] where practitioners use easier-to-obtain properties, such as molecular structures, to infer materials' properties quantitatively and statistically.[3] This approach is particularly useful for modeling pertinent properties of organic semiconductor materials[1,4−6] that are difficult to experimentally measure or require extensive quantum chemistry calculations for predictions. These materials have applications in various technologies, such as photovoltaics,[7−9] light-emitting diodes,[10,11] and transistors.[12,13]

Electronic properties, such as band gap, frontier orbitals' energies, electron affinity (EA), and ionization potential (IP), of organic semiconductors correlate with the materials' stability, charge transport properties, and other devices' qualities[12,14,15] and, thus, make pragmatic and vital objective functions for data-driven optimization of materials. Fortunately, these properties correlate well with the molecule's 2-dimensional (2D) structure, especially for polycyclic aromatic hydrocarbons (PAHs) by the use of the degree of π-orbital overlap (DPO) descriptor, which is based on the quantum mechanic 2D particle-in-the-box model.[16−19] Nevertheless, the attachment of electron acceptors/donors to PAH molecules complicates such a particle-in-the-box physical model and, hence, undermines the simple "box" model. This calls for a better approach to this problem.

Modern cheminformatics provides several widely used, general-purpose 2D molecular fingerprints that summarize

Received: July 19, 2023
Accepted: September 15, 2023
Published: October 2, 2023

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

molecular structures as collections of substructures. These fingerprints include the Molecular ACCess System (MACCS)[20,21] and the extended connectivity fingerprint (ECFP) algorithm[22] for structural representation in various modeling disciplines.[5,6,21,23−27] However, these fingerprints are binary vectors that indicate the presence or absence of certain structural features to fully capture the structures of molecules that contain repetitive or similar subunits, such as polymers.[27] This is a particularly relevant concern for polycyclic aromatic hydrocarbons (PAHs) which are constructed by attaching a number of benzene units, and their electronic properties are strongly influenced by their size and shape.

Alternatively, models that operate on graph data structures can be used to learn chemical data. These include graph kernel methods,[28] which employ a kernel function[29] that computes the similarity between two graphs. There are a variety of such methods, including the marginalized graph kernel, which compares graphs on the basis of collections of random walks on both of them.[30] The Weisfeiler−Lehman (WL) graph kernels[31,32] introduce a kernel function that compares graphs by comparing substructures generated by the Weisfeiler−Lehman algorithm, which is similar to chemists' ECFP. Moreover, different Weisfeiler−Lehman kernels are introduced that incorporate different elements of the graphs into the comparison, such as shortest paths and edges.

This work focuses on developing a generalized atom-based QSPR model for electronic properties of PAHs, thienoacenes, and PAH derivatives that are substituted with nitrile (−CN) and nitro ($-NO_2$) groups using the cheminformatic and graph machine learning (ML) tools mentioned above. These diverse data sets are gathered from our previous works, as well as those created for the nitro-substituted PAHs in this work. All gathered data are resampled into three different balanced data sets, which are used to assess the accuracy of several WL kernel methods and the ECFP fingerprint, used in conjunction with the Gaussian process regression (GPR) model. Furthermore, active learning protocols are also implemented and assessed in this work. We also discuss the linear-model-based interpretations of the WL methods and the WL kernel methods' potential pitfalls.

2. METHODOLOGY

2.1. Weisfeiler−Lehman Algorithm and Kernel. In cheminformatics, molecules can be represented as undirected graphs composed of a set of nodes ($\mathcal{V}$) and a set of edges ($\mathcal{E}$), which represent a set of atoms ($\mathcal{A}$) and bonds ($\mathcal{B}$), respectively. A labeled graph has the addition of a labeling function (l), which maps each atom to a label, which is an integer in our work: l: $\mathcal{A} \rightarrow N$. The Weisfeiler−Lehman (WL) algorithm is a process that recursively computes the new labeling function of a graph. Initially, atoms are labeled on the basis of their intrinsic properties: $l^{[0]}(a) = \text{hash}[f(a)]$, where "hash" is a hash function and $f(a)$ is a list of atom a's properties, including but not limited to its atomic numbers, chirality, degree (number of non-hydrogen neighbor atoms), formal charge, number of attached hydrogens, hybridization, whether it is in a ring or an aromatic ring, and the number of radical electrons.[33] The recursive updating of each atom's label can be described succinctly using the message passing framework,[34] as follows:

$$m_a^{[t+1]} = \text{sort}\{[l^{[t]}(a')|a' \in \mathcal{N}(a)]\} \quad (1)$$

$$l^{[t+1]}(a) = \text{hash}\{[l^{[t]}(a)] \oplus m_a^{[t+1]}\} \quad (2)$$

where the superscript $[t]$ denotes the t-th iteration, $a \in \mathcal{A}$ is an atom, $\mathcal{N}(a)$ is a set of atoms that a bonds with, "sort" is a function for sorting lists in ascending order, and $\oplus$ denotes the concatenation of lists. In other words, for each atom, t-th labels (i.e., of t-th iteration) of its neighbor atoms are gathered into an ascendingly sorted list, which is inserted at the beginning with the t-th label of the main atom and then hashed to obtain the $(t+1)$-th label of the main atom. The number of iterations of this step is chosen by the users. For comparison, the ECFP method has a similar style of updating atoms' labels (which are called identifiers), except that bonds are also used for sorting neighbor atoms' labels.

The WL kernel methods use the result of the WL algorithm to formulate a kernel function for a pair of molecular graphs $\mathcal{M}$ and $\mathcal{M}'$ as a kernel function for a pair of vectors, as follows:

$$k^{[T]}(\mathcal{M}, \mathcal{M}') = k[\phi_{\text{WL-A}}^{[T]}(\mathcal{M}), \phi_{\text{WL-A}}^{[T]}(\mathcal{M}')] \quad (3)$$

The simplest feature vector $\phi_{\text{WL-A}}^{[T]}(\mathcal{M})$ for a molecule $\mathcal{M}$ is a vector whose entries count the number of atoms with certain labels, including all initial labels and all labels at the end of every iteration up to the T-th iteration of the Weisfeiler−Lehman (WL) algorithm. To construct this feature vector, we first determine all the unique atomic labels that appear in the data set and assign each of them an arbitrary index between zero and the total number of unique atomic labels. Then, the ith entry of $\phi_{\text{WL-A}}^{[T]}(\mathcal{M})$ is the count of the number of atoms in $\mathcal{M}$ that have the label corresponding to the index i. In mathematical terms, if we denote all the indexed labels as $\sigma_1, \sigma_2, \cdots, \sigma_N$, then the i-th entry of the feature vector $\phi_{\text{WL-A}}^{[T]}(\mathcal{M})$ can be defined as follows:

$$[\phi_{\text{WL-A}}^{[T]}(\mathcal{M})]_i = |\{a \in \mathcal{A} | l^{[t]}(a) \equiv \sigma_i \text{ and } t < T\}| \quad (4)$$

where $||$ returns the number of elements in a set and $\mathcal{A}$ is the set of all atoms in the molecule. Since this method is mainly concerned with atoms characterized by their surrounding topology, in this work, we call it atom-based WL (WL-A), instead of the original name subtree WL, kernel to facilitate understanding in chemical applications.

The vector $\phi_{\text{WL-A}}^{[T]}(\mathcal{M})$ shares many similarities with ECFP vectors, as both record atoms' labels. However, ECFP ignores atom labels that represent duplicate substructures and optionally, but frequently, ignores duplicate atom labels, which results in bit fingerprint vectors. Additionally, ECFP maps identifiers to vectors using a feature hashing process; therefore, ECFP fingerprint lengths are customizable.[35,36] We believe that this approach could be adapted by the WL kernel to improve the memory and runtime efficiency.

Alternatively, atoms' labels can be used to define other elements of the graphs, such as edges for bond types and shortest paths for interatomic distances,[37] which can be counted and vectorized in the atom, bond-based WL kernel (WL-AB) method and the atom, distance-based WL kernel (WL-AD) method, respectively. These items are characterized by labels of atoms at their two ends and their properties. We define labels for bonds and shortest paths, respectively, as $l_{\text{B}}^{[t]}(a, a') = [l^{[t]}(a), l^{[t]}(a')] \oplus f(a, a')$ and $l_{\text{D}}^{[t]}(a, a') = [l^{[t]}(a), l^{[t]}(a')] \oplus \text{len}(a, a')$. Here, $l^{[t]}(a) > l^{[t]}(a')$ to avoid double-counting, $f(a, a')$ is the list of properties of the bond between a and a', and $\text{len}(a, a')$ is the smallest number of bonds between a and a', i.e., the length of the shortest path. The properties of a bond that we consider in this work include bond type (e.g., single, double, ...), bond stereo

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

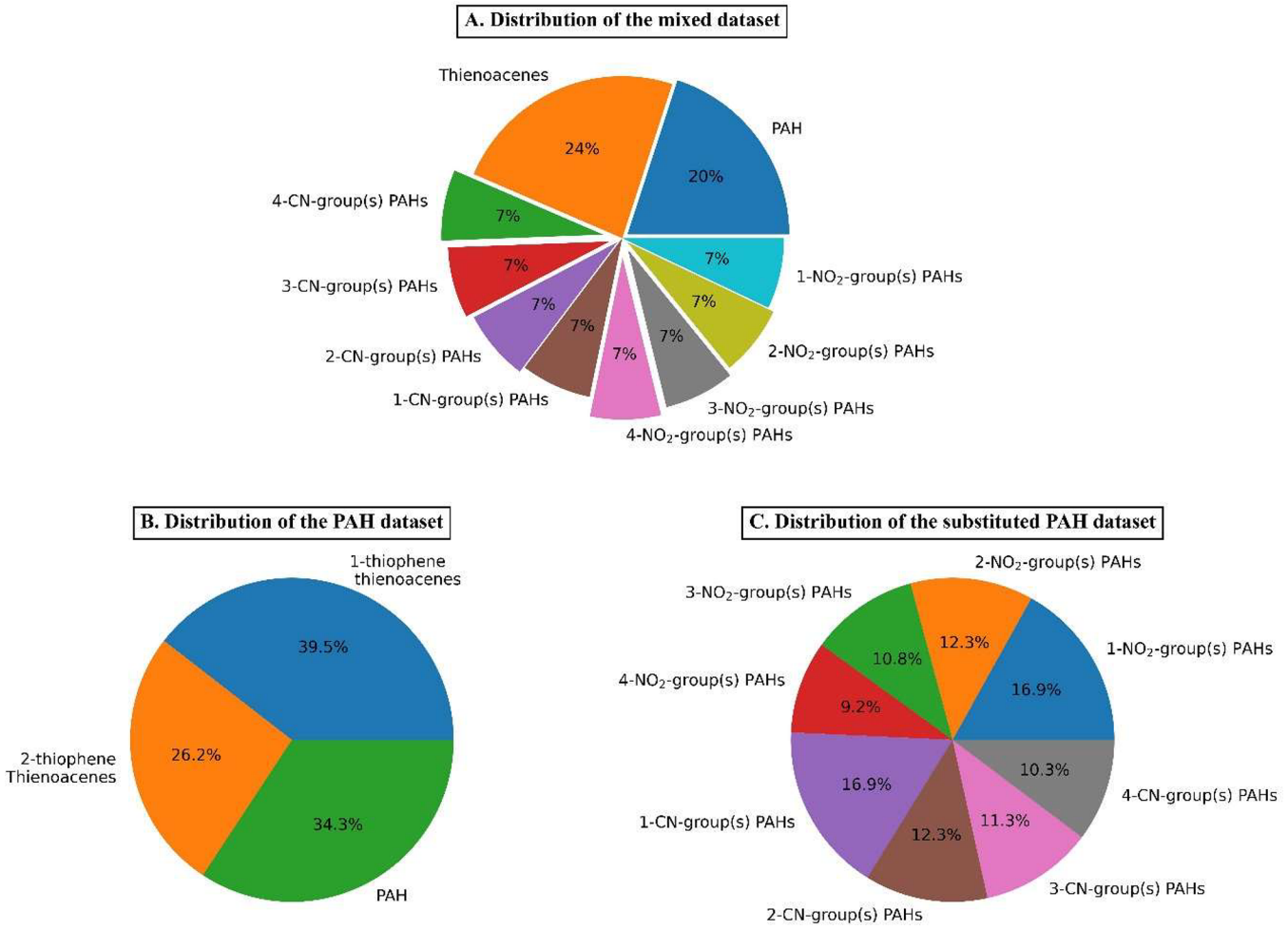


Figure 1. Distribution of molecule classes for all three data sets.

(e.g., *E*, *Z*, cis, trans, or none), and whether the bond is conjugated.[33] Similar to above, these methods can also be defined mathematically as

$$[\phi^{[T]}_{\mathrm{WL-AB}}(\mathcal{M})]_i = |\{(a, a') \in \mathcal{B} | l^{[t]}_{\mathrm{B}}(a, a') \equiv \sigma_i \text{ and } t < T\}| \quad (5)$$

$$[\phi^{[T]}_{\mathrm{WL-AD}}(\mathcal{M})]_i = |\{(a, a') \in (\mathcal{A}, \mathcal{A}) | l^{[t]}_{\mathrm{D}}(a, a') \equiv \sigma_i \text{ and } t < T\}| \quad (6)$$

As with the WL-A, all labels σ_1, σ_2,⋯ are determined beforehand and assigned indices, $\mathcal{B}$ is a set of tuples of atom pairs that bond with each other, and $(\mathcal{A}, \mathcal{A})$ is a set of tuples of all possible pairs of atoms in the molecule.

Using the Rdkit package,[38] the SMILES strings are converted to graph representations of lists of node properties and adjacency lists. From the graph representations, the described algorithms for extracting WL vectors are implemented in Python. The CRC32 hash function from the Zlib library of Python is used. The Rdkit's implementation of bit ECFP fingerprint[38] is used in this study.

2.2. Machine Learning Models and Active Learning. We used the Gaussian process regression (GPR)[39] in combination with various WL kernels or the ECFP fingerprint to predict electronic properties, which are denoted as GPR/WL or GPR/ECFP. For the kernel function of the GPR model (i.e., the function k in eq 3), we used both a linear function and a nonlinear radial basis function (RBF). One advantage of the GPR model is that it can provide predictive uncertainty,[30] which is useful for use with an active learning protocol.[30,40] This protocol deals with the problem of selecting which samples to add to the training set in order to improve the model the most when the true values of the samples are unknown. We implemented this active learning protocol[30] with our WL kernel and GPR model. Note that samples with the highest predictive standard deviations computed by the trained GPR model are assumed to have the most uncertain predictions and, thus, including these samples in the training set would improve the model.

In addition to the GPR model, we also used the ridge regression (RR) method to map ECFP fingerprints and ϕ vectors generated from WL algorithms [i.e., $\phi^{[T]}_{\mathrm{WL-A}}(\mathcal{M})$, $\phi^{[T]}_{\mathrm{WL-AB}}(\mathcal{M})$, or $\phi^{[T]}_{\mathrm{WL-AD}}(\mathcal{M})$ vectors] to electronic properties. These models are denoted as RR/ECFP and RR/WL, respectively. RR is the same as linear regression except for the use of regularization techniques to prevent overfitting.[29] In this study, implementations of these models from the scikit-learn package[41] were used.

2.3. Physical Interpretation of the Model. An advantage of using the ridge regression model is its ability to provide some physical insight into the model. In particular, the contributions of individual atoms to the property predictions can be extracted. Note that ridge regression is simply a linear regression whose loss function includes a regularization term.[29,42] Thus, the

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

prediction from the T-iteration WL-A vector of the ith compounds has the form:

$$\hat{y}_i = W\phi^{[T]}_{\mathrm{WL-A}}(\mathcal{M}) + b \tag{7}$$

Without losing the generality, suppose that an atom v has $T+1$ labels $l^{[0]}(v) \equiv \sigma_j, l^{[1]}(v) \equiv \sigma_k, \cdots, l^{[T]}(v) \equiv \sigma_p$ where $\sigma_j, \sigma_k, \cdots, \sigma_p$ are labels whose counts are kept track of at $j, k, \cdots, p$th entries of the $\phi^{[T]}_{\mathrm{WL-A}}(\mathcal{M})$ vector. Then, the contribution of atom v to the prediction according to eq 7 is

$$\hat{y}(v) = W_j + W_k + \ldots + W_p \tag{8}$$

where $W_j, W_k, \cdots, W_p$ are the $j, k, \cdots, p$th components of the parameters vector W of the RR model. Henceforth, the sum of contributions of all atoms in the molecule plus a bias constant b is the prediction of the model:

$$\hat{y}_i = \sum_{v\in\mathcal{A}} \hat{y}(v) + b \tag{9}$$

2.4. Data. In this study, we use a data set of a total of 2131 molecules from four classes of polycyclic aromatic compounds: PAHs, thienoacenes, singly to quadruply cyano-substituted PAHs, and singly to quadruply nitro-substituted PAHs. Electronic properties data on PAHs, thienoacenes, and cyano-substituted PAH molecules are from our previous studies.[16−18] Electronic properties of the nitro-substituted PAH chemical class were calculated in this study at the same level of theory as other chemical classes, namely the B3LYP-D3/6-31+G(d) level of theory, by using the GAUSSIAN package.[43] Electron affinity (EA) and ionization potential (IP) can be respectively approximated to be the highest occupied molecular orbital (HOMO) and the lowest occupied molecular orbital (LUMO) energy levels according to Koopman's theorem. The band gap values are approximated to be the HOMO−LUMO gaps.[16−18]

Because of the fact that some substituted PAH chemical classes have more structural variations than others, such as the base PAH class, the four original data sets are severely imbalanced, as illustrated in Figure S1 and Table S1, if fully used to assess the performance of different ML models in this study. To address this issue, we performed random and stratified sampling on each original chemical class to create three different balanced composite data sets, where representations of each chemical class in a composite one are similar in magnitude.

The first data set is a mixed data set, which comprises 425 data points from all four classes of molecules with similar proportions. The distribution of these four classes is shown in the pie chart of Figure 1A, and the number of data points per chemical class is given in Table S2. The second data set, named the PAH data set, is composed of all 246 PAH and thienoacene class molecules from the original data pool, as shown in the pie chart of Figure 1B and in Table S3. Finally, the substituted PAH data set consists of 887 cyano- and nitro-substituted PAH molecules that are resampled to balance the number of molecules with different types and numbers of substituents, as shown in the pie chart of Figure 1C and in Table S4.

For each experiment, we randomly split the data set into two separate subsets, namely, the training and test sets, with approximately 70% and 30% of the data, respectively. To ensure that the training and test sets have similar distributions of band gap values, we employed a band gap−bins-wise stratified splitting method, as previously described in our work.[18,19]

3. RESULTS AND DISCUSSION

3.1. Effect of the Number of Iterations. Figure 2 shows three WL vectors, $\phi^{[T]}_{\mathrm{WL-A}}$, $\phi^{[T]}_{\mathrm{WL-AB}}$, and $\phi^{[T]}_{\mathrm{WL-AD}}$, as functions of

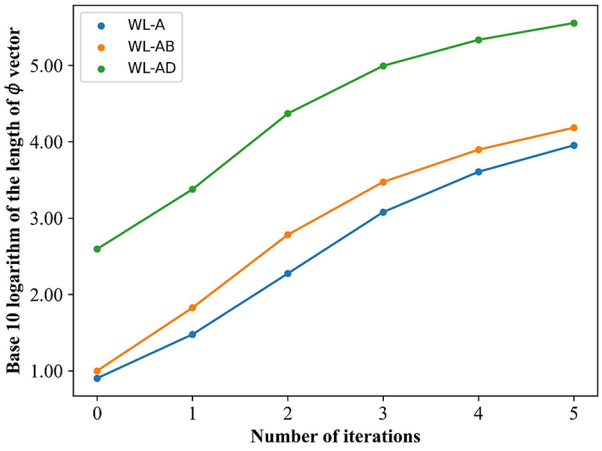


Figure 2. Plot of the base 10 logarithms of lengths of feature vectors of WL-A, WL-AB, and WL-AD as a function of the number of iterations.

the number of WL iterations. It demonstrates an exponentially increasing relationship, thereby indicating that the number of WL iterations should be carefully selected to avoid model overfitting and a long training time, especially for the GPR model.

Figure 3 shows the root-mean-square deviations (RMSD) of band gaps for the ECFP-based models (Figure 3A−C) and WL-based model (Figure 3D−I) as functions of the number of WL iterations. Because of overfitting, some GPR/WL kernel methods with high numbers of iterations produce RMSD values larger than 2.00 eV, and therefore, they are not presented in Figures 3D−F for the sake of visibility of other models' performances. Such extreme overfitting issues are also observed for the GPR/WL-AD model with radial basis function (RBF) kernel, and henceforth, only the linear kernel function is considered for these methods. However, Figure 3D−I also suggests that if the number of iterations is too low (e.g., <2 for most models), models are inaccurate because of underfitting. Note that the accuracy of the ECFP model depends on its radius of fingerprint, which is equivalent to the number of iterations of WL methods, and thus, Figure 3A−C also displays the relationship between ECFP-based model accuracy versus the number of iterations for simple comparisons. As shown in Figure 3, ECFP-based models take more iterations to converge, and the converged errors are generally higher than those of the WL kernel methods with the same regressing learning model.

From this point onward, the number of iterations of the WL model and the radius of ECFP as hyperparameters are optimized via grid search and cross-validation. Table 1 lists the search range for values of the number of WL iterations or ECFP radius and the GPR kernel function used for each method, which is based on the findings in this section.

3.2. Accuracy of GPR/WL Kernel Methods and ECFP Model. Figure 4 shows the RMSDs of several models with GPR/WL kernels or ECFP for all three data sets obtained from 20 different runs. In addition, Figure S6 similarly shows corresponding R-squared values for these models, and Figures S2−S4 plot the predictions by GPR/ECFP and GPR/WL models against the DFT-calculated values. The results indicate that three WL kernel methods demonstrate similar accuracy and

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

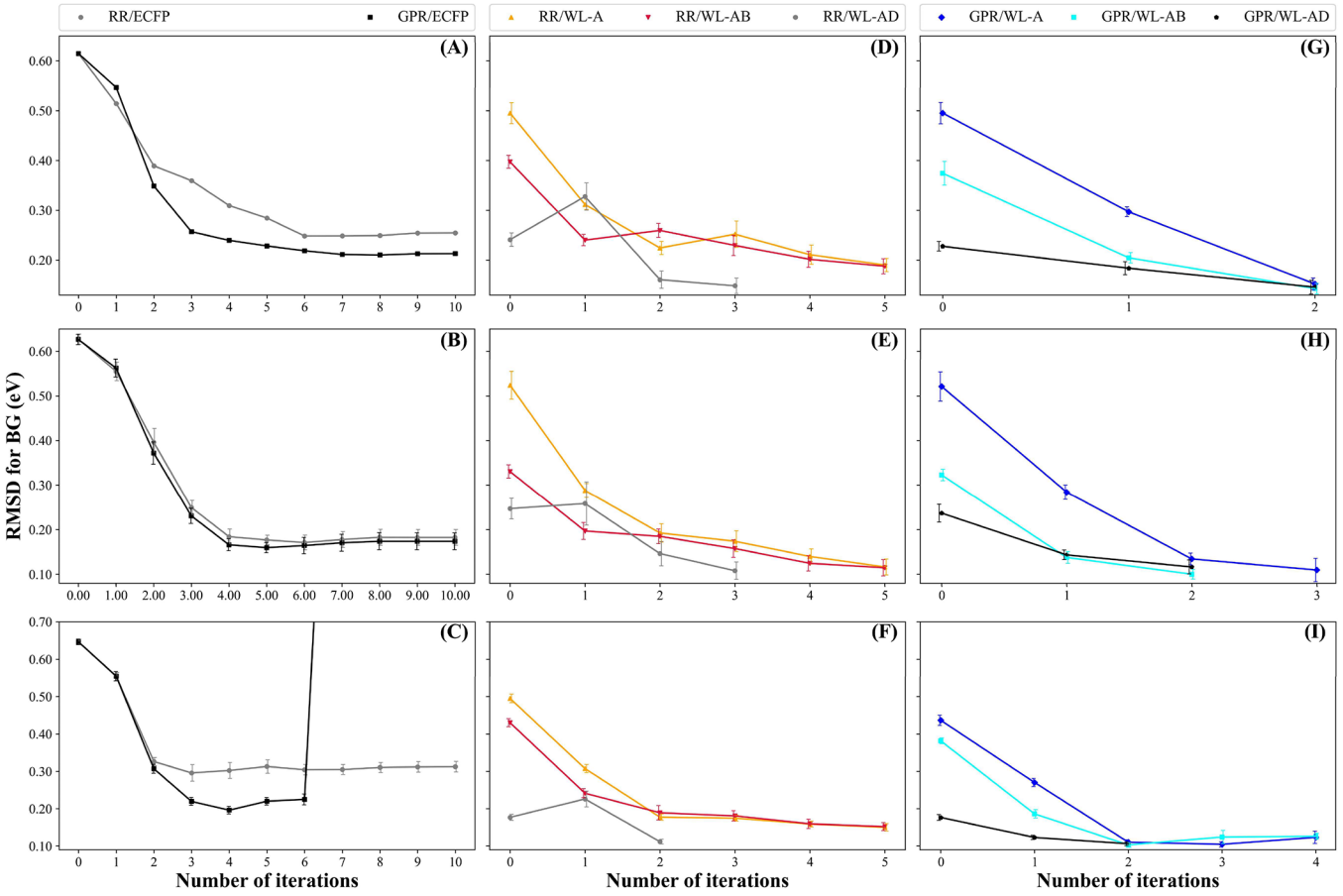


Figure 3. Plots of root-mean-square deviations for band gap as the function of the number of iterations for the mixed data set, the PAH data set, and the substituted PAH data set, respectively, from top to bottom. Plots (A−C) are for ECFP-based methods, plots (D−F) are for RR/WL, and plots (G−I) are for GPR/WL kernel methods.

Table 1. Table of Search Ranges for the Number of WL Iterations and GPR Kernel Function Used for Each Method

model	data set	number iterations	GPR kernel function
GPR/WL-A	all	[2, 3]	RBF
GPR/WL-AB	all	[1, 2]	RBF
GPR/WL-AD	all	[1, 2]	Linear
GPR/ECFP	mixed	[4−8]	RBF
	PAH	[3−6]	RBF
	substituted PAH	[2−5]	RBF
RR/WL-A	all	[3, 4, 5]	
RR/WL-AB	all	[3, 4, 5]	
RR/WL-AD	mixed and PAH	[1, 2, 3]	
	substituted PAH	[1, 2]	
RR/ECFP	mixed	[4−8]	
	PAH	[3−6]	
	substituted PAH	[2−5]	

significantly outperform the model with ECFP. For the mixed data set, the RMSD values for band gap, EA, and IP properties are around 0.15−0.16 eV, 0.12−0.13 eV, and 0.08−0.09 eV, respectively. The errors are lower for the PAH and substituted PAH data sets with 0.10−0.11 eV for band gap and less than 0.10 eV for EA and IP properties, as these data sets are more homogeneous. These results suggest that GPR/WL kernel methods can model the electronic properties of fusene compounds with reasonable accuracy given the 0.1 eV uncertainty of the DFT method.[16]

The GPR/ECFP model has an average RMSD value of ∼0.1 eV higher for the band gap property compared with the GPR/WL kernel methods. This difference can be attributed to the fact that bit ECFP vectors only register the presence of molecular fragments with 0 or 1 bit, thereby leaving out critical information regarding the size of the molecule, which is important for modeling electronic properties of PAHs. On the basis of this finding, we discourage the use of bit ECFP for modeling properties that are correlated with the size of molecules.

Since the results on the performance of different ML models are similar for the band gap, EA, and IP electronic properties, for simplicity, we use only the band gap results to present the performance of the active learning protocol and physical interpretations of the learned models below. Results for the EA and IP can be found in the Supporting Information.

3.3. Active Learning. To evaluate the active learning capability of our WL-A/GPR model, we begin with a training set that comprises 20% of the full-sized training set, while the remaining data points are placed in a test set. We then select samples from the test set, either randomly or using the active learning protocol described earlier, and then add them to the training set. We repeat this process iteratively until the training set reaches its full size. The accuracies of the model are plotted as a function of the training set size for both methods of training set augmentation, namely, randomly and actively, in Figure 5 for the

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

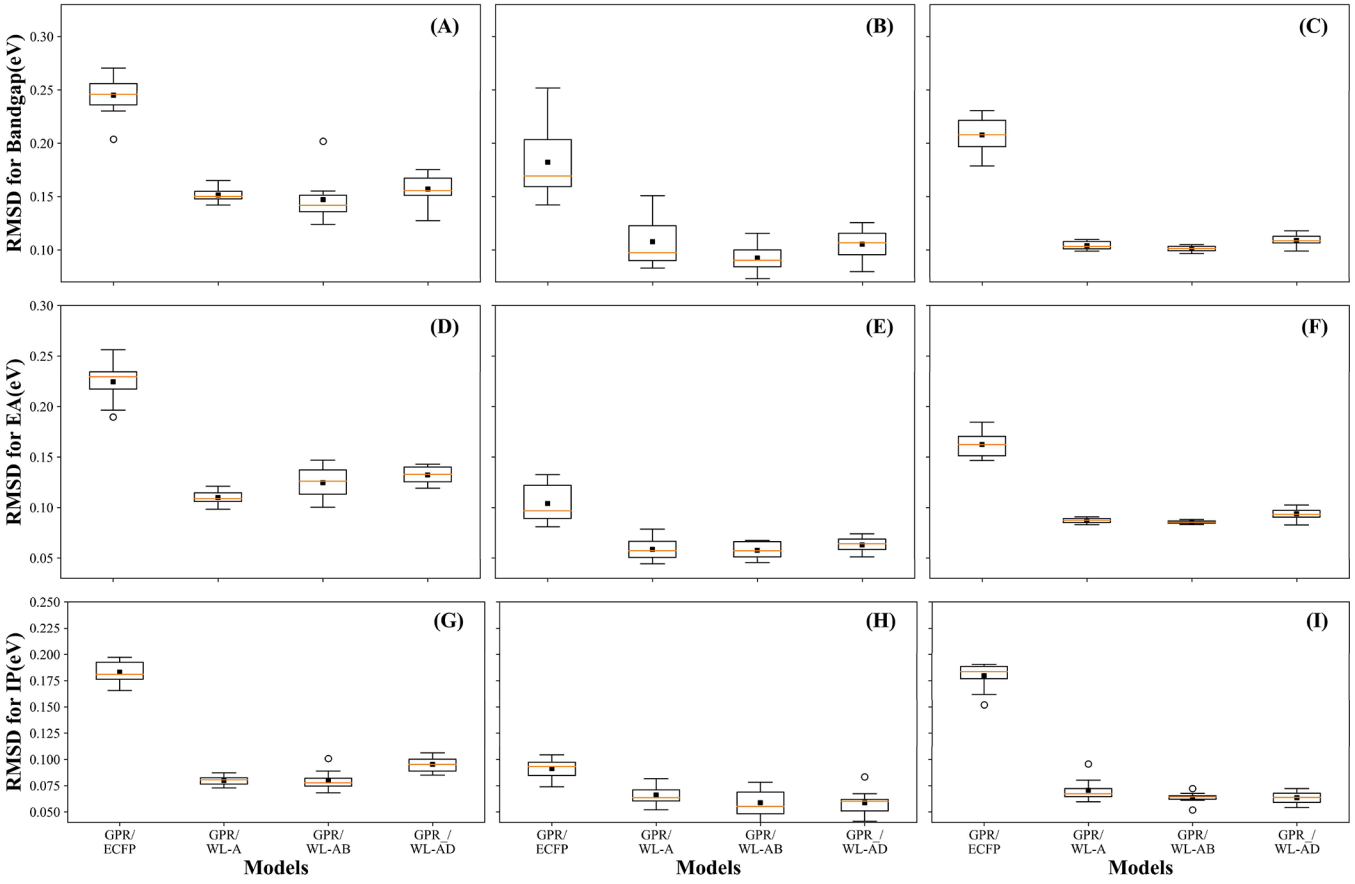


Figure 4. Boxplots for RMSDs obtained over 20 runs for GPR/WL kernel methods and GPR/ECPF model for band gap (A−C), EA (D−F), and IP (G−I) from top to bottom. Square scatters are average values of RMSDs. Figures in the leftmost (A, D, G), middle (B, E, H), and rightmost (C, F, I) columns are for the mixed data set, PAH data sets, and substituted PAH data sets, respectively.

band gap property and Figures S7 and S8 for the EA and IP properties, respectively.

In general, the models trained with actively built training sets outperform those trained on randomly built training sets in most cases, although not by a large margin, of 0.01−0.03 eV in RMSD. This is particularly true for the mixed data set, which is the most diverse data set among the three. For the substituted PAH data set, the active learning protocol only demonstrates its superiority for two kernel methods out of three (WL-A and WL-AB) and performs comparably for the other. Finally, for the PAH data set, which is the smallest of the three, the active protocol slightly underperforms for one method and performs comparably for the other two. In summary, the effectiveness of the active learning protocol seems to significantly correlate with the degree of diversity of the data sets. The result indicates that building or refining the model via training set augmentation with active learning can be achieved with slightly better results, which subsequently may reduce the expenses associated with data gathering.

3.4. Linear Models and Physical Interpretation of Learned Models. Figure 6 presents the errors determined over 20 runs of models with RR regressors and WL ϕ vectors (RR/WL) or ECFP (RR/ECFP) as feature vectors for all three data sets. Similarly, models' accuracies in terms of R-squared are presented in Figure S13, while Figures S9−S12 plot RR/ECFP or RR/WLs model predictions versus DFT calculated values for all three electronic properties. Linear models generally exhibit lower accuracy than their nonlinear GPR-based counterparts. However, the WL-based RR models demonstrate similar accuracy to GPR/WL kernel methods for the PAH data set. Additionally, RR/WL-AD exhibits better accuracy than RR/WL-A or WL-AB and roughly the same accuracy as GPR/WL models for all data sets, thereby suggesting that $\phi_{\mathrm{WL-AD}}$ demonstrates a better linear relationship with electronic properties than $\phi_{\mathrm{WL-A}}$ or $\phi_{\mathrm{WL-AB}}$. This can be attributed to the fact that the shortest paths encoded in $\phi_{\mathrm{WL-AD}}$ (or an atom pair's distance) represent the molecules' sizes, which linearly correlate with the electronic properties according to the quantum mechanical particle-in-the-box model. These results support our expectation that WL-based methods can be engineered with domain knowledge to better represent molecules for modeling certain properties.

Predictions by models that are trained in the PAH data set and the substituted PAH data set are analyzed and visualized in Figures 7 and 8. Each illustrates structures with the highest band gaps (top row), lowest band gaps (bottom row), and median band gaps (middle row). Overall, the lower the band gap, the denser and more intense the red orbs are that mark atoms that contribute negatively to the band gap properties or decrease the band gap. The appearance of the darker red orbs (atoms that contribute more negatively) in a molecule correlates with the length of its longest segment in the molecule. This result supports the concept of the reference segment of the DPO model.[16−18] Visualization for predictions of EA and IP properties are respectively given in Figures S14 and S15 for PAH data sets and Figures S16 and S17 for substituted PAH data

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

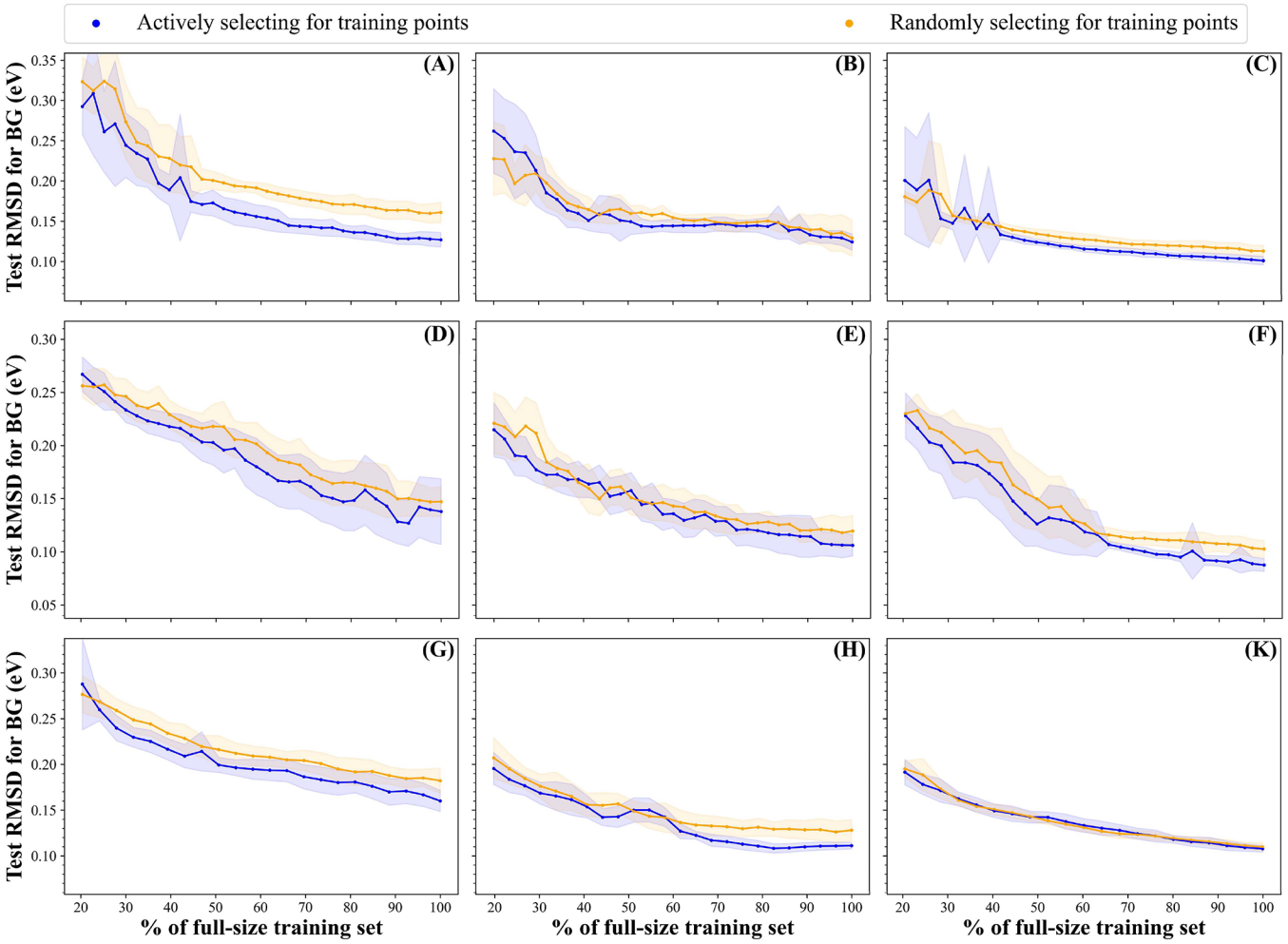


Figure 5. Plots of test RMSDs for the band gap as a function of the training set size for active learning and random selection. The leftmost, middle, and rightmost columns are for the mixed, PAH, and substituted PAH data sets, respectively, while the top row (A−C), the middle row (D−F), and the bottom row (G−K) are for the WL-A, WL-AB, and WL-AD kernel methods.

sets. In general, the visualization of the models' interpretation for EA and IP properties is mostly the same as for band gap. Note that since the EA trend is opposite to band gap or IP's, the signs of atomic contributions to IP are of opposite sign to the other two.

3.5. Erroneous Out-of-Plane Structures. The leave-one-out (LOO) method involves training a model with all samples except for one and then using that sample as a test set to assess the model's error. This procedure is repeated for each sample, thereby allowing the LOO error to be determined for all samples. Therefore, we propose using this method to identify structures that the present models predict with the largest errors. The mean of LOO RMSDs for each sample in the mixed data set of the WL-A/GPR model for band gap is 0.10 ± 0.09 eV. Note that LOO error is an overtly optimistic indicator of performance[2] and also is highly variable. Figure 9 shows nine structures with high LOO RMSD for band gap in the mixed data set. The common feature among them is that they are likely to be out-of-plane because of steric hindrance between substituents (compounds 1, 3, 6, 7, 8, 9), between substituents and fused rings (compounds 2,5), and between fused rings (compounds 4).

The planarity of PAHs is a recurring source of error for QSPR models for the electronic properties of aromatic molecules. In our previous work on the DPO-based model, we discussed this issue in detail.[16] Since the electronic properties of aromatic molecules depend strongly on their planarity, which is not explicit in 2D structural representation, the WL-A kernel method, as well as other WL-AB and WL-AD methods, also reflect this issue. The correction for this problem would be developing additional descriptors that describe the structural nonplanarity, which is a convoluted task given that the input is 2-dimensional, and thus, will be added in future study.

4. CONCLUSION

This paper presents a study of various WL kernel models and ECFP-based learning models for predicting the electron affinity, ionization potential, and band gap of fusenes. We utilized three balanced data sets from a diverse pool of more than 2000 PAHs, thienoacenes, and −CN or $-NO_2$-substituted PAHs to evaluate the models' accuracies. Our results demonstrate that WL kernel models significantly outperform ECFP-based methods and achieve accuracies within 0.15 eV, which is reasonably accurate given the DFT uncertainty of 0.10 eV. We also implemented the active learning protocol with the Gaussian process regressor, which performs favorably, especially against diverse and large data sets. Furthermore, the linear-model-based interpretation of WL methods provides a physical picture supporting our previous quantum-mechanics-based DPO descriptor.[16,17]

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

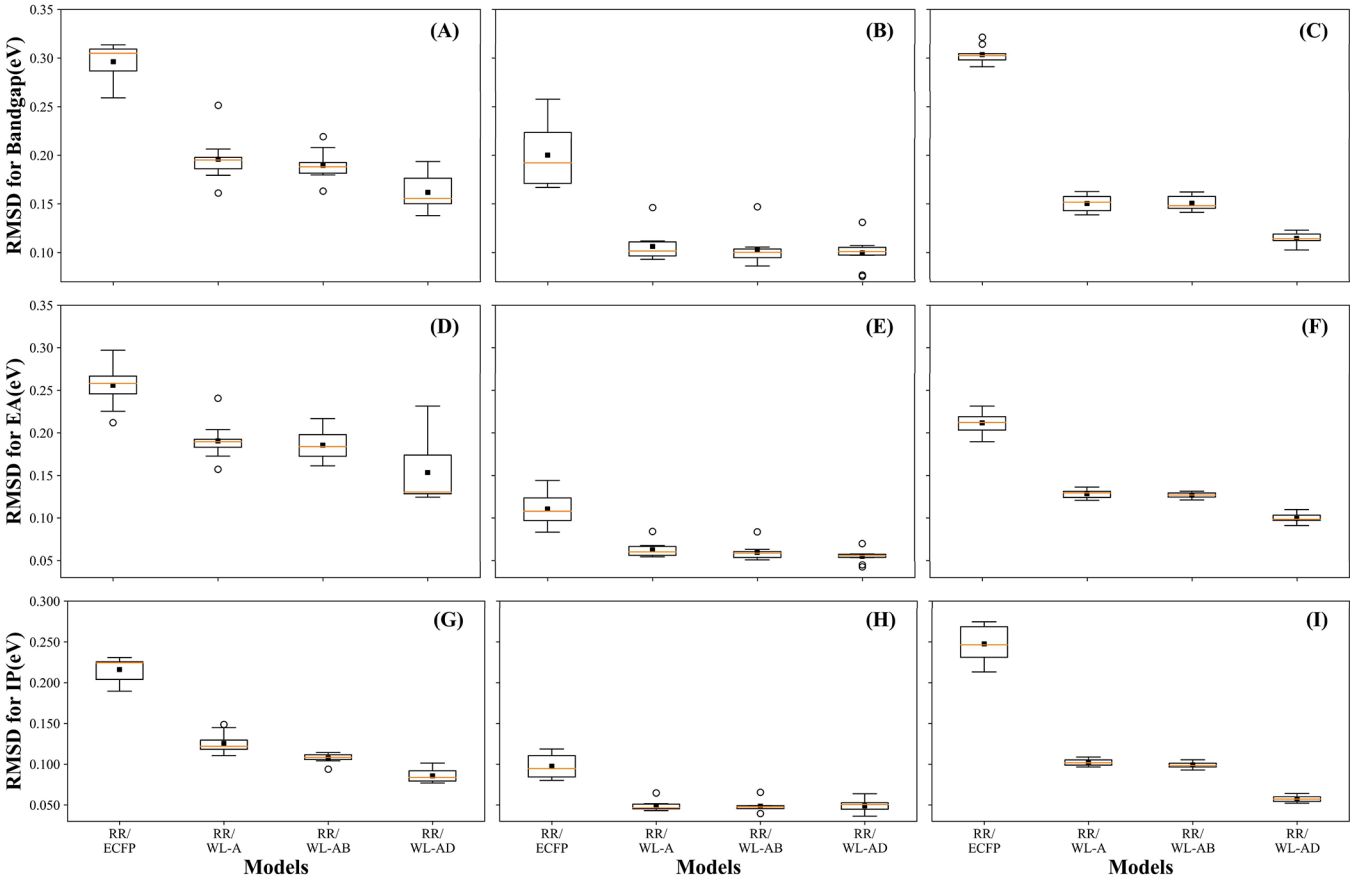


Figure 6. Box plots for RMSDs obtained over 20 runs for the RR model with either ϕ_{WL} or ECFP as the feature vector. Figures in the leftmost column, namely (A, D, G), are for the mixed data set; the middle column (B, E, H) is for the PAH data set; and the rightmost column (C, F, I) is for the substituted PAH data sets.

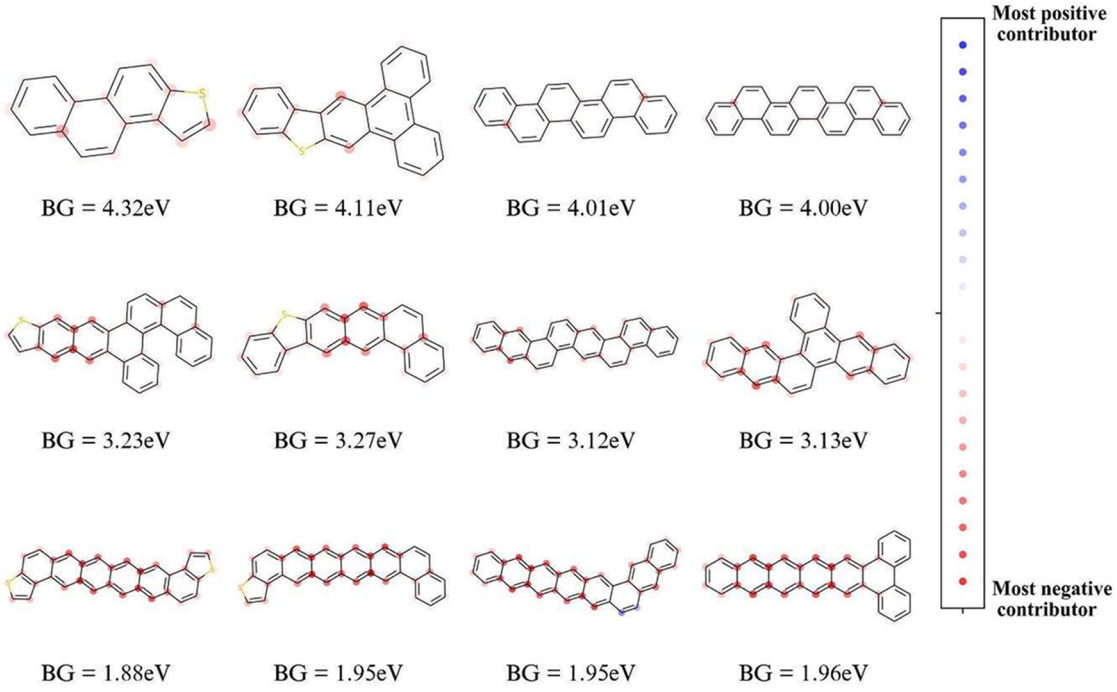


Figure 7. Visualizations of the contributions of atoms to the prediction for band gaps of several PAHs and thienoacenes with the highest (top rows), median (middle row), and lowest (bottom row) band gap values.

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

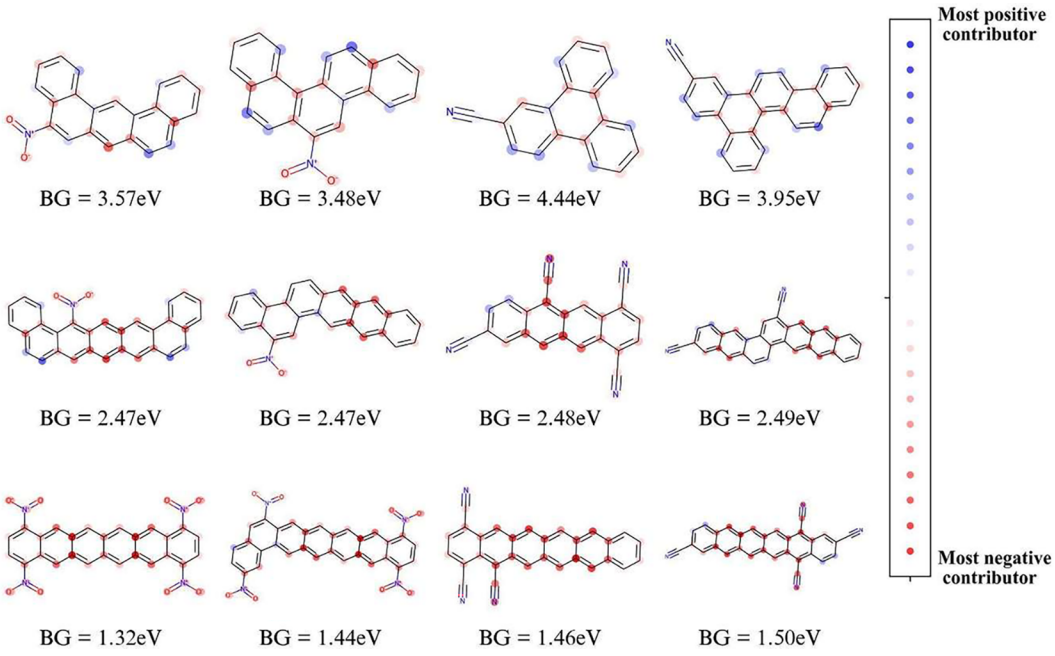


Figure 8. Visualizations of the contributions of atoms to the prediction for band gaps of several cyano- and nitro-substituted PAHs with the highest (top rows), median (middle row), and lowest (bottom row) band gap values.

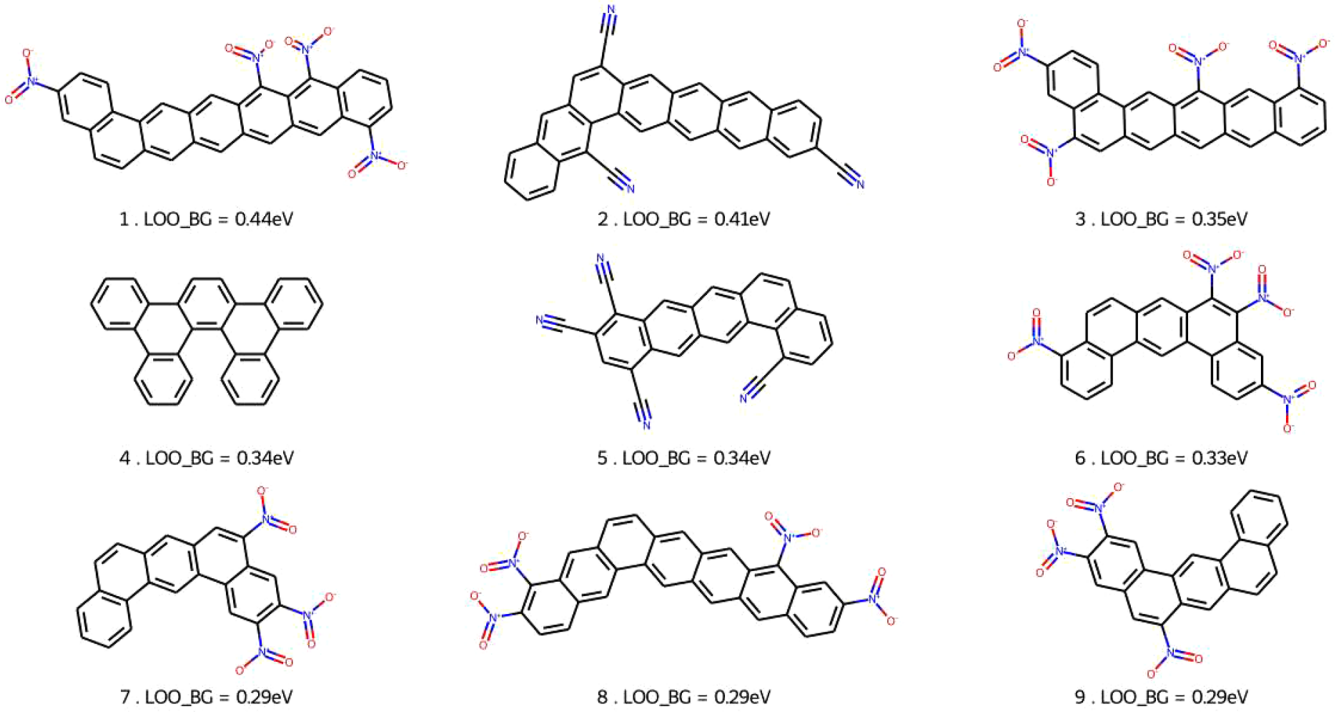


Figure 9. Compounds with the largest leave-one-out (LOO) band gap errors.

Furthermore, the WL kernels/methods come in three basic variants that have different expressive powers and complexity, as demonstrated in this work. We expect that these methods will benefit from future research that explores the customization possibilities of the presented variants, particularly the shortest-path-based WL-AD method. Moreover, the WL algorithm and kernel methods have been a source of inspiration for graph neural networks.[44,45] Therefore, we hope that our work will serve as inspiration for future studies on graph neural networks.

■ ASSOCIATED CONTENT

Data Availability Statement

Source code and data described in this paper are available at https://github.com/Tuan-H-Nguyen/Atomic-based-QSPR-for-fusenes-and-derivatives.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05212.

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451

Pie charts and tabulated data for data; parity plots for all models; box plots of R-squared values of models; plots of models' accuracy for EA and IP as a function of training set size; and visualizations of interpretations for models' predictions on EA and IP properties (PDF)

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 − *Faculty of Chemical Engineering, Ho Chi Minh City University of Technology, Ho Chi Minh City 7000000, Vietnam;* Present Address: Department of Chemistry, Emory University, Atlanta, GA 30322, USA

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

Lam H. Nguyen − *Faculty of Chemistry, VNUHCM-University of Science, Ho Chi Minh City 700000, Vietnam; 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.3c05212

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

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

EA: electron affinity
ECFP: extended connectivity fingerprint
GPR: Gaussian process regression
IP: ionization potential
LOO: leave-one-out
PAH: polycyclic aromatic hydrocarbons
RMSD: root-mean-square deviation
RR: ridge regression
WL: Weisfeiler−Lehman
WL-A: atom-based Weisfeiler−Lehman kernel method
WL-AB: atom, bond-based Weisfeiler−Lehman kernel method
WL-AD: atom, distance-based Weisfeiler−Lehman kernel method

REFERENCES

(1) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 195−216.

(2) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure−Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112* (5), 2889−2919.

(3) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121* (16), 9759−9815.

(4) 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* (12), 4849−4861.

(5) Atahan-Evrenk, S.; Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *J. Phys. Chem. A* **2019**, *123* (36), 7855−7863.

(6) Wen, Y.; Liu, Y.; Yan, B.; Gaudin, T.; Ma, J.; Ma, H. Simultaneous Optimization of Donor/Acceptor Pairs and Device Specifications for Nonfullerene Organic Solar Cells Using a QSPR Model with Morphological Descriptors. *J. Phys. Chem. Lett.* **2021**, *12* (20), 4980−4986.

(7) Xu, C.; Zhao, Z.; Yang, K.; Niu, L.; Ma, X.; Zhou, Z.; Zhang, X.; Zhang, F. Recent progress in all-small-molecule organic photovoltaics. *Journal of Materials Chemistry A* **2022**, *10* (12), 6291−6329.

(8) Zhang, T.; An, C.; Cui, Y.; Zhang, J.; Bi, P.; Yang, C.; Zhang, S.; Hou, J. A Universal Nonhalogenated Polymer Donor for High-Performance Organic Photovoltaic Cells. *Adv. Mater.* **2022**, *34* (2), No. 2105803.

(9) Kippelen, B.; Brédas, J.-L. Organic photovoltaics. *Energy Environ. Sci.* **2009**, *2* (3), 251−261.

(10) Geffroy, B.; Le Roy, P.; Prat, C. Organic light-emitting diode (OLED) technology: materials, devices and display technologies. *Polymer international* **2006**, *55* (6), 572−582.

(11) Shi, Y.-Z.; Wu, H.; Wang, K.; Yu, J.; Ou, X.; Zhang, X. Recent progress in thermally activated delayed fluorescence emitters for nondoped organic light-emitting diodes. *Chemical Science* **2022**, *13*, 3625.

(12) Schwarze, M.; Tress, W.; Beyer, B.; Gao, F.; Scholz, R.; Poelking, C.; Ortstein, K.; Günther, A. A.; Kasemann, D.; Andrienko, D.; Leo, K. Band structure engineering in organic semiconductors. *Science* **2016**, *352* (6292), 1446−1449.

(13) Tang, W.; Huang, Y.; Han, L.; Liu, R.; Su, Y.; Guo, X.; Yan, F. Recent progress in printable organic field effect transistors. *Journal of Materials Chemistry C* **2019**, *7* (4), 790−808.

(14) Griggs, S.; Marks, A.; Bristow, H.; McCulloch, I. n-Type organic semiconducting polymers: stability limitations, design considerations and applications. *Journal of Materials Chemistry C* **2021**, *9* (26), 8099−8128.

(15) Brédas, J. L.; Calbert, J. P.; da Silva Filho, D. A.; Cornil, J. Organic semiconductors: A theoretical characterization of the basic parameters governing charge transport. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99* (9), 5804−5809.

(16) Nguyen, L. H.; Truong, T. N. Quantitative Structure−Property Relationships for the Electronic Properties of Polycyclic Aromatic Hydrocarbons. *ACS Omega* **2018**, *3* (8), 8913−8922.

(17) 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* (4), 7516−7523.

(18) Nguyen, T. H.; Le, K. M.; Nguyen, L. H.; Truong, T. N. Machine Learning-Based Quantitative Structure−Property Relationships for the Electronic Properties of Cyano Polycyclic Aromatic Hydrocarbons. *ACS Omega* **2023**, *8*, 464−472.

(19) 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* (26), 22879−22888.

(20) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002**, *42* (6), 1273−1280.

(21) 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 Applied Polymer Materials* **2020**, *2* (12), 5592−5601.
(22) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(23) Yang, C.; Chen, J.; Wang, R.; Zhang, M.; Zhang, C.; Liu, J. Density Prediction Models for Energetic Compounds Merely Using Molecular Topology. *J. Chem. Inf. Model.* **2021**, *61* (6), 2582−2593.
(24) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki−Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819−4827.
(25) Gawriljuk, V. O.; Foil, D. H.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Riabova, O.; Makarov, V.; Godoy, A. S.; Oliva, G.; Ekins, S. Development of Machine Learning Models and the Discovery of a New Antiviral Compound against Yellow Fever Virus. *J. Chem. Inf. Model.* **2021**, *61* (8), 3804−3813.
(26) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22* (16), 8373−8390.
(27) Lee, F. L.; Park, J.; Goyal, S.; Qaroush, Y.; Wang, S.; Yoon, H.; Rammohan, A.; Shim, Y. Comparison of Machine Learning Methods towards Developing Interpretable Polyamide Property Prediction. *Polymers* **2021**, *13*, 3653.
(28) Kriege, N. M.; Johansson, F. D.; Morris, C. A survey on graph kernels. *Applied Network Science* **2020**, *5* (1), 6.
(29) Bishop, C. M.; Nasrabadi, N. M. *Pattern recognition and machine learning*, Vol. *4*; Springer, 2006.
(30) Tang, Y.-H.; de Jong, W. A. Prediction of atomization energy using graph kernel and active learning. *J. Chem. Phys.* **2019**, *150* (4), No. 044107.
(31) Shervashidze, N.; Schweitzer, P.; Van Leeuwen, E. J.; Mehlhorn, K.; Borgwardt, K. M. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research* **2011**, *12* (9), 2539−2561.
(32) Togninalli, M.; Ghisu, E.; Llinares-López, F.; Rieck, B.; Borgwardt, K., Wasserstein weisfeiler-lehman graph kernels. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, December 8−14, 2019.
(33) Hu, W.; Fey, M.; Ren, H.; Nakata, M.; Dong, Y.; Leskovec, J., OGB-LSC: A large-scale challenge for machine learning on graphs. *arXiv*, October 20, 2021, 2103.09430, ver. 3. DOI: 10.48550/arXiv.2103.09430.
(34) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural message passing for quantum chemistry*; PMLR, 2017; pp 1263−1272.
(35) Weinberger, K.; Dasgupta, A.; Langford, J.; Smola, A.; Attenberg, J. Feature hashing for large scale multitask learning. *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning* **2009**, 1113−1120.
(36) 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*, November 3, 2015, 1509.09292, ver. 2. DOI: 10.48550/arXiv.1509.09292
(37) Pires, D. E. V.; Blundell, T. L.; Ascher, D. B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* **2015**, *58* (9), 4066−4072.
(38) *RDKit: Open-source cheminformatics*. http://www.rdkit.org (accessed 2022-12-06).
(39) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
(40) Tynes, M.; Gao, W.; Burrill, D. J.; Batista, E. R.; Perez, D.; Yang, P.; Lubbers, N. Pairwise Difference Regression: A Machine Learning Meta-algorithm for Improved Prediction and Uncertainty Quantification in Chemical Search. *J. Chem. Inf. Model.* **2021**, *61* (8), 3846−3857.
(41) 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. *Journal of Machine Learning Research* **2011**, *12* (85), 2825−2830.
(42) Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
(43) 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.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, Jr., J. A.; 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*; Gaussian Inc.: Wallingford, CT, 2016.
(44) Lei, T.; Jin, W.; Barzilay, R.; Jaakkola, T. Deriving neural architectures from sequence and graph kernels. In *Proceedings of the 34th International Conference on Machine Learning*; Precup, D., Teh, T. W., Eds.; Proceedings of Machine Learning Research, Vol. *70*; PMLR, 2017; pp 2024−2033.
(45) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, December 4−9, 2017.

https://doi.org/10.1021/acsomega.3c05212
ACS Omega 2023, 8, 38441−38451