Environmental Science & Technology

pubs.acs.org/est

Article

Predicting Solute Descriptors for Organic Chemicals by a Deep Neural Network (DNN) Using Basic Chemical Structures and a Surrogate Metric

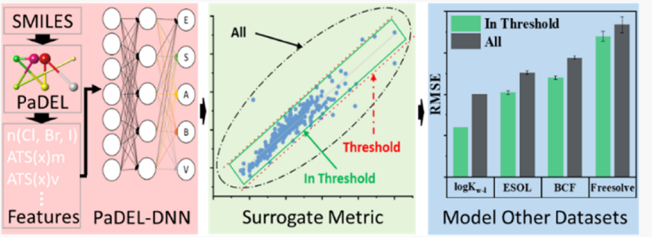
Kai Zhang and Huichun Zhang*

Cite This: *Environ. Sci. Technol.* 2022, 56, 2054−2064

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Solute descriptors have been widely used to model chemical transfer processes through poly-parameter linear free energy relationships (pp-LFERs); however, there are still substantial difficulties in obtaining these descriptors accurately and quickly for new organic chemicals. In this research, models (PaDEL-DNN) that require only SMILES of chemicals were built to satisfactorily estimate pp-LFER descriptors using deep neural networks (DNN) and the PaDEL chemical representation. The PaDEL-DNN-estimated pp-LFER descriptors demonstrated good performance in modeling storage-lipid/water partitioning coefficient (log $K_{storage\text{-}lipid/water}$), bioconcentration factor (BCF), aqueous solubility (ESOL), and hydration free energy (freesolve). Then, assuming that the accuracy in the estimated values of widely available properties, e.g., logP (octanol−water partition coefficient), can calibrate estimates for less available but related properties, we proposed logP as a surrogate metric for evaluating the overall accuracy of the estimated pp-LFER descriptors. When using the pp-LFER descriptors to model log $K_{storage\text{-}lipid/water}$, BCF, ESOL, and freesolve, we achieved around 0.1 log unit lower errors for chemicals whose estimated pp-LFER descriptors were deemed "accurate" by the surrogate metric. The interpretation of the PaDEL-DNN models revealed that, for a given test chemical, having several (around 5) "similar" chemicals in the training data set was crucial for accurate estimation while the remaining less similar training chemicals provided reasonable baseline estimates. Lastly, pp-LFER descriptors for over 2800 persistent, bioaccumulative, and toxic chemicals were reasonably estimated by combining PaDEL-DNN with the surrogate metric. Overall, the PaDEL-DNN/surrogate metric and newly estimated descriptors will greatly benefit chemical transfer modeling.



KEYWORDS: *chemical similarity, chemical transfer modeling, evaluation metric, model interpretation, octanol−water partition coefficient, PaDEL, pp-LFER descriptors, RDKit*

INTRODUCTION

Although chemical modeling using available molecular representations such as molecular fingerprints or molecular images has shown enormous potentials,[1−3] the application of these models remains limited due to either little mechanistic meanings of these molecular features or too many input features. A lack of mechanistic meanings of input features restricts the derivation of clear relationships between the features and the output predictions.[4,5] Too many features would complicate the modeling process, especially if the modeling is based on experimental data which are almost always limited. The solute descriptors used in poly-parameter linear free-energy relationships (pp-LFER, eq 1) provide a promising solution to addressing the above limitations, particularly when it comes to chemical transfer processes. There are only five pp-LFER descriptors for each chemical, each with a clear mechanistic meaning.[5] This makes it much easier to develop/interpret models and identify driving forces for a chemical transfer process. For example, the paired parameters in eq 1 can quantify the contributions in a specific chemical transfer process of polarization-induced interactions (eE), dipole−dipole/induced−dipole interactions (sS), hydrogen-bond interactions (aA and bB), and cavity formation energy (vV).[6]

$$\log \mathrm{SP} = e{\cdot}E + s{\cdot}S + a{\cdot}A + b{\cdot}B + v{\cdot}V + c \quad (1)$$

where SP represents a specific chemical transfer process; E, S, A, B, and V are the solute descriptors—or pp-LFER descriptors—that quantify the excess molar refraction, dipolarity/polarizability, hydrogen-bond donating, hydrogen-bond accepting, and excess molar volume of a solute,

Received: August 10, 2021
Revised: November 3, 2021
Accepted: December 27, 2021
Published: January 7, 2022

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

separately; e, s, a, b, and v are fitting coefficients; and c is a constant.

The contribution of these interactions can also be quantified by interpreting the obtained machine learning models, such as when predicting single or binary aqueous adsorption.[7,8] Indeed, plenty of studies have used the pp-LFER descriptors to model physicochemical processes and gained considerable insights. As chemical transfer between two phases taken as an example, the pp-LFER descriptors have been successfully used in modeling partitioning between water and various organic solvents,[9] adsorptions of organic chemicals onto different adsorbents,[10] partitioning between blood and body tissue/fluid,[11] and bioconcentration from water to aquatic organisms.[12] The pp-LFER descriptors have also shown promising applications in diverse fields ranging from chemical separation, chemical engineering, and toxicology, to pharmacology.[13]

Although the pp-LFER descriptors have found numerous applications, only around 2000 chemicals have experimental values for all of the five descriptors.[14,15] The E, S, A, and B (V can be easily obtained) are generally obtained by measuring multiple partition coefficients or solubility of chemicals in different biphase systems, based on which researchers continue to expand the database of the pp-LFER descriptors for a few chemicals per year.[16,17] However, these experimental approaches are time consuming and can hardly catch up with the rapidly increasing number of organic chemicals.[18,19] The lack of available pp-LFER descriptors has greatly limited their applications to emerging chemicals.[14]

To increase the availability of pp-LFER descriptors, studies have tried to estimate them by various models, among which group-contribution and quantum chemical calculation methods are the most widely employed. The group-contribution method builds predictive models by allocating the values of pp-LFER descriptors to certain local functional groups or the fine structure of a molecule; however, this approach mostly focuses on discrete constitutional molecular information with little attention to global molecular characteristics (e.g., topological, electrostatic, and geometric information), so estimates by this approach are not always satisfactory.[20,21] Besides, this approach generally covers limited types of functional groups as they are constrained by chemicals having known pp-LFER descriptors, so new functional groups in complex chemicals cannot be included. The second approach uses quantum chemical calculations to build predictive models and has achieved superior predictions.[6,22−24] However, direct estimation by quantum chemical calculations is mostly focused on E, A, and B, while S needs to be obtained through complex modeling or even experiments.[6,24,25] The quantum chemical approach also needs substantial computational skills which are not easily accessible for many researchers and are time consuming for high-precision calculations. Therefore, a simple predictive model that can accurately estimate the pp-LFER descriptors but requires only the most basic chemical information is highly desirable.[14]

Another problem with estimated pp-LFER descriptors is that most available models do not provide convenient ways to determine the accuracy of their estimates, while poorly estimated descriptors may not provide reliable modeling results. For example, ABSOLV, a commercial software that was built based on existing pp-LFER descriptors and only requires SMILES as the input, can predict solute descriptors; however, using these predicted descriptors without accuracy evaluation to model $\log P$ for some munition compounds yielded much higher prediction errors (RMSE: 3.56, N = 8) than using experimentally measured descriptors (RMSE: 0.37).[26] On the other hand, plenty of studies on quantitative structure−activity relationships (QSARs) have tried to address a similar problem by defining applicability domains (ADs) using methods such as convex hull, distance-based (leverage and K-nearest neighbor), probability density distribution-based, and random forest-based methods.[27] However, four ADs are needed for E, S, A, and B if following the traditional AD strategy, which would lead to an unavoidable dilemma that a chemical may be within the ADs for some of the descriptors but not for all. When this happens, it becomes challenging to evaluate the applicability of the obtained pp-LFER descriptors for new chemicals. Moreover, the underlying assumption of some ADs may not apply to the pp-LFER descriptors. For example, the commonly used leverage approach generally assumes a normal data distribution, which is apparently violated for the A and B terms because values for these two descriptors are close to zero for most chemicals (Figure S1).

In this work, we developed predictive models (PaDEL-DNN) that only require SMILES (simplified molecular-input line-entry system) of chemicals to accurately estimate pp-LFER descriptors using a deep neural network (DNN) and an open-source chemical package (PaDEL).[28] A pp-LFER data set containing all five descriptor values for 1978 chemicals was first compiled. During modeling, three commonly used chemical representations—molecular fingerprints (MFs), RDKit,[29] and PaDEL—were first compared regarding the model accuracy. Dimension reduction was then performed on the best chemical representation by three different approaches—two commonly used correlation coefficient-based methods and one LASSO method. Another four chemical transfer data sets, namely, $\log K_{\text{storage-lipid/water}}$, ESOL, freesolve, and BCF, with 327−1128 chemicals were collected for further model evaluation, where the newly estimated pp-LFER descriptors by the PaDEL-DNN were used to model the above four chemical data sets, and the comparison between the above-obtained results and reported modeling results served as an indirect metric to evaluate the PaDEL-DNN models. Moreover, instead of evaluating every individual estimated pp-LFER descriptor, we for the first time proposed to evaluate the overall accuracy of the estimated pp-LFER descriptors by one "surrogate metric," such as the octanol−water partition coefficient ($\log P$), which correlates well with many chemical transfer processes such as bioconcentration or adsorption.[12,30] The surrogate metric was validated by comparing the modeling performance on the four chemical data sets with/without applying the surrogate metric. Next, post hoc interpretation was performed on the PaDEL-DNN models to explore how training chemicals contributed to predictions and to provide insights for further improving the PaDEL-DNN models. Lastly, the PaDEL-DNN models and the surrogate metric were coupled to estimate the pp-LFER descriptors for over 4000 PBT (persistent, bioaccumulative, and toxic) chemicals. As PBT chemicals have almost no experimental properties reported but the evaluation of their risks is necessary, the estimated pp-LFER descriptors will help obtain many other properties.

MATERIALS AND METHODS

Data Collection. We compiled a data set (referred to as the LFER data set hereafter) that contains 1978 chemicals with known experimental values for all of the five descriptors.[14,15] Four additional chemical data sets were also collected,

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

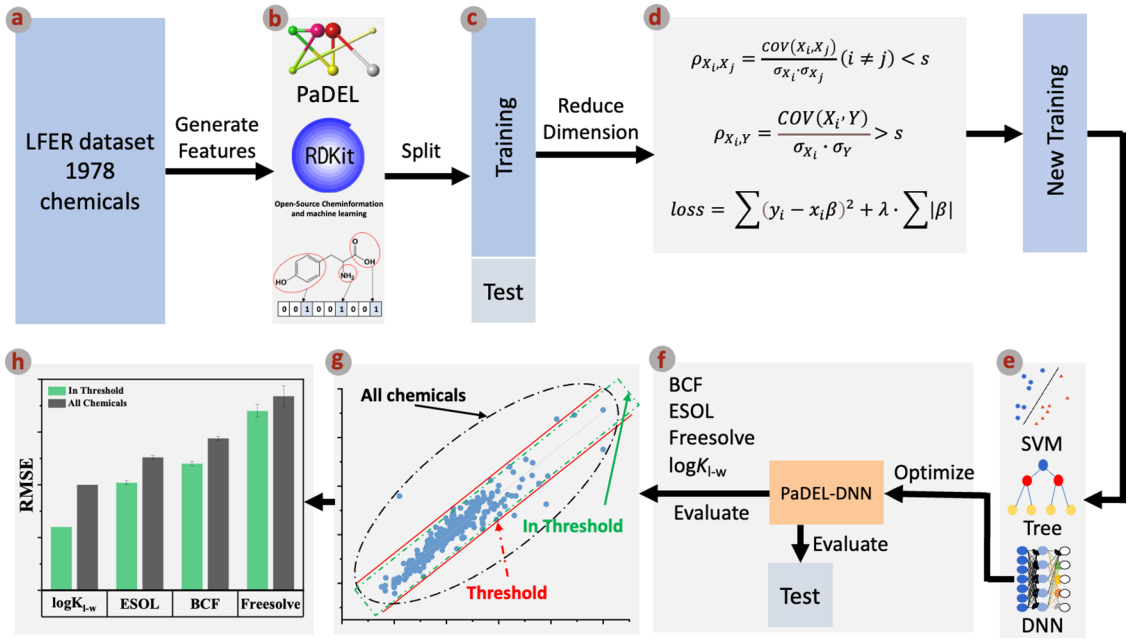


Figure 1. Workflow of this research. After collecting the LFER data set (a), three different types of chemical representations were derived for all 1978 chemicals (b). (c) These chemicals were then randomly split into training and test data sets in a ratio of 8:2. (d) For the training data set, three types of dimension reduction methods were employed to reduce the number of features in the input. (e) Dimension-reduced training data set was employed to compare the performance of three ML algorithms (support vector machine (SVM), tree, and DNN), and DNN was identified to be the best. (f) Optimized PaDEL-DNN models were first evaluated on the above test data sets and then the other four external data sets. (g) Scheme of the surrogate metric (predicted vs real values). The two parallel red lines are the 3×SD threshold; the data points in the black dashed eclipse are all of the chemicals from a chemical data set; and the points within the green dashed rectangular represent chemicals that have the pp-LFER descriptors accurately estimated. (h) Modeling performance on the four external data sets using either only the accurately estimated pp-LFER descriptors (in threshold) or all estimated pp-LFER descriptors regardless of the estimation accuracy (all chemicals).

including $\log K_{\text{storage-lipid/water}}$ (N = 327),[31] estimated SOLubility (ESOL, N = 1128),[32] free solvation (freesolve, N = 639),[33] and bioconcentration factor (BCF, N = 1034).[12,30,34,35] In addition, a list of 4020 PBT chemicals was collected from the literature.[36−38] A summary of all of the above data sets is in Table S1.

Model Development and Validation. During the model development (see the workflow in Figure 1), molecular fingerprints with lengths ranging from 512 to 2560 bits and the RDKit molecular representation (referred to as RDKit hereafter) with 1249 features were derived using the RDKit package.[29] The PaDEL molecular representation with 1444 features (referred to as PaDEL hereafter) was derived using the PaDEL package (more details in Text S1.1).[28] The chemicals in the LFER data set were then randomly split into training and test data sets in a ratio of 8:2. The best model was first selected by performing fivefold cross-validation on training data sets and then evaluated on test data sets (details in Texts S1.2 and 1.3). When estimating the pp-LFER descriptors for the chemicals in the $\log K_{\text{storage-lipid/water}}$, ESOL, freesolve, BCF, or PBT data sets, all 1978 chemicals were employed as the training data set to maximize the model performance.[39]

Although ML algorithms are good at extracting useful information from high-dimensional inputs, a simpler input is still quite attractive because it may not only increase the efficiency of model training but also simplify the interpretation and application of the model. Thus, three methods were used to reduce the input dimension (reduce the number of input features) of the deep neural network (DNN) model. The first method (referred to as the input−input coefficient) dropped highly correlated input features according to their correlation coefficients (ρ), by selecting only one feature from any pair of features whose ρ-value was greater than a certain threshold (0.6−0.8 depending on the descriptors). The second method (referred to as the input−output coefficient) selected features that correlated well—correlation coefficient > 0.6−0.9—with the pp-LFER descriptors. The third method (referred to as the LASSO coefficient) dropped features according to the coefficients of LASSO, which uses L1 regularization to ensure the sparsity of models while achieving small prediction errors (details in Text S2).[40,41] Using these three sets of reduced inputs, we followed the aforementioned model development and validation procedure to develop dimension-reduced PaDEL-DNN models, and the reduced PaDEL-DNN models were also employed to estimate the pp-LFER descriptors for the four chemical data sets.

Modeling Chemical Transfer Processes. In addition to directly evaluating the PaDEL-DNN models on the reserved test data sets, we indirectly evaluated them on the additional four chemical data sets. Briefly, we used the PaDEL-DNN-estimated solute descriptors to model the aforementioned four chemical data sets, and modeling results could indirectly indicate the prediction performance of the PaDEL-DNN models on new chemicals, as a low modeling error would suggest good estimation of the pp-LFER descriptors by the PaDEL-DNN models. Because pp-LFER models have been well developed for $\log K_{\text{storage-lipid/water}}$,[42] we inputted the estimated pp-LFER descriptors into reported pp-LFER equations to calculate $\log K_{\text{storage-lipid/water}}$ for (1) all reported 305 chemicals, (2) a subgroup of 51 chemicals, and (3) another 22 complex chemicals (more details in Text S3.1).[31] We then examined the accuracy of the obtained

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

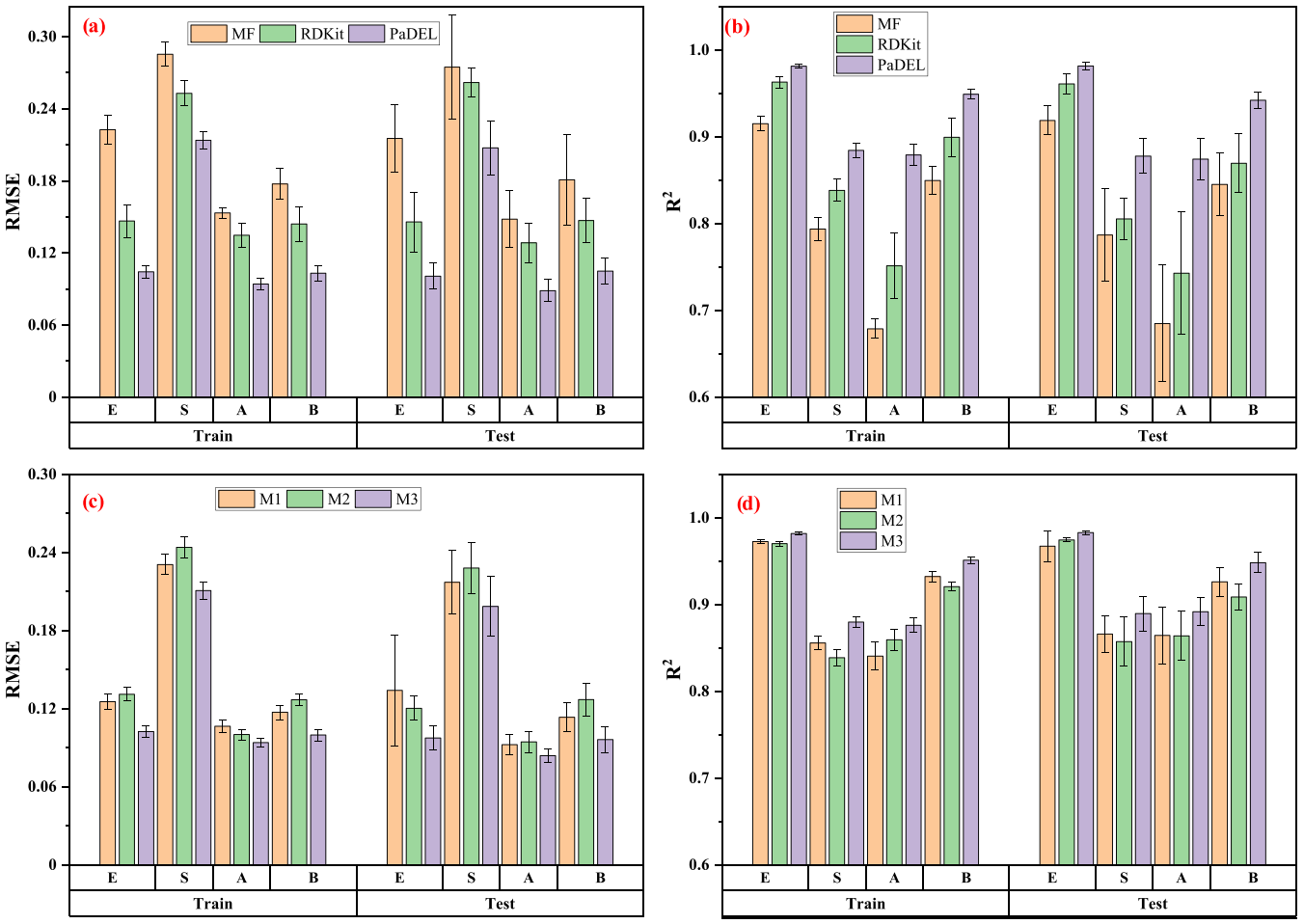


Figure 2. Performance of DNN models in predicting the pp-LFER descriptors. (a, b) Models based on three different chemical representations. (c, d) Models with different input dimension reduction methods. M1−M3 represent the PaDEL-DNN models using dimension-reduced inputs based on the input−input, input−output, and LASSO methods, respectively. MF = molecular fingerprints, Train = training data sets, and Test = test data sets.

$\log K_{\text{storage-lipid/water}}$ values. This accuracy was further compared with those by ABSOLV (also estimates the pp-LFER descriptors first and then calculates the $\log K_{\text{storage-lipid/water}}$) and by three other common methods—KOWWIN, SPARC, and COSMOtherm, which predict $\log K_{\text{storage-lipid/water}}$ directly using chemical structures like SMILES.

Unlike $\log K_{\text{storage-lipid/water}}$, there are no well-established models for ESOL, BCF, or freesolve using the pp-LFER descriptors. Therefore, new models were first built for these three processes using the small data set-friendly Gaussian process regression (GPR).[43] The inputs for these three data sets were the five PaDEL-DNN-estimated pp-LFER descriptors, and the output was the corresponding property—ESOL, BCF, or freesolve. During the modeling, the data set was first randomly split into training and test data sets (8:2). The best model was selected based on fivefold cross-validation on training data sets and was then evaluated on reserved test data sets by examining the RMSE and R^2 values (more details in Text S3.2). These models were then paired with the surrogate metric to validate the PaDEL-DNN models, as discussed below.

Surrogate Metric. For chemicals with estimated pp-LFER descriptors, a reported multilinear regression (MLR, eq 2, Texts S4.1,4.2)[17] equation was employed to estimate their $\log P$ values using the obtained pp-LFER descriptors.

$$\log P = 0.562 \cdot E - 1.054 \cdot S + 0.034 \cdot A - 3.46 \cdot B + 3.814 \cdot V + 0.088 \quad (2)$$

The differences between the estimated and reported $\log P$ values were used as a surrogate metric to evaluate the overall accuracy of the estimated pp-LFER descriptors. The reported standard deviation (SD) for $\log P$ is around 0.2 log unit based on eq 2,[9] so around (3 × SD, Text S4.3) was used here as the threshold for acceptable accuracy. If the absolute estimation errors for $\log P$ for certain chemicals are smaller than the threshold, the obtained pp-LFER descriptors are deemed "accurate", otherwise "inaccurate." The basic concept behind the surrogate metric is that accurate pp-LFER descriptors would most likely predict $\log P$ well while inaccurate descriptors would not. For test chemicals from the LFER data set, this idea can be easily verified (Texts S4.4 and 4.5). For chemicals from the other four chemical transfer data sets, the surrogate metric was validated indirectly because their pp-LFER descriptors are mostly unknown. As the BCF data set is taken as an example, PaDEL-DNN was first used to estimate the pp-LFER descriptors for all chemicals in the data set. Then, we used the pp-LFER descriptors as inputs and modeled BCF values only for the chemicals whose pp-LFER descriptors were deemed accurate (referred to as "Accurate Estimations"). Meanwhile, the same number of chemicals (referred to as

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

"Random Estimations") as in the Accurate Estimations was randomly chosen from the BCF data set to perform the same modeling. This was to exclude the influence of fewer chemicals involved in modeling. If the modeling errors for the Accurate Estimations were considerably smaller than those for the Random Estimations, it is likely that the surrogate metric has identified chemicals with accurate pp-LFER descriptors.

Post Hoc Interpretation of PaDEL-DNN Models. The chemical similarity is generally thought the key to the performance of QSARs and machine learning models;[44−46] however, traditional chemical similarity mostly focuses on the overall structural similarity, while some pp-LFER descriptors are largely determined by certain functional groups (e.g., A and B are mostly determined by O/N-containing groups). Finding a suitable similarity metric—measures to quantify similarity among chemicals—is a critical step in understanding how the PaDEL-DNN models use the chemical similarity to make estimates. To this end, the obtained models were analyzed in two steps (details in Texts S5.1 and 5.2): (1) finding the optimal similarity metric through developing *K*-nearest neighbor models (*K* ranging from 3 to 50, meaning the top 3−50 most similar chemicals to a test chemical) through a suitable similarity metric and examining their performance. The better the used similarity metric, the lower the prediction errors and (2) exploring the importance of similar (*K*-nearest neighbors) versus less similar training chemicals in the model performance using the selected similarity metric.[47]

Application of PaDEL-DNN Models to PBT Chemicals (Details in Text S6). The PaDEL package was first used to generate the PaDEL representation for over 4000 PBT chemicals. The PaDEL representation was then inputted to the PaDEL-DNN models to estimate pp-LFER descriptors for these PBT chemicals. The surrogate metric was further employed to evaluate the overall quality of these new estimates.

RESULTS AND DISCUSSION

Comparison of Three Chemical Representations. The obtained DNN models based on the PaDEL representation (PaDEL-DNN) provided the best estimations for all of the pp-LFER descriptors, followed by the models based on the RDKit representation (RDKit-DNN, 21−34% higher RMSE than PaDEL-DNN), and then the models based on molecular fingerprints (MF-DNN, 41−116% higher RMSE than PaDEL-DNN, Figure 2). The R^2 for the estimated pp-LFER descriptors decreased from PaDEL-DNN to RDKit-DNN and then MF-DNN for all of the descriptors. Meanwhile, all of these models provided low variance and consistent modeling results for both the training and test data sets, suggesting the robustness and generalization ability of these models.

For the MF-DNN models, the general trend is that the longer the fingerprints the better the estimations, but the improvement becomes negligible when the length exceeds 1536 bits (Table S2). This agrees well with our previous findings that used molecular fingerprints to model aqueous reaction rate constants for organic pollutants with OH radicals.[1] The molecular fingerprint consists of either 1s or 0s to indicate whether a structure or functional group exists. The conversion from SMILES to molecular fingerprints is not entirely reversible, and a considerable amount of chemical information may be lost during the conversion. This resembles the commonly used group-contribution method, which attributes the desired property to contributions from functional groups or substructures of a molecule.[48] In other words, molecular fingerprints can only represent local constitutional information of chemical structures but lack global parameters[49] for describing possible interactions among local chemical features. This may be the reason for the worst performance by MF-DNN.

Compared with MF-DNN, RDKit-DNN showed improved estimations for the *E* term and considerable improvements for *S*, *A*, and *B* terms (Figure 2a,b). The RDKit representation is essentially a combination of 1D functional groups/structures and common 2D/3D descriptors such as "Asphericity" and "Topological Polar Surface Area" (Table S3).[31,50,51] 1D descriptors count the number of different functional groups or structures in molecules, for example, the number of heavy atoms or hydrogen donating/accepting groups, while 2D and 3D descriptors cover some global parameters considering possible intramolecular interactions between various functional groups.

The PaDEL representation shares some features with RDKit, such as the functional group/structure counting, but includes many other unique features such as "Van der Waals volume," "Vertex adjacency information," and "Zagreb index."[28] The PaDEL-DNN achieved significant improvements over the best RDKit-DNN models with the RMSEs/R^2 of 0.1/0.98 and 0.1/0.95 for *E* and *B*, separately (Figure S3). These are considerably better than the most recent quantum chemical calculation-based multilinear regression (QC-MLR, RMSE of 0.17 and 0.12) or ABSOLV (RMSE of 0.15 and 0.15).[6] The estimates for *S* (RMSE: 0.2) were also better than ABSOLV (RMSE: 0.22) and the same as the QC-MLR (RMSE: 0.2).[6] The estimates for *A* were comparable among the three methods (RMSE: 0.09−0.07). Similarly, other studies also found that the PaDEL representation achieved good predictions for physical properties.[52]

Reduced PaDEL-DNN Models Based on Dimension-Reduced Inputs. To simplify model interpretation and application, three new reduced PaDEL-DNN models (M1−M3) were built to reduce the input dimension by applying the input−input, input−output, and LASSO dimension reduction methods, separately. It was found that dimension reduction by the input−input (M1) or input−output (M2) coefficients almost always increased the estimation errors (Figure 2c,d). Only the LASSO method (M3) achieved comparable performance with the PaDEL-DNN models (Figure 2). However, one problem emerged when M3 was employed to estimate the pp-LFER descriptors for the chemicals in the four chemical transfer data sets. For example, only 132 out of the 153 reduced PaDEL features can be obtained for all of the chemicals in the $\log K_{\text{storage-lipid/water}}$ data set—the other 21 of the 153 features cannot be calculated for some chemicals. Because the training and test data sets must have the same number of input features to ensure proper model training/estimation, M3 has to be trained by reducing the number of input features from 153 to 132. As each feature in dimension-reduced inputs likely captures certain critical chemical information (otherwise, they would not have been selected), any missing feature would inevitably lead to key information loss and hence worse estimates for solute descriptors. Therefore, using the M3-estimated pp-LFER descriptors to calculate $\log K_{\text{storage-lipid/water}}$ achieved worse performance than using the PaDEL-DNN models' estimated pp-LFER descriptors (MSE increased from 0.117 to 0.155).

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

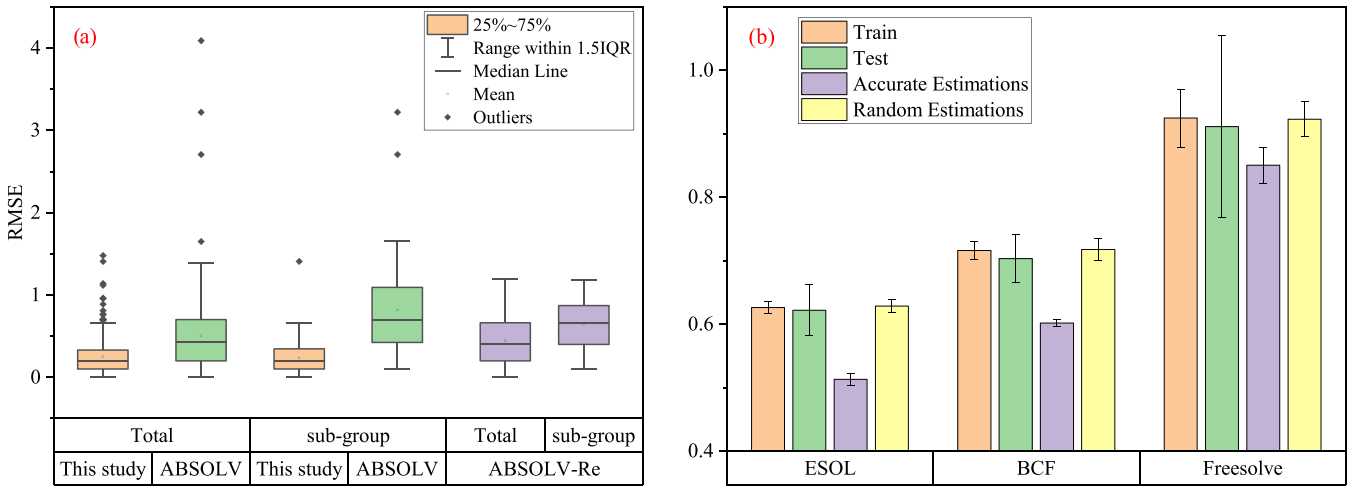


Figure 3. (a) Comparison of RMSEs in predicting log $K_{storage\text{-}lipid/water}$ based on the pp-LFER descriptors estimated by the PaDEL-DNN models versus by ABSOLV. "Total" and "subgroup" indicate the chemical group with 305 and 51 chemicals, separately. ABSOLV-Re represents the results excluding extremely poor predictions. (b) Modeling performance (RMSE) for different data sets using the PaDEL-DNN-estimated pp-LFER descriptors. Train and Test represent the training and test data sets after data splitting without applying the surrogate metric. Accurate Estimations and Random Estimations have the same number of chemicals but consider chemicals having accurate pp-LFER descriptors and randomly selected chemicals from each data set, respectively.

We further indirectly evaluated the PaDEL-DNN models by examining how well the PaDEL-DNN-estimated pp-LFER descriptors can model external chemicals in additional four chemical data sets. This approach differs significantly from many QSARs studies that focus only on one data set for the model development/validation and reduce the input dimension as much as possible. Many chemicals in the four chemical data sets have no experimental pp-LFER descriptors but may be well estimated by the PaDEL-DNN models. However, if we only consider chemicals in the LFER data set during the input dimension reduction, some selected features may not always be available for other chemicals. As a result, rebuilt models (M3) may perform well on the LFER data set but not on other data sets due to missing features. As for the PaDEL-DNN models that used all of the features in the PaDEL representation, some features may seem repetitive for the LFER data set, but they may serve as backups for missing features, which provide the PaDEL-DNN models some tolerance to missing values. In fact, the worse performance of the reduced PaDEL-DNN models is similar to the reported model for diabetic retinopathy detection, where the built model achieved satisfactory performance during the model development but did not perform well in real-world clinical settings when some essential input information was missing due to nonideal environmental factors.[53]

Overall, satisfactory models need to consider not only the data sets used for the modeling but also possible nonideal conditions during model applications. Otherwise, a model that achieves excellent accuracy on the training/test data sets may become less favorable during real applications. Therefore, all features in the PaDEL representation are used in the following sections.

Comparison with ABSOLV in Estimating the pp-LFER Descriptors for New Chemicals. The above comparison among different models was mostly based on known pp-LFER descriptors or the reserved test data sets. The primary usage of models is however to make predictions for unknowns; thus, it becomes important to examine how the models perform on new chemicals. Unfortunately, the lack of experimental descriptors for chemicals makes it impossible to directly compare different models based on RMSEs or R^2. We propose to address this problem through indirect comparison—examining the modeling performance of the estimated pp-LFER descriptors on known chemical transfer processes. The more accurate the estimated pp-LFER descriptors, the lower the errors for the subsequently modeled chemical transfer processes. Following this idea, the PaDEL-DNN models were compared with the ABSOLV based on the prediction errors for log $K_{storage\text{-}lipid/water}$ using the pp-LFER descriptors estimated by either model. We first estimated the pp-LFER descriptors for 305 chemicals using the PaDEL-DNN models. The estimated descriptors were then employed to calculate their log $K_{storage\text{-}lipid/water}$ values and compared with the experimental values.[42] The overall RMSE for the calculated log $K_{storage\text{-}lipid/water}$ was 0.34 log unit, much smaller than when using the ABSOLV-estimated pp-LFER descriptors (RMSE = 0.61). When the comparison was made on a subgroup of 305 chemicals (mostly H-bond donor substances which were harder to predict, N = 51) or additional 22 complex chemicals with more than one functional group, the PaDEL-DNN-estimated pp-LFER descriptors still showed considerable improvement (RMSE = 0.35 or 0.89) over the ABSOLV-estimated pp-LFER descriptors (RMSE = 0.91 or 1.29).[31] Note that these errors are also mostly smaller than those for directly calculated log $K_{storage\text{-}lipid/water}$ by KOWWIN, SPARC, or COSMOtherm, with the RMSE of 0.6, 0.54, and 0.45, respectively, for all 305 chemicals; 0.84, 0.42, and 0.35, respectively, for 51 chemicals; and 1.6, 1.25, and 0.71, respectively, for 22 chemicals.[31]

In addition, prediction errors in log $K_{storage\text{-}lipid/water}$ using the PaDEL-DNN-estimated pp-LFER descriptors showed lower variance (Figure 3a) than those by ABSOLV in both 305-full and 51-subgroup log $K_{storage\text{-}lipid/water}$ data sets. Even with the worst predictions being excluded, prediction errors by PaDEL-DNN (RMSE = 0.33 and 0.32 log unit for the total and subgroup separately) were still much smaller than those by ABSOLV (RMSE = 0.54 and 0.71, ABSOLVE-Re in Figure 3a). These comparisons well suggest the potential of PaDEL-

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

Table 1. RMSEs of the Models Using the PaDEL-DNN-Estimated pp-LFER Descriptors with/without Applying the Surrogate Metric and Comparison with Reported Model Performance[a]

data set (# of chemicals)	train	test	surrogate metric	random selection	reported results
log $K_{storage\text{-}lipid/water}$ (305)	*	0.34	0.3 (274)	0.5	0.45−0.61[31]
ESOL (1128)	0.63 ± 0.01	0.62 ± 0.04	0.51 ± 0.01(865)	0.63 ± 0.01	0.5 and 0.6[32]
BCF (1034)	0.72 ± 0.01	0.70 ± 0.04	0.60 ± 0.01(767)	0.72 ± 0.02	0.77 and 0.84[55]
Freesolve (639)	0.92 ± 0.05	0.91 ± 0.14	0.85 ± 0.03(549)	0.92 ± 0.03	1.2 and 1.5[32]

[a]Note: Train and Test represent training and test data sets after data splitting without applying the surrogate metric. *: There is a reported model for log $K_{storage\text{-}lipid/water}$; thus, no chemicals were used to train models, and all 305 chemicals were used in the test data set.

DNN to accurately estimate pp-LFER descriptors and the capability of these newly estimated pp-LFER descriptors to model chemical transfer processes, as discussed below.

Modeling Three Chemical Transfer Processes Using Estimated pp-LFER Descriptors. To further evaluate the modeling capability of the PaDEL-DNN-estimated pp-LFER descriptors, we used them to model bioconcentration factor (BCF), ESOL, and freesolve. BCF is a useful parameter for evaluating the potential risk of chemicals in the aquatic environment. However, BCF is generally obtained through time-consuming experiments and is only available for a very small portion of chemicals. ESOL aims to estimate the aqueous solubility of a chemical directly from its structure, while freesolve aims to estimate the hydration free energy of small molecules in water. ESOL and freesolve have become benchmark data sets for evaluating chemical descriptors.[32,33] Because all of these processes involve the transfer of chemicals between two phases, the pp-LFER descriptors should be able to capture molecular-level interactions in these processes. For example, it was found that models incorporating the pp-LFER descriptors can well predict BCF (N = 305, R^2 = 0.72), and the prediction was much better than models without incorporating the pp-LFER descriptors (R^2 from 0.52 to 0.71 for a subgroup of 305 chemicals with log P values between 4 and 5).[54]

The above BCF modeling was however only applied to a limited number of chemicals due to the lack of pp-LFER descriptors. With the PaDEL-DNN model-estimated pp-LFER descriptors, we expanded the modeling of BCF to a new data set of 1034 chemicals, and the results showed consistently good performance on training and test data sets (RMSE = 0.72 ± 0.01 vs 0.70 ± 0.04). Such performance is quite satisfactory and even better than that of the most often used CAESAR (RMSE = 0.84, N = 851) or Meylan (RMSE = 0.77, N = 851) models.[55,56] The PaDEL-DNN-estimated pp-LFER descriptors also achieved satisfactory prediction for ESOL and freesolve (Table 1). For ESOL, the performance (RMSE = 0.63−0.62) is comparable with the widely used benchmark prediction (RMSE ≈ 0.6) using complex graphic neural networks. The prediction errors for training and test data sets (RMSE = 0.92 ± 0.05 and 0.91 ± 0.14) of the freesolve were also considerably smaller than the benchmark result (RMSE ≈ 1.2) and *ab initio* predictions (RMSE = 1.5).[32] The consistent prediction between training and test data sets suggests the generalization ability of models using the estimated pp-LFER descriptors. Besides, the simplicity (only five variables) and clear physical meanings of the pp-LFER descriptors will make the model interpretation convenient, which cannot be easily achieved using other chemical descriptors.

Surrogate Metric. For new chemicals, knowing the accuracy of the estimated pp-LFER descriptors is highly desirable no matter which model is used. The surrogate metric may provide a convenient, effective way to evaluate the accuracy of estimated pp-LFER descriptors. To validate this approach, we first used it to evaluate the performance of the estimated pp-LFER descriptors in modeling log $K_{storage\text{-}lipid/water}$ for 305 chemicals. Based on the surrogate metric, we obtained "accurate" and "inaccurate" pp-LFER descriptors for 274 and 31 chemicals, respectively; using these descriptor values to estimate log $K_{storage\text{-}lipid/water}$ had an RMSE value of 0.3 (N = 274) and 0.5 (N = 31) log unit, respectively. It is clear that using the accurate pp-LFER descriptors yielded more accurate log$K_{storage\text{-}lipid/water}$ values.

We then tested the surrogate metric on BCF, ESOL, and freesolve data sets (Figure 3b). In this test, only chemicals with accurate pp-LFER descriptors were selected for the modeling. The fivefold cross-validation results showed that all of the RMSEs were considerably reduced (Table 1, Surrogate metric). The prediction for BCF can compare well with the reported integrated models.[55] The prediction for ESOL was better than the benchmark prediction and almost equal to the *ab initio* prediction (SD = 0.5).[32] The modeling for freesolve data set also achieved improvement after applying the surrogate metric. Meanwhile, for randomly chosen chemicals, there were negligible changes in the prediction accuracy (Table 1, Random selection) compared with those using all available chemicals. These comparisons strongly suggest that the surrogate metric can indeed select chemicals whose pp-LFER descriptors are accurately estimated, so using these estimated descriptors can improve the prediction accuracy of the models. Overall, these results support the proposed surrogate metric to evaluate the overall quality of the estimated pp-LFER descriptors.

Post Hoc Interpretation of the PaDEL-DNN Models. Although considerable improvements have been achieved in estimating the pp-LFER descriptors through combining PaDEL-DNN with the surrogate metric, it is crucial to understand how the PaDEL-DNN models make predictions using the learned chemical information. In QSARs applications, the chemical similarity between a target chemical and the chemicals employed in the QSARs development is generally believed to be the key to good model performance.[44] This is also true in chemical-related machine learning models.[45,46] The calculated chemical similarity mostly considers the entire molecules; however, such an overall similarity does not always apply to the pp-LFER descriptors because E relies on the entire chemical structure while S, A, and B are highly related to specific functional groups, such as polar or hydrogen-bond accepting or donating groups. In other words, the level of similarity between two chemicals may vary from one pp-LFER descriptor to another. However, the traditional molecular fingerprints-based chemical similarity (e.g., Tanimoto similarity[57]) between two chemicals is the same regardless of the pp-LFER descriptors. To find a good similarity metric for all of the pp-LFER descriptors, we compared different similarity

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

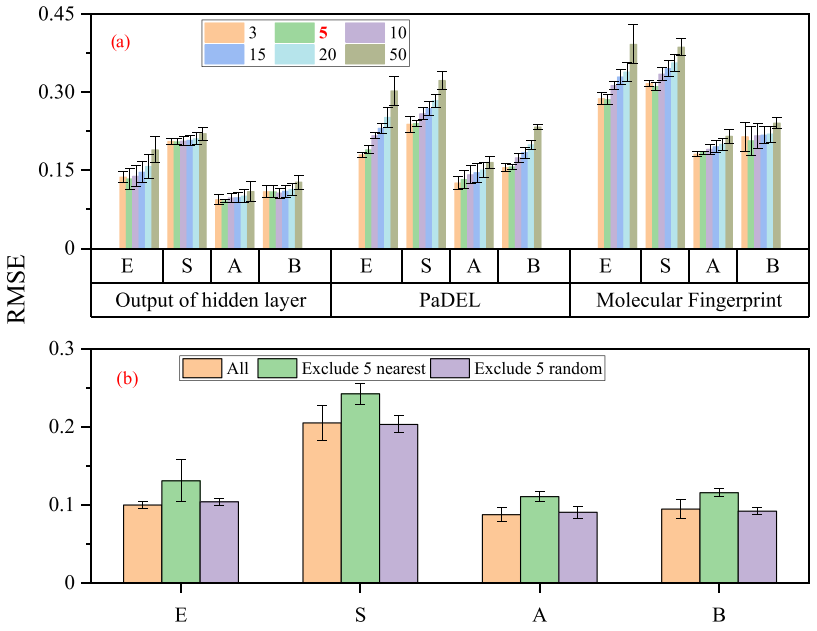


Figure 4. Post hoc interpretation of the PaDEL-DNN models. (a) Estimation errors for the pp-LFER descriptors based on three *K*-nearest neighbor models. (b) Based on the output of the last hidden layer (exclude five nearest) or random selection (exclude five random), estimation errors for the pp-LFER descriptors by the retrained PaDEL-DNN models with five chemicals excluded from the training data set for each test chemical. "All" estimation errors are based on the full PaDEL-DNN models.

strategies by examining the performance of three *K*-nearest neighbor models (details in Text S5.1).

The first model (output of the hidden layer, Figure 4a) used the output of the last hidden layer[58,59] of the PaDEL-DNN models to select *K*-nearest neighbors. The output of the last hidden layer is directly related to the prediction target and is generally believed to capture the essential knowledge of a DNN model.[59] The second model (PaDEL) and the third model (Molecular Fingerprint) selected *K*-nearest neighbors based on the PaDEL representation and the Tanimoto similarity between molecular fingerprints,[60] respectively. The results (Figure 4a) indicated that the first model performed the best for all of the tested *K* numbers and all of the pp-LFER descriptors. The second method performed better than the third one, which agrees with the better performance of the PaDEL representation than molecular fingerprints during the DNN modeling. The better performance of the PaDEL representation here is probably because it considered both local and global chemical information, which is beneficial for estimating the pp-LFER descriptors.

The performance of the first *K*-nearest models, although not as good as the PaDEL-DNN models, is still quite satisfactory considering the simplicity of the models. This good performance demonstrated that the output from the last hidden layer of the PaDEL-DNN models was indeed highly related to the pp-LFER descriptors. In other words, these PaDEL-DNN models have learned critical chemical information for estimating the pp-LFER descriptors. With the increasing number of *K*-nearest chemicals for the first models (Figure 4a), estimation errors first decreased slightly or remained stable and then increased. This trend could be because a good prediction needs enough chemical information, whereas too many chemicals would inevitably introduce some noise or less relevant information to make the prediction worse. Around five nearest chemicals may be the optimum for the first *K*-nearest models considering both the accuracy and the model complexity.

The good performance of the first *K*-nearest models raised another question about whether other less similar training chemicals also contributed to the estimation or not. To address this question, the influence function idea[47] was applied to the PaDEL-DNN models. This approach traces a model's prediction back to the training data set by comparing model predictions with/without certain training data records, thereby identifying the training records that are most responsible for given predictions.[47] For each test chemical, we focused on the *K*-nearest training chemicals and compared the predictions by the PaDEL-DNN models with/without those nearest training chemicals. One PaDEL-DNN model per pp-LFER descriptor was retrained after excluding five nearest chemicals in the training data set for each test chemical. As a control to account for the possible influence of a slightly smaller training data set, the PaDEL-DNN models were also retrained by excluding five random chemicals. When the five nearest chemicals were excluded, the estimation errors by the retrained PaDEL-DNN models (Figure 4b) were considerably greater than those by the original PaDEL-DNN models. However, the dropping of five random chemicals showed negligible influence on the estimation errors. These results suggest that the five nearest chemicals are important in estimating the pp-LFER descriptors and the PaDEL-DNN models indeed relied on similar chemicals to make accurate predictions. Nevertheless, the retrained PaDEL-DNN models (exclude five nearest) still maintained considerable prediction capability. In other words, the PaDEL-DNN models rely on both the most similar chemicals and all other chemicals in training data sets to make predictions. A large number of less similar training chemicals provide a reasonable baseline estimate, while a few most similar chemicals can considerably improve the overall accuracy. Therefore, the PaDEL-DNN models indeed learned

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

essential chemical information from all of the training chemicals.

Application of the PaDEL-DNN Models to PBT Chemicals. As mentioned in the Introduction section, there is a large gap between the limited number of the pp-LFER descriptors and the ever-increasing number of organic chemicals. The well-trained PaDEL-DNN models together with the surrogate metric can help narrow this gap, by not only accurately estimating the pp-LFER descriptors but also evaluating the overall goodness of the estimates. Thus, we used the PaDEL-DNN models and the surrogate metric to predict and evaluate the pp-LFER descriptors for over 4000 PBT chemicals. Most of these chemicals have few experimental properties available notwithstanding the pp-LFER descriptors. The lack of known physicochemical properties makes it difficult to assess their possible environmental behaviors or risks. Accurate estimates of the pp-LFER descriptors for these PBT chemicals would greatly help address these problems. After we applied the PaDEL-DNN models and the surrogate metric to these chemicals (Text S6), the estimated pp-LFER descriptors are within the threshold of 0.5 log units for 1798 chemicals and are between 1 and 2 times the threshold (1 log unit) for additional 1095 chemicals. We believe that future modeling would benefit tremendously from these estimated pp-LFER descriptors.

To make the PaDEL-DNN models easily accessible to users who may not have ample modeling experiences, we developed a web predictor (Figure S14, code and user guide uploaded to GitHub). For a new chemical, predictions can be achieved by simply inputting the SMILES and clicking "Submit," and predictions and surrogate metric results for the pp-LFER descriptors will be displayed in a table.

Despite the promising results from the PaDEL-DNN models and the surrogate metric, there are still three major limitations: (1) This research only covered neutral chemicals while there are a large number of ionizable chemicals. However, it is difficult to build predictive models for charged chemicals because (a) the available *E*, *S*, *A*, and *B* values are mostly for neutral chemicals; (b) there are different opinions for charged chemicals regarding the *E*, *S*, *A*, *B*, and *V* values. Some researchers believe that the influences of charge in chemicals can be described by adding the J^-/J^+ terms,[61] whereas others think that the presence of charge will also change the values of *E*, *S*, *A*, *B*, and *V*.[62] (c) There are many missing values among the features of charged chemicals when deriving PaDEL/RDkit chemical representations. For most models, we need to have the same number of input features for all of the chemicals, but because of the missing values, we will have to discard some important features for neutral chemicals to ensure the uniform length of inputs when combining neutral and charged chemicals in one model. (2) Although the PaDEL-DNN models and the surrogate metric could accurately estimate the pp-LFER descriptors for many new chemicals, descriptors for many other chemicals still cannot be accurately estimated. The post hoc model analysis found that the PaDEL-DNN models relied on chemicals with similar functional groups when making predictions. The LFER data set, although covering nearly 2000 chemicals, is still limited in the diversity of chemical structures. Selectively performing experiments/computation on some chemicals will be needed for further expanding the applicability of the PaDEL-DNN models. (3) Although the surrogate metric was used to replace traditional ADs and achieved some promising results, one should realize that the surrogate metric evaluates the overall prediction for solute descriptors rather than one by one. A good prediction of solute descriptors based on the surrogate metric does not mean that all of the descriptors are equally well predicted. In addition, the end point (e.g., logP) of the surrogate metric is different from the prediction target (e.g., solute descriptors). Although they are similar, there are always some differences among them such that the accuracy evaluation based on the surrogate metric cannot entirely replicate the prediction accuracy in the target.

SIGNIFICANCE

In this research, the pp-LFER descriptors were accurately predicted by the PaDEL-DNN models—requiring only SMILES—and the surrogate metric. During modeling, it was found that PaDEL and RDKit representations achieved better performance than the molecular fingerprints. Also, proper input dimension reduction may not affect the model performance on the LFER data set but would make worse estimates on external data sets. This suggests that eliminating input covariables as many as possible may limit the applicability of the built models. Future chemical-related modeling should consider both local and global chemical information and should be able to handle chemicals beyond the initial data sets.

The PaDEL-DNN-estimated pp-LFER descriptors demonstrated promising modeling performance on four chemical transfer data sets. The simplicity and interpretability as well as the satisfactory modeling performance make the pp-LFER descriptors promising chemical descriptors in modeling many other processes. The proposed surrogate metric reduced the RMSEs by around 0.1 log units for chemical transfer modeling. We believe that the surrogate metric provides a new, simple way to evaluate the model performance when applying a model to new targets that do not have available data for a direct evaluation.

The interpretation of the PaDEL-DNN models provided useful insights into how PaDEL-DNN used the training chemical information to estimate test chemicals. By understanding the contribution of the bulk, less similar chemicals versus that of a few most similar chemicals to the model predictions, we provided a new direction for improving the model performance, that is, in addition to increasing the sample size of training data sets, we need to employ a carefully selected similarity metric to select some nearest chemicals (increasing the chemical similarity) in the training data set for a target chemical. Overall, this research will greatly expand chemical modeling not only in the environmental field but likely in many other disciplines.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.1c05398.

More detailed explanations of the methods and training process in this study, the figures and tables mentioned in the main text, and additional figures and tables to support the training process (PDF)

Source data compiled for this research (XLSX)

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064

AUTHOR INFORMATION

Corresponding Author

Huichun Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0002-5683-5117; Phone: (216) 368-0689; Email: hjz13@case.edu

Author

Kai Zhang − *Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, Ohio 44106, United States;* orcid.org/0000-0003-4058-6512

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.1c05398

Author Contributions

The web predictor and user guide can be found at: https://github.com/cwrukaizhang/LFERsPredictor.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant #CBET-1804708.

REFERENCES

(1) Zhong, S.; Hu, J.; Fan, X.; Yu, X.; Zhang, H. A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants. *J. Hazard. Mater.* **2020**, *383*, No. 121141.

(2) Zhong, S.; Zhang, K.; Wang, D.; Zhang, H. Shedding light on "Black Box" machine learning models for predicting the reactivity of HO radicals toward organic compounds. *Chem. Eng. J.* **2021**, *405*, No. 126627.

(3) Zhong, S.; Hu, J.; Yu, X.; Zhang, H. Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation. *Chem. Eng. J.* **2021**, *408*, No. 127998.

(4) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; Shi, W.; Tan, H.; Wang, T.; Wang, X.; Wong, B. M.; Xiao, X.; Yu, X.; Zhu, J.-J.; Zhang, H. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741−12754.

(5) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6*, 1379−1390.

(6) Liang, Y.; Xiong, R.; Sandler, S. I.; Di Toro, D. M. Quantum Chemically Estimated Abraham Solute Parameters Using Multiple Solvent-Water Partition Coefficients and Molecular Polarizability. *Environ. Sci. Technol.* **2017**, *51*, 9887−9898.

(7) Zhang, K.; Zhang, H. Coupling a Feedforward Network (FN) Model to Real Adsorbed Solution Theory (RAST) to Improve Prediction of Bisolute Adsorption on Resins. *Environ. Sci. Technol.* **2020**, *54*, 15385−15394.

(8) Zhang, K.; Zhong, S.; Zhang, H. J. Predicting aqueous adsorption of organic compounds onto biochars, carbon nanotubes, granular activated carbons, and resins with machine learning. *Environ. Sci. Technol.* **2020**, *54*, 7008−7018.

(9) Schwarzenbach, R. P.; Gschwend, P. M.; Imboden, D. M. *Environmental Organic Chemistry*, 3rd ed.; John Wiley & Sons: Hoboken, New Jersey, 2016; p 1024.

(10) Shih, Y. H.; Gschwend, P. M. Evaluating activated carbon-water sorption coefficients of organic compounds using a linear solvation energy relationship approach and sorbate chemical activities. *Environ. Sci. Technol.* **2009**, *43*, 851−857.

(11) Abraham, M. H.; Ibrahim, A.; Acree, W. E., Jr Air to lung partition coefficients for volatile organic compounds and blood to lung partition coefficients for volatile organic compounds and drugs. *Eur. J. Med. Chem.* **2008**, *43*, 478−485.

(12) Grisoni, F.; Consonni, V.; Vighi, M. Detecting the bioaccumulation patterns of chemicals through data-driven approaches. *Chemosphere* **2018**, *208*, 273−284.

(13) Arey, J. S.; Green, W. H.; Gschwend, P. M. The electrostatic origin of Abraham's solute polarity parameter. *J. Phys. Chem. B* **2005**, *109*, 7564−7573.

(14) Endo, S.; Goss, K. U. Applications of polyparameter linear free energy relationships in environmental chemistry. *Environ. Sci. Technol.* **2014**, *48*, 12477−12491.

(15) Ulrich, N.; Endo, S.; Brown, T.; Watanabe, N.; Bronner, G.; Abraham, M.; Goss, K. *UFZ-LSER Database v 3.2.1 [Internet]*; Helmholtz Centre for Environmental Research-UFZ: Leipzig, Germany, 2017.

(16) Acree, W. E.; Che, M.; Lee, G.; Abraham, M. H. Calculation of the Abraham model solute descriptors for the pharmaceutical compound acipimox based on experimental solubility data. *Phys. Chem. Liq.* **2019**, *57*, 382−387.

(17) Abraham, M. H.; Acree, W. E., Jr Determination of the hydrogen-bond acidity and basicity for un-dissociated hydrazoic acid, isocyanic acid and isothiocyanic acid. *J. Mol. Liq.* **2019**, *294*, No. 111666.

(18) Abraham, M. H.; Acree, W. E., Jr; Cometto-Muniz, J. E. Descriptors for terpene esters from chromatographic and partition measurements: Estimation of human odor detection thresholds. *J. Chromatogr. A* **2020**, *1609*, No. 460428.

(19) Poole, C. F. Wayne State University experimental descriptor database for use with the solvation parameter model. *J. Chromatogr. A* **2020**, *1617*, No. 460841.

(20) Brown, T. N. Predicting hexadecane-air equilibrium partition coefficients (L) using a group contribution approach constructed from high quality data. *SAR QSAR Environ. Res.* **2014**, *25*, 51−71.

(21) Platts, J. A.; Butina, D.; Abraham, M. H.; Hersey, A. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.

(22) Rahaman, O.; Doren, D. J.; Di Toro, D. M. Quantum mechanical estimation of Abraham hydrogen bond parameters using 1:1 donor-acceptor complexes. *J. Phys. Org. Chem.* **2014**, *27*, 783−793.

(23) Davis, C. W.; Di Toro, D. M. Predicting solvent-water partitioning of charged organic species using quantum-chemically estimated Abraham pp-LFER solute parameters. *Chemosphere* **2016**, *164*, 634−642.

(24) Schwöbel, J. A. H.; Ebert, R.-U.; Kühne, R.; Schüürmann, G. Prediction models for the Abraham hydrogen bond donor strength: comparison of semi-empirical, ab initio, and DFT methods. *J. Phys. Org. Chem.* **2011**, *24*, 1072−1080.

(25) Bauer, C. A.; Schneider, G.; Göller, A. H. Machine learning models for hydrogen bond donor and acceptor strengths using large and diverse training data generated by first-principles interaction free energies. *J. Cheminf.* **2019**, *11*, 1−16.

(26) Liang, Y.; Kuo, D. T.; Allen, H. E.; Di Toro, D. M. Experimental determination of solvent-water partition coefficients and Abraham parameters for munition constituents. *Chemosphere* **2016**, *161*, 429−437.

(27) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012**, *17*, 4791−810.

(28) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.

(29) Landrum, G.. *Rdkit Documentation. Release*, 2013.

(30) Grisoni, F.; Consonni, V.; Vighi, M.; Villa, S.; Todeschini, R. Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environ. Int.* **2016**, *88*, 198−205.

(31) Geisler, A.; Oemisch, L.; Endo, S.; Goss, K.-U. Predicting storage−lipid water partitioning of organic solutes from molecular structure. *Environ. Sci. Technol.* **2015**, *49*, 5538−5545.
(32) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(33) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711−720.
(34) Petoumenou, M. I.; Pizzo, F.; Cester, J.; Fernandez, A.; Benfenati, E. Comparison between bioconcentration factor (BCF) data provided by industry to the European Chemicals Agency (ECHA) and data derived from QSAR models. *Environ. Res.* **2015**, *142*, 529−34.
(35) Grisoni, F.; Consonni, V.; Villa, S.; Vighi, M.; Todeschini, R. QSAR models for bioconcentration: is the increase in the complexity justified by more accurate predictions? *Chemosphere* **2015**, *127*, 171−9.
(36) Sun, X.; Zhang, X.; Muir, D. C.; Zeng, E. Y. Identification of Potential PBT/POP-Like Chemicals by a Deep Learning Approach Based on 2D Structural Features. *Environ. Sci. Technol.* **2020**, *54*, 8221−8231.
(37) Zhang, X.; Sun, X.; Jiang, R.; Zeng, E. Y.; Sunderland, E. M.; Muir, D. C. Screening new persistent and bioaccumulative organics in China's inventory of industrial chemicals. *Environ. Sci. Technol.* **2020**, *54*, 7398−7408.
(38) Strempel, S.; Scheringer, M.; Ng, C. A.; Hungerbühler, K. Screening for PBT chemicals among the "existing" and "new" chemicals of the EU. *Environ. Sci. Technol.* **2012**, *46*, 5680−5687.
(39) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.* **2020**, *6*, No. eaaz4301.
(40) Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal Stat. Society: Series B* **1996**, *58*, 267−288.
(41) Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 3932−3937.
(42) Geisler, A.; Endo, S.; Goss, K.-U. Partitioning of organic chemicals to storage lipids: Elucidating the dependence on fatty acid composition and temperature. *Environ. Sci. Technol.* **2012**, *46*, 9519−9524.
(43) Quiñonero-Candela, J.; Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *J. Machine Learning Res.* **2005**, *6*, 1939−1959.
(44) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1044.
(45) Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2019**, *59*, 181−189.
(46) Gao, Y.; Zhong, S.; Torralba-Sanchez, T. L.; Tratnyek, P. G.; Weber, E. J.; Chen, Y.; Zhang, H. Quantitative structure activity relationships (QSARs) and machine learning models for abiotic reduction of organic compounds by an aqueous Fe(II) complex. *Water Res.* **2021**, *192*, No. 116843.
(47) Koh, P. W.; Liang, P. In *Understanding Black-Box Predictions via Influence Functions*, International Conference on Machine Learning, Sydney, Aastralia, 2017; International Machine Learning Society (IMLS): Sydney, Aastralia, 2017; pp 1885−1894.
(48) Wang, Z.; Su, Y.; Jin, S.; Shen, W.; Ren, J.; Zhang, X.; Clark, J. H. A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties. *Green Chem.* **2020**, *22*, 3867−3876.
(49) 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.
(50) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(51) Todeschini, R.; Lasagni, M.; Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom.* **1994**, *8*, 263−272.
(52) Stepišnik, T.; Škrlj, B.; Wicker, J.; Kocev, D. A comprehensive comparison of molecular feature representations for use in predictive modeling. *Comput. Biol. Med.* **2021**, *130*, No. 104197.
(53) Beede, E.; Baylor, E.; Hersch, F.; Iurchenko, A.; Wilcox, L.; Ruamviboonsuk, P.; Vardoulakis, L. M. In *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy*, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu,HI, USA, 2020; Association for Computing Machinery: Honolulu, HI, USA, 2020; pp 1−12.
(54) Zhao, S.; Jones, K. C.; Sweetman, A. J. Can poly-parameter linear-free energy relationships (pp-LFERs) improve modelling bioaccumulation in fish? *Chemosphere* **2018**, *191*, 235−244.
(55) Gissi, A.; Nicolotti, O.; Carotti, A.; Gadaleta, D.; Lombardo, A.; Benfenati, E. Integration of QSAR models for bioconcentration suitable for REACH. *Sci. Total Environ.* **2013**, *456−457*, 325−332.
(56) Ai, H.; Wu, X.; Zhang, L.; Qi, M.; Zhao, Y.; Zhao, Q.; Zhao, J.; Liu, H. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicol. Environ. Saf.* **2019**, *179*, 71−78.
(57) Bajusz, D.; Racz, A.; Heberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, No. 20.
(58) Caruana, R.; Kangarloo, H.; Dionisio, J. D.; Sinha, U.; Johnson, D. In *Case-Based Explanation of Non-Case-Based Learning Methods*, Proceedings of the AMIA Symposium, 1999; American Medical Informatics Association, 1999; pp 212−215.
(59) Papernot, N.; McDaniel, P. *Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning*, 2008, arXiv:1803.04765. arXiv.org e-Print archive. https://arxiv.org/abs/1803.04765 (accessed on 13 March, 2008).
(60) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 1−13.
(61) Jadbabaei, N.; Zhang, H. Sorption Mechanism and Predictive Models for Removal of Cationic Organic Contaminants by Cation Exchange Resins. *Environ. Sci. Technol.* **2014**, *48*, 14572−14581.
(62) Abraham, M. H.; Acree, W. E., Jr Descriptors for ions and ion-pairs for use in linear free energy relationships. *J. Chromatogr. A* **2016**, *1430*, 2−14.

https://doi.org/10.1021/acs.est.1c05398
Environ. Sci. Technol. 2022, 56, 2054−2064