JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

# In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning

Qingda Zang,[†] Kamel Mansouri,[‡] Antony J. Williams,[‡] Richard S. Judson,[‡] David G. Allen,[†] Warren M. Casey,[∥] and Nicole C. Kleinstreuer*[,∥]

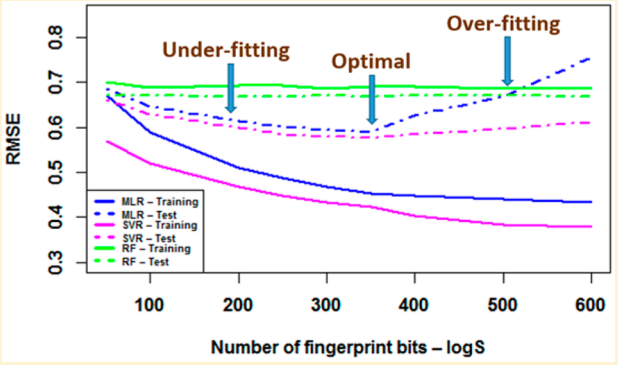
[†]Integrated Laboratory Systems, Inc., Research Triangle Park, North Carolina 27709, United States

[‡]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

[∥]National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States

Supporting Information

**ABSTRACT:** There are little available toxicity data on the vast majority of chemicals in commerce. High-throughput screening (HTS) studies, such as those being carried out by the U.S. Environmental Protection Agency (EPA) ToxCast program in partnership with the federal Tox21 research program, can generate biological data to inform models for predicting potential toxicity. However, physicochemical properties are also needed to model environmental fate and transport, as well as exposure potential. The purpose of the present study was to generate an open-source quantitative structure−property relationship (QSPR) workflow to predict a variety of physicochemical properties that would have cross-platform compatibility to integrate into existing cheminformatics workflows. In this effort, decades-old experimental property data sets available within the EPA EPI Suite were reanalyzed using modern cheminformatics workflows to develop updated QSPR models capable of supplying computationally efficient, open, and transparent HTS property predictions in support of environmental modeling efforts. Models were built using updated EPI Suite data sets for the prediction of six physicochemical properties: octanol−water partition coefficient (logP), water solubility (logS), boiling point (BP), melting point (MP), vapor pressure (logVP), and bioconcentration factor (logBCF). The coefficient of determination ($R^2$) between the estimated values and experimental data for the six predicted properties ranged from 0.826 (MP) to 0.965 (BP), with model performance for five of the six properties exceeding those from the original EPI Suite models. The newly derived models can be employed for rapid estimation of physicochemical properties within an open-source HTS workflow to inform fate and toxicity prediction models of environmental chemicals.




## INTRODUCTION

The U.S. Environmental Protection Agency (EPA) has identified ∼32 000 chemicals with the potential for human exposure, ranging from pesticides and industrial chemicals to food additives and personal care products.[1−6] As only a fraction of these chemicals have been fully assessed and characterized,[7−9] there is a need for more rapid and inexpensive approaches to prioritize thousands of chemicals for mechanistically relevant toxicity testing. The cross-agency U.S. federal Tox21 and EPA's ToxCast research programs have developed promising tools for chemical hazard characterization and prioritization, such as in vitro high-throughput screening (HTS) assays, computational toxicology approaches, and quantitative structure activity relationship (QSAR) models.[10−18] What these tools lack, however, is the ability to provide insight into the fate and transport of the chemicals, which is needed for chemical risk assessment.

The behavior of chemicals in humans and the environment often depends on some key physicochemical properties, such as octanol−water partition coefficient (logP), water solubility (logS), melting point (MP), boiling point (BP), vapor pressure (VP), and bioconcentration factor (BCF).[19−23] These properties affect bioavailability, permeability, absorption, transport, and persistence of chemicals in the body and in the environment and are used extensively in exposure, toxicological hazard, and risk assessments of organic chemicals. Certain properties, such as BCF, are required by regulations such as REACH (Registration, Evaluation, Authorization and Restric-

Received: October 15, 2016
Published: December 22, 2016






DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

Table 1. Distribution of Six Physicochemical Property Values for Training and Test Sets

| property | data set[a] | minimum | maximum | mean | median | standard deviation |
|---|---|---|---|---|---|---|
| logP | training (11370) | −5.40 | 11.29 | 2.07 | 1.99 | 1.83 |
| | test (2837) | −5.08 | 9.36 | 2.06 | 1.96 | 1.82 |
| logS | training (1507) | −12.06 | 1.58 | −2.56 | −2.40 | 2.15 |
| | test (503) | −11.25 | 1.35 | −2.71 | −2.39 | 2.30 |
| logBCF | training (456) | −0.35 | 5.97 | 1.88 | 1.71 | 1.25 |
| | test (152) | −0.30 | 5.82 | 1.90 | 1.68 | 1.29 |
| BP | training (4074) | −88.60 | 548.00 | 188.09 | 188.50 | 85.20 |
| | test (1358) | −84.70 | 536.00 | 188.66 | 190.80 | 84.69 |
| MP | training (6485) | −196.00 | 385.00 | 79.60 | 79.00 | 98.45 |
| | test (2163) | −187.00 | 385.00 | 82.60 | 83.00 | 101.09 |
| logVP | training (2034) | −13.68 | 5.67 | −2.01 | −1.22 | 3.58 |
| | test (679) | −11.80 | 4.72 | −2.15 | −1.40 | 3.56 |

[a]Numbers of chemicals in each set are indicated in parentheses.

tion of Chemicals) and the United Nations Globally Harmonized System of Classification and Labeling of Hazardous Chemicals.[24−26] These properties are often used in evaluating new or problematic chemicals and are valuable parameters in developing QSAR models for toxicity end points.[27−29] These parameters and models support initiatives being driven by the OECD (Organisation of Economic Cooperation and Development) and ICATM (International Cooperation on Alternative Test Methods) to reduce or waive animal tests using alternative methods such as in silico modeling.[30,31]

Whereas physicochemical properties have been experimentally determined for some chemicals, the majority lack freely available experimental data. In the USA, for new chemicals submitted for regulatory approval, such data are often considered confidential business information and are thus not available to regulatory authorities. Obtaining needed data via experimental measurements can be expensive and time-consuming, it may be difficult to handle hazardous or reactive chemicals, and some premanufacturing chemicals are unavailable for testing.

Quantitative structure−property relationship (QSPR) methods are designed to identify the relationship between the physicochemical property of interest and the chemical molecular structure without testing and are widely used to provide inputs for toxicity prediction models.[32−35] Numerous computational approaches have been proposed to construct QSPR models, and these methods can be generally categorized into three classes: models based on other experimentally determined physicochemical properties;[35−37] models based on calculated molecular descriptors;[32−34,38] and models based on group contributions.[39−43] The third approach was the foundation of initial work in QSPR development. In this approach, a molecule is divided into basic structural building blocks such as atoms or larger functional groups that constitute unique descriptors. Such methods are conceptually simple and computationally efficient for a wide range of chemicals because they only require counting occurrence of functional fragments in a molecule. Nevertheless, the model may suffer from the "missing fragment problem" when new fragments are encountered that were not available in the training set. This issue may be addressed by employing a large learning data set and considering the applicability domain (AD) of the model. The predictive performance of published models in the literature depends on the size, diversity, and composition of the data, rendering it difficult to make direct comparisons of models built using different training sets.[44]

EPA's Estimation Program Interface (EPI) Suite provides QSPR models to predict a variety of key fate and transport parameters for environmental chemicals.[45] However, the decades-old data sets upon which EPI Suite models were originally built contain numerous errors and the models used in predictions are not open-source. Accordingly, we used structure-curation and data-curation workflows to reanalyze and update EPI Suite experimental property data sets,[46,47] with which to build open source QSPR models capable of supplying computationally efficient and transparent high-throughput property predictions for environmental chemicals.

The goal of the present study was to develop QSPR models for in silico prediction of six physicochemical properties using diverse data sets of environmental chemicals exclusively based on analysis of their binary molecular fingerprints. We aimed to improve upon the existing EPI Suite platform and provide the community with an open-source model built in R that can be leveraged in workflows for several computational languages (e.g., Python and Java). We applied various computational methods, ranging from simple linear regression to sophisticated machine learning approaches and make recommendations on which models are more appropriate to predict each property. Ultimately, we sought to develop an open-source approach that provides reliable and accurate estimation of physicochemical properties for a wide range of environmental chemicals requiring only input of chemical structures in SMILES (Simplified Molecular Identification and Line Entry System) notation[48] to generate predictions. To facilitate their acceptance and application, we have adhered to the validation principles defined by the Organisation for Economic Cooperation and Development (OECD) when building and evaluating the QSPR models.[49] These models allow physicochemical property predictions to be readily generated and integrated with other types of information for regulatory and research purposes.

## MATERIALS AND METHODS

**Data Sets.** The experimentally measured physicochemical property values of structurally diverse sets of environmental chemicals used in this study were taken from a publicly available data source: Estimation Program Interface (EPI) Suite Data.[50] These organic chemicals cover a broad variety of product classes, including industrial chemicals, antimicrobials, colorants, fertilizers, flame retardants, fragrances, pharmaceut-



DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

icals, herbicides, pesticides, inert ingredients, petrochemicals and food additives (Table S1). Since most chemicals have multiple product classes, the built models cannot statistically distinguish for test chemicals from these different classes. Prior to modeling, the chemical information was processed by structure-curation and data-curation workflows developed at EPA to provide a QSAR-ready data set free from structural ambiguities (e.g., unconnected fragments, mixtures, salts, inorganics, duplicates), and to reconcile inconsistencies in the EPI Suite data set.[46,47] This process found and fixed a number of errors in the EPI Suite chemical library, and discarded chemicals that were not able to be fixed (e.g., entries for which the chemical name and structure could not be reconciled). All salts were stripped and then the main structure was neutralized when possible.[6]

The chemical names, CAS Registry Numbers, SMILES notations, and experimental data corresponding to the six properties are given in the Supporting Information. The data sets were randomly partitioned into training sets (75% of the chemicals) and test sets (25% of the chemicals) to build the models and externally validate their predictive power, respectively. Table 1 lists the summary statistics for the training and test sets. As shown in Figures 1 and S1, several property values are approximately normally distributed, where logP spans nearly 17 log units from −5.40 to 11.29 with a median of 1.99 while MP ranges from −196 to 385 °C and is centered at 79 °C. The data for logS, logVP, and logBCF are skewed since there is an upper limit to logS and logVP and a lower limit to logBCF.

**Molecular Fingerprints.** The chemicals were represented by fingerprints derived from their molecular structures. Fingerprints were calculated using a wide variety of publicly available SMARTS systems implemented in PaDEL:[51,52] Estate (79 bits), Extended (1024 bits), Substructure (307 bits), Klekota Roth (4860 bits), PubChem (881 bits), Atom Pairs 2D (780 bits), and MACCS (166 bits). A total of 8097 binary bits were generated, with 1 and 0 denoting the presence or absence, respectively, of a specific structural fragment. Fingerprint bits with zero variance (i.e., uniform observations across the set) were removed. To obtain reliable models, sufficient occurrences of the fingerprint bits throughout the entire data sets are necessary and thus, bits with low occurrences (<2%) were eliminated. Following removal of highly correlated and infrequently occurring bits, the resulting numbers of bits retained and employed to build the regression models were: 1681 for logP; 1061 for logS; 450 for logBCF; 1050 for BP; 1424 for MP; and 1145 for logVP. A genetic algorithm (GA)[53,54] was used to reduce the feature space by assigning an initial population of chromosomes to two times the number of variables (fingerprint bits). The crossover probability on each chromosome in a population and mutation rate on each gene in a chromosome were set to 50% and 1%, respectively. There were no improvements in the fitness score after 1000 generations.

**Multiple Linear Regression.** Multiple linear regression (MLR) is widely used in the modeling of property data.[40,55] We used MLR to produce a linear model to describe the relationship between a physicochemical property and the molecular fingerprint bits:

$$\text{property} = \sum_{j=1}^{m} c_j f_j \qquad (1)$$

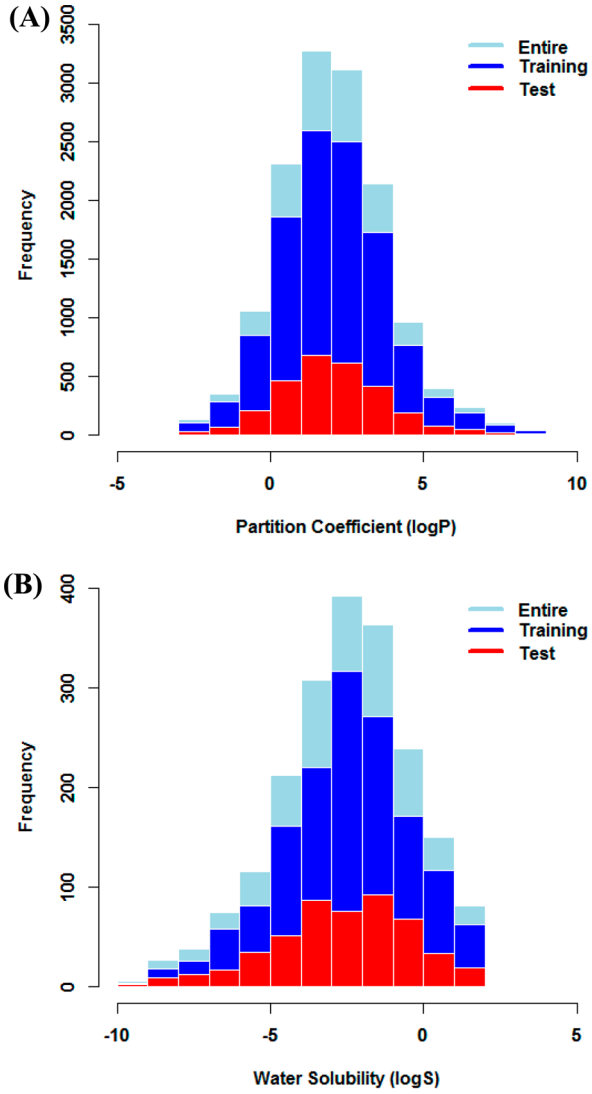



**Figure 1.** Data distribution of logP (A) and logS (B).

In eq 1, property is one of the six physicochemical properties (logP, logS, logBCF, BP, MP or logVP); $c_j$ is the contribution coefficient, which is determined by regression analysis; and $f_j$ is the binary bit of the $j$th fingerprint, with its presence or absence represented by the numeric values 1 or 0 respectively. Any fragment that occurred in a molecule was counted only once for that molecule, no matter how many times it occurred in the molecule.

**Partial Least Squares Regression.** Partial least-squares regression (PLSR) is a widely used multivariate analytical technique in QSPR studies.[56,57] The advantage of PLSR over MLR lies in its ability to build a regression model based on highly correlated descriptors, extract the relevant information, and reduce data dimensions. We employed PLSR to generate linear statistical models based on the fingerprint bits and the physicochemical property being predicted. A set of orthogonal latent variables or principal components (PCs) were first generated through a linear combination of the original molecular fingerprint bits, which served as new variables for regression with the response variables (i.e., the physicochemical properties) to build QSPR models. The optimal number of PCs was determined by 10-fold cross-validation (CV).



DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

**Random Forest Regression.** Random forest (RF) is a nonlinear consensus method based upon an ensemble of decision trees which are grown from separate bootstrap samples of the training data.[58] Bootstrap sampling is conducted via random selection with replacement from the training chemicals during tree growth. The chemicals that are not selected in the construction of the forest are called out-of-bag (OOB) samples, which are used to evaluate the prediction accuracy as trees are added to the forest. Each tree gives a prediction for its OOB chemicals, and the average of these results over all trees provides an overall unbiased external validation. There are three possible model parameters for RF regression: ntree—the number of trees in the forest; mtry—the number of variables randomly sampled at each tree node; and nodesize—the minimum node size below which nodes are not further subdivided. In the present study, the RF model was trained based upon a parameter combination of ntree = 500, nodesize = 5, and mtry = 1/3 the number of fingerprint bits.

**Support Vector Regression.** Support vector regression (SVR) models a nonlinear relationship between the property and molecular descriptors by utilizing an appropriate kernel function to map the input variables from a lower dimensional space to a higher dimensional feature space and transform the nonlinear relationship into a linear form.[35,59,60] An $\varepsilon$-insensitive loss function was used for the SVR modeling, in which the training chemical samples were represented as a tube with radius $\varepsilon$ and a Gaussian radial basis function (RBF) was employed as a kernel function. The accuracy of SVR relies on the optimization of the model parameters. An $\varepsilon$-based SVR analysis needs to tune the RBF kernel parameter $\gamma$, the radius of the tube $\varepsilon$, and the regularization parameter $C$ which determines the trade-off between model complexity and the training error. Thus, 10-fold CV via parallel grid search was performed in order to find the optimal combination of the three parameters.

**Model Validation.** The performance of each QSPR model was evaluated by examining the correlation between the experimental and predicted values using the following parameters:[61,62] $R^2$ (coefficient of determination) and RMSE (root mean squared error) for training or test sets with $n$ chemicals; $Q^2$ (coefficient of determination) and RMSEcv for 10-fold CV with $v$ chemicals not included in the CV model building set. The 10-fold CV procedure was completed using only the training set.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \tag{2}$$

$$Q^2 = 1 - \frac{\sum_{i=1}^{v}(p_i - \hat{p}_i)^2}{\sum_{i=1}^{v}(p_i - \bar{p})^2} \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2} \tag{4}$$

$$\mathrm{RMSEcv} = \sqrt{\frac{1}{v}\sum_{i=1}^{v}(p_i - \hat{p}_i)^2} \tag{5}$$

In eqs 2−5, $p_i$ and $\hat{p}_i$ are the measured and predicted property values for chemical $i$, respectively, and $\bar{p}$ is the mean of all chemicals in the data set. In addition, standard error of prediction (SEP) was employed as a criterion to select the optimal principal components in the PLSR analysis.[56]

$$\mathrm{SEP} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \hat{p}_i - \mathrm{bias})^2} \tag{6}$$

$$\mathrm{bias} = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i) \tag{7}$$

**Applicability Domain.** Three distance-based measures, i.e., leverage, distance from centroid and $k$-nearest neighbors (kNN), were applied to assess the applicability domain (AD) of each regression model. The distance of a test chemical from a defined point in the descriptor space of the training set was calculated and compared to a predefined threshold.[63,64] The test chemical is considered to be within AD if its distance is less than or equal to the threshold. Leverage is defined as the diagonal element of the covariance matrix for a given data set, and the leverage of a test chemical is proportional to Hotellings $T^2$ statistic and its Mahalanobis distance. The threshold was set to three times the average of the leverage ($3m/n$, with $m$ being the number of variables and $n$ the number of training chemicals). For the measure of distance from centroid, the distance of a test chemical from the training set centroid is compared with a threshold, which is determined as follows: (1) calculate the distances of training chemicals from their centroid; (2) sort the vector of distances in ascending order; (3) set the distance value corresponding to 95th percentile as the threshold. The kNN measure defines the model's AD based on the similarity between a test chemical and the training chemicals. The average distance of the test chemical from its five nearest neighbors in the training set is compared with a threshold, which is the 95th percentile of average distance of training chemicals from their five nearest neighbors.[65]

**Statistical Analysis.** Mathematical processing for data standardization, multivariate regression analysis, and statistical model building were performed using the *R* statistical computing environment for Windows (version 3.2.1).[66] Genetic algorithm, multiple linear regression, partial least-squares regression, random forest regression, support vector regression and distance of $k$-nearest neighbors were implemented by the *R* packages *subselect*, *stats*, *pls*, *randomForest*, *e1071*, and *FNN*, respectively. The *R* code for feature selection and regression analysis is provided in the Supporting Information.

**Online Resource.** The CompTox Dashboard from the EPA National Center for Computational Toxicology integrates experimental and predicted physicochemical end point data for over 700 000 individual chemical structures (https://comptox.epa.gov/dashboard/). It also meshes together other data including bioassay screening data, exposure models, and product categories. The dashboard contains tens of thousands of physicochemical property experimental data points that have been curated as described in thorough detail elsewhere.[47] These data have also been used to develop a new suite of prediction models across the database (the so-called OPEn saR Application (OPERA) models). Data curation is also underway for property end points, including toxicity values, which will be released to the community in later versions of the application. Integrated calculation reports are available for each property prediction associated with a chemical and these provide details regarding the performance of the prediction algorithm, associated information regarding whether the chemical is


DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

Table 2. Correlations ($r$) among Molecular Weight (MW) and the Six Physicochemical Properties[a]

| | MW | logP | logS | logBCF | BP | MP | logVP |
|---|---|---|---|---|---|---|---|
| MW | 1.000 | 0.256 | −0.648 | 0.367 | 0.475 | 0.460 | −0.721 |
| logP | | 1.000 | −0.873 | 0.830 | 0.365 | −0.043 | −0.387 |
| logS | | | 1.000 | −0.825 | −0.444 | −0.285 | 0.564 |
| logBCF | | | | 1.000 | 0.355 | 0.163 | 0.351 |
| BP | | | | | 1.000 | 0.733 | −0.959 |
| MP | | | | | | 1.000 | −0.833 |
| logVP | | | | | | | 1.000 |

[a]The number of chemicals for each pair: 1664 (logP vs logS); 482 (logP vs logBCF); 1609 (logP vs BP); 3531 (logP vs MP); 1560 (logP vs logVP); 330 (logS vs logBCF); 975 (logS vs BP); 1811 (logS vs MP); 1156 (logS vs logVP); 285 (logBCF vs BP); 473 (logBCF vs MP); 334 (logBCF vs logVP); 3421 (BP vs MP); 1775 (BP vs logVP); 2143 (MP vs logVP).

Table 3. Regression Statistics of LogP Using Subsets of Fingerprint Bits and MW

| variable | model statistics | data set | MLR | PLSR | RF | SVR |
|---|---|---|---|---|---|---|
| 1681 FP bits | $R^2$ | training | 0.916 | 0.915 | 0.880 | 0.991 |
| | | test | 0.879 | 0.878 | 0.876 | 0.920 |
| | RMSE | training | 0.509 | 0.510 | 0.552 | 0.176 |
| | | test | 0.607 | 0.608 | 0.564 | 0.502 |
| 600 FP bits | $R^2$ | training | 0.901 | 0.902 | 0.885 | 0.983 |
| | | test | 0.888 | 0.889 | 0.879 | 0.932 |
| | RMSE | training | 0.546 | 0.546 | 0.548 | 0.264 |
| | | test | 0.576 | 0.575 | 0.569 | 0.457 |
| 600 FP bits + MW | $R^2$ | training | 0.904 | 0.905 | 0.891 | 0.987 |
| | | test | 0.891 | 0.892 | 0.886 | 0.935 |
| | RMSE | training | 0.539 | 0.537 | 0.539 | 0.207 |
| | | test | 0.569 | 0.568 | 0.562 | 0.451 |

contained within the local applicability domain for the algorithm and a series of nearest neighbors based on the descriptors. The data generated from the predictions reported in this manuscript are available via the CompTox dashboard and are listed as NICEATM Models.

## RESULTS AND DISCUSSION

Prior to regression modeling, we investigated the correlation among the six properties as well as their relationship with molecular weight (MW) to potentially inform additional feature sets for model building. For example, if the BP prediction is very accurate and has high correlation to MP, one could use the BP as a predictor for MP modeling. Table 2 gives Pearson correlation coefficients ($r$) for each combination of physicochemical properties, which were calculated using the following formula:[62]

$$r = \frac{n\sum p_k p_l - \sum p_k \sum p_l}{\sqrt{n\sum p_k^2 - (\sum p_k)^2}\sqrt{n\sum p_l^2 - (\sum p_l)^2}} \quad (8)$$

In eq 8, $p_k$ and $p_l$ represent different physicochemical properties and $n$ is the number of chemicals in each pair of properties. As shown in Table 2, MW is moderately correlated to logVP ($r$ = −0.721) and logS ($r$ = −0.648), poorly correlated to BP ($r$ = 0.475), MP ($r$ = 0.460) and logBCF ($r$ = 0.367), and nearly uncorrelated with logP ($r$ = 0.256). According to the correlation results reported in Table 2, the six properties can be divided into two groups, one including log P, logS, and logBCF with $r$ of −0.873, 0.830, and −0.825 for logP vs logS, logP vs logBCF, and logS vs logBCF, respectively; another including BP, MP and logVP, where MP is significantly correlated to BP and logVP with $r$ of 0.733 and −0.833, respectively. BP is also highly correlated to logVP with $r$ of −0.959. In contrast, the experimental data of logP, logS and logBCF are uncorrelated with those of MP, BP and logVP, with low $r$ for each pair.

Our data set included a large number of independent variables (molecular fingerprint bits) but not all of them made substantial contributions to modeling the physicochemical properties. The presence of irrelevant or redundant features limits the applicability of a model and may result in overfitting. When too many variables are employed, a complicated regression model can fit the training data extremely well with very low deviations between experimental values and predicted values. However, an overadapted or overfitted model yields large prediction errors for test chemicals and thus loses its generalization. Therefore, it is critical to identify appropriate subsets of informative variables from the original set of fingerprint bits. We applied the well-established GA dimension reduction method to our data sets for feature selection. Prediction models were built by regression against the training sets with different subsets of fingerprint bits using four contrasting approaches, and the model performance was evaluated using 10-fold cross-validation and an independent external test set. The results of each predicted property and regression statistics are discussed in detail below.

**Octanol−Water Partition Coefficient (logP) Model.** Among the six properties, logP had the largest data set of chemicals (over 14000) and fingerprint bits (1681). As shown in Table 3 and Figure 2A, model performance varied with respect to the number of fingerprint bits used. When models were trained using the entire set of fingerprint bits, some of the variables were unrelated to the variation of the property. Using 600 fingerprint bits selected by GA, the MLR results show a significant correlation between the estimated and measured values on the test set with an $R^2$ of 0.888 and a minimum


DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

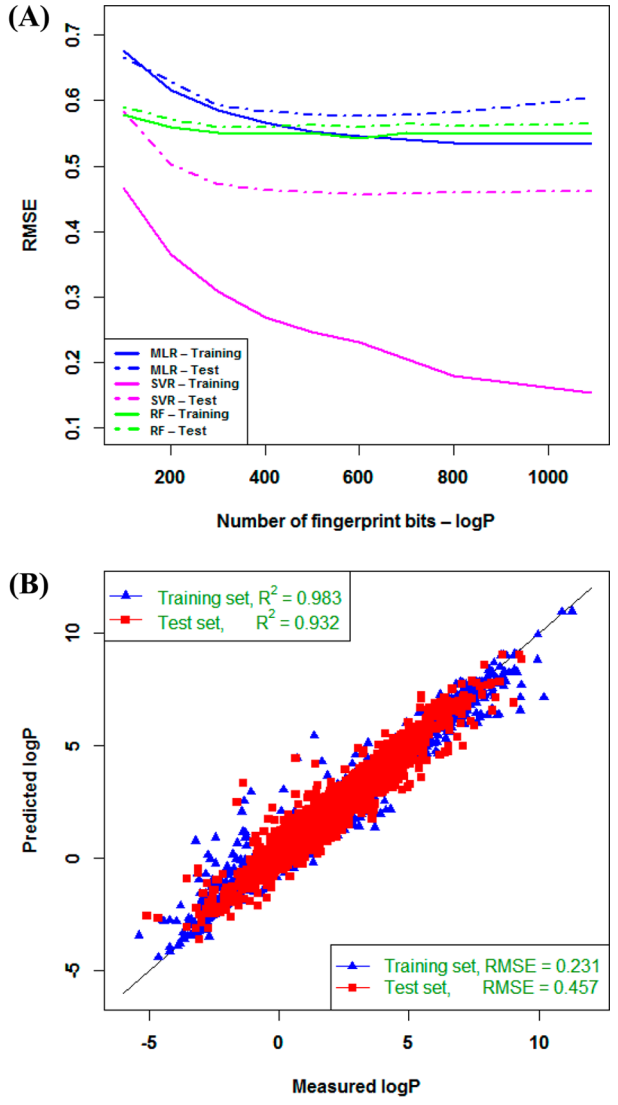



**Figure 2.** Relationship between model complexity and prediction errors (RMSE) (A) and plot of experimental data versus estimated values by SVR using 600 fingerprint bits (B) for logP.

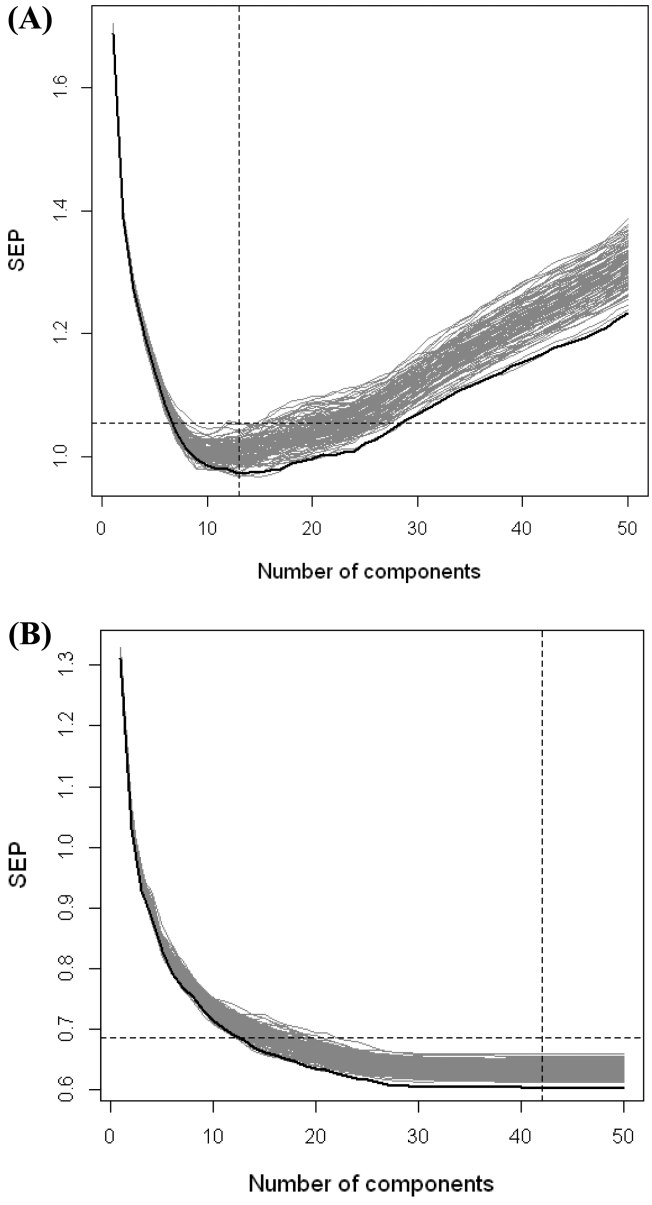



**Figure 3.** Relationship between the number of principal components (PCs) and the standard error of prediction (SEP) for the PLSR model of logP. The black lines were produced from a single 10-fold CV while the gray lines correspond to 100 repetitions of the 10-fold CV. (A) Plot of SEP versus PCs for the all-bit model. (B) Plot of SEP versus PCs for the 600-bit model selected by GA.

RMSE of 0.576. These statistics are very similar to those for the training set ($R^2$ of 0.901 and RMSE of 0.546), indicating the stability of the model. The inclusion of MW marginally improved the prediction on the test set with $R^2$ increased to 0.891.

When PLSR was utilized to build models, the number of significant principal components (PCs) was determined using a 10-fold cross-validation (CV) procedure on the training set. The relationship of the standard error of prediction (SEP) versus the number of PCs is displayed in Figure 3. The gray lines were produced by repeating this procedure 100 times, while the black line depicts the lowest SEP value from a single 10-fold CV; the dashed vertical lines represent the optimal number of PCs and the dashed horizontal lines indicate the SEP value for the test set when the optimal PCs are applied. The variation of SEP is much larger for the all-variable (1681 fingerprint bits) model than the model of 600 fingerprint bits selected by GA, implying that the optimized feature set exhibits greater model stability. For the all-variable model, SEP initially decreases with PCs; the trend then reverses when noise emerges as the complexity of the model increases. Although more components improve the fitting quality of the model, they also lower the predictive power due to overfitting; the minimal cross-validation error was obtained using 13 PCs. For the 600-bit model, the SEP decreases monotonically and gradually approaches a stable value, and the use of 42 PCs yielded the optimal model, which gave a minimum RMSE of 0.575 corresponding to an $R^2$ of 0.889 for the test chemicals (Table 3).

Unlike other modeling approaches, RF regression was insensitive to the number of fingerprint bits. Figure 2A shows that the model statistics do not vary with the number of variables. An RF model trained on the 600 fingerprint bits produced nearly the same prediction as that trained on all bits. In each case the correlation coefficient of the test set was found to be similar to that of the training set, and the RMSE values


DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

**Table 4. Regression Statistics of LogS Using Subsets of Fingerprint Bits, MW, and LogP**

| variable | model statistics | data set | MLR | PLSR | RF | SVR |
|---|---|---|---|---|---|---|
| 1061 FP bits | $R^2$ | training | 0.983 | 0.933 | 0.878 | 0.995 |
| | | test | 0.604 | 0.869 | 0.896 | 0.910 |
| | RMSE | training | 0.275 | 0.540 | 0.673 | 0.156 |
| | | test | 1.805 | 0.791 | 0.663 | 0.649 |
| 350 FP bits | $R^2$ | training | 0.953 | 0.952 | 0.873 | 0.959 |
| | | test | 0.931 | 0.931 | 0.892 | 0.932 |
| | RMSE | training | 0.454 | 0.456 | 0.687 | 0.422 |
| | | test | 0.588 | 0.587 | 0.671 | 0.579 |
| 350 FP bits + MW | $R^2$ | training | 0.957 | 0.956 | 0.887 | 0.960 |
| | | test | 0.932 | 0.933 | 0.899 | 0.933 |
| | RMSE | training | 0.436 | 0.437 | 0.648 | 0.414 |
| | | test | 0.584 | 0.580 | 0.655 | 0.575 |
| 350 FP bits + MW + logP | $R^2$ | training | 0.961 | 0.960 | 0.925 | 0.966 |
| | | test | 0.935 | 0.938 | 0.932 | 0.939 |
| | RMSE | training | 0.415 | 0.416 | 0.547 | 0.388 |
| | | test | 0.552 | 0.548 | 0.554 | 0.542 |

were also very close to each other. Since RF is not sensitive to the number of variables, feature selection did not improve the predictive performance, and hence it was unnecessary to remove irrelevant fingerprint bits from the model. The RF algorithm encompassed a large number of simple tree models, and thus greatly mitigated overfitting. Additionally, it was not essential to perform cross or independent validation on RF models as this was inherently provided by the OOB estimate.

The optimum RBF kernel parameter $\gamma$, the radius of the tube $\varepsilon$, and the regularization parameter $C$ that collectively yielded the lowest RMSE from cross-validation were chosen to build the SVR model. SVR is robust against the existence of irrelevant or of mutually correlated variables, and here even using all fingerprint bits yielded satisfactory results. Feature selection did further improve the predictive performance with test set $R^2$ increasing from 0.920 to 0.932 using 600 fingerprint bits. Figure 2B shows a scatter plot of predicted versus experimental values from SVR modeling. In general, the measured and predicted values were highly correlated and the majority of the data points are concentrated around the regression line with a small deviation over a large range for both training and test sets.

**Water Solubility (logS) Model.** Compared to the logP data set, the logS data has a low ratio of training chemicals to bits (1507/1061). As shown in Table 4, small sets of variables achieved better predictive power than large ones. When all fingerprint bits were employed, the model yielded $R^2$ of 0.983, 0.933, and 0.995 for the training set, but only 0.604, 0.869, and 0.910 for the test set corresponding to MLR, PLSR, and SVR, respectively, implying that overfitting occurred. The predictive performance of the models was improved remarkably when a more appropriate number of fingerprint bits was selected using GA (Figure 4A). The optimal model with 350 bits produced test $R^2$ of 0.931 for both MLR and PLSR, and 0.932 for SVR, which were very close to the training set values (Figure 4B). These results confirm the effectiveness of the GA feature selection in capturing the relevant information and building stable models. Using an optimal combination of fingerprint bits selected using GA, the results from MLR and PLSR are almost identical.

Many factors can affect the solubility of a molecule in an aqueous solution, including the molecule's size, shape, polarity, and hydrophobicity. For example, a cavity must be formed in water for a chemical to dissolve in an aqueous solution. The

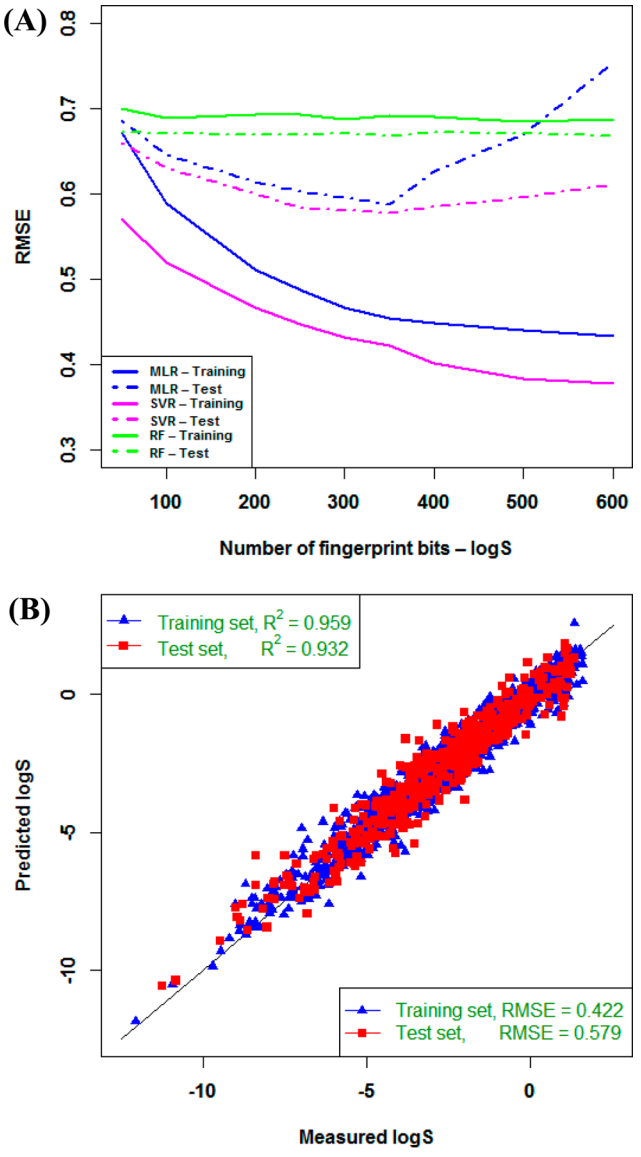



**Figure 4.** Relationship between model complexity and prediction errors (RMSE) (A) and plot of experimental data versus estimated values by SVR using 350 fingerprint bits (B) for logS.



DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

larger the molecular weight or the molecular size, the larger the required cavity is, so the greater the energy needed to make a bigger cavity and thus the lower the solubility. Hence, logS is negatively correlated to MW with $r = -0.648$. The partition coefficient measures how strongly a chemical molecule interacts with *n*-octanol in comparison to water. The aqueous solubility reduces with increasing hydrophobicity, and hence logS has a negative correlation with logP ($r = -0.873$). When MW and logP were incorporated as additional descriptors to model water solubility, both were found to play important roles and led to improved predictions compared to the pure fingerprint bit model. A significant improvement was achieved for all four regression approaches, and SVR yielded the highest test $R^2$ of 0.939 and lowest RMSE of 0.542. This improvement was most pronounced for RF with $R^2$ increasing from 0.892 to 0.932, and it performed nearly as well as the other models, suggesting that the use of numeric descriptors favors RF modeling.

**Bioconcentration Factor (logBCF) Model.** As the smallest data set with only 456 training chemicals, the validation statistics of logBCF models rely highly on the ratio of the number of chemicals to the fingerprint bits. When all 450 bits were employed for MLR modeling, for example, the fit to the training data was excellent with $R^2$ close to 1, but for the test chemicals $R^2$ was close to zero (Table S2). This was expected, as overly complex models have a tendency to overfit the training data and lose their generalization capability. Unlike the all-variable model from MLR, which yielded extremely low prediction accuracy for the test set, PLSR, RF, and SVR models achieved acceptable regression statistics with test $R^2$ between 0.723 and 0.725.

As with the previous results, the quality of the model depended heavily on the number of selected fingerprint bits. Figure S2A shows the initial decline of the RMSE for the test set until attaining a minimum at a medium number of bits, and then a gradual increase with the number of bits. While using too many variables resulted in overfitting and hence lowered the predictive ability of the model, too few variables cannot capture sufficient structural information, and with a smaller set of chemicals it was more challenging to derive a stable and meaningful model due to underfitting. When the most information-rich variables were retained and redundant ones were discarded via feature selection, the predictive performance was enhanced significantly. As shown in Figure S2A, the models with 100−250 fingerprint bits achieved very similar performance and the lowest prediction errors occurred on the models with moderate complexity around 200 bits. In the development of QSPR models, it is recommended that the ratio of number of training chemicals to number of descriptors be greater than 4:1.[61] In the logBCF modeling, this rule is broken due to the use of binary fingerprint bits as descriptors and the models are not overfitted even if excessive number of descriptors are employed.

Using optimal subset of fingerprint bits, MLR, PLSR, and SVR yielded very similar results in terms of the predictive ability on the test set with SVR achieving the highest $R^2$ of 0.879 and lowest RMSE of 0.418 (Figure S2B). Including MW and logP as additional descriptors to model logBCF did not improve model performance to a significant extent except RF modeling, which yielded an increase in test $R^2$ from 0.734 to 0.816. RF regression is not sensitive to the number of descriptors, but it is very sensitive to if the descriptors are discrete or continuous. Continuous values, such as MW and logP, can remarkably improve the performance of RF models.

**Boiling Point (BP) Model.** Since the ratio of training chemicals to bits (4074/1050) was large for the BP data set, the modeling statistics were not sensitive to the bit number, and the model performance did not vary considerably with different subsets of fingerprint bits for the test set (Figure S3A). The BP model with 400 bits had the highest test $R^2$ of 0.943 and lowest RMSE of 19.72 using SVR (Table S3). Our regression analysis suggested that MW is an important contributing factor for boiling point. When incorporating MW into the model, the predictive performance was improved with test $R^2$ increasing to 0.965 for SVR model, which outperformed the other three approaches with test $R^2$ ranging from 0.935 to 0.940. The model had difficulty accurately predicting high boiling points; large errors were observed near the upper extreme of the experimental range, and chemicals with medium and low experimental values were predicted more accurately (Figure S3B). These errors may result from experimental measurements or are due to fewer data points at the high end of the range of values and resulting lack of model coverage.

**Melting Point (MP) Model.** When modeling MP, feature selection did not have a significant impact on the external validation results (Figure S4A) due to a large ratio of training chemicals to bits (6485/1424). Similar to the other physicochemical property modeling, SVR achieved the best results with a test $R^2$ of 0.813 (Figure S4B), followed by RF (0.802), PLSR (0.781), and MLR (0.780) (Table S4) using 500 fingerprint bits. To improve the predictive ability, we considered employing estimated BP as an additional descriptor to predict MP. This consideration was based on the fact that the BP data set was large (5400 chemicals) and good predictive performance with test $R^2$ greater than 0.960 was achieved from the fingerprint model, leading to reliable BP estimates.

A series of models were derived through correlating the experimental MP with MW and estimated BP. Table S4 compares the regression statistics obtained using various combinations of fingerprint bits, MW and BP. It is evident that the regression models were very sensitive to BP, which could provide useful information and more accurate estimation for MP. The inclusion of both BP and MW as two descriptors into MP models enhanced the predictive performance for all four regression approaches. The test $R^2$ increased to 0.826 while RMSE decreased to 39.14 using SVR. Overall, SVR regression substantially outperformed the other approaches and RF was slightly superior to MLR and PLSR. These facts reflect the advantage of nonlinear approaches over linear approaches for modeling MP. The RMSE reported here compares well with a recent report that modeled on a much larger set of over 200 000 data points extracted from patents.[67]

**Vapor Pressure (logVP) Model.** The data set of logVP was relatively small with 2034 training chemicals. Hence, feature selection by GA remarkably improved the predictive ability of the model when fewer fingerprint bits were used instead of all 1145 bits, particularly for MLR modeling (Figure S5A). The four approaches had low prediction errors and comparatively similar statistical performance when 350 bits were employed to build the models, with $R^2$ values of the test set ranging from 0.902 to 0.930 (Table S5). A significant correlation between logVP and BP was observed with an $r$ of −0.959, and meanwhile MW was also correlated to logVP with an $r$ of −0.721 (Table 2). After introducing MW as an additional descriptor, the correlation of the model was greatly improved, and the inclusion of BP further enhanced the predictive performance with test $R^2$ values between 0.941 and


DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

Table 5. Correlation Statistics between Experimental Data for Chemicals Used in This Study and EPI Suite Predictions

| | all chemicals | | | training set | | | test set | | |
|---|---|---|---|---|---|---|---|---|---|
| property | number | $R^2$ | RMSE | number | $R^2$ | RMSE | number | $R^2$ | RMSE |
| logP | 14207 | 0.895 | 0.605 | 11370 | 0.893 | 0.612 | 2837 | 0.904 | 0.576 |
| logS | 2010 | 0.948 | 0.490 | 1507 | 0.945 | 0.495 | 503 | 0.955 | 0.472 |
| logBCF | 608 | 0.818 | 0.484 | 456 | 0.820 | 0.485 | 152 | 0.813 | 0.481 |
| BP | 5432 | 0.937 | 21.73 | 4074 | 0.938 | 21.56 | 1358 | 0.937 | 21.78 |
| MP | 8648 | 0.635 | 57.25 | 6485 | 0.634 | 57.58 | 2163 | 0.638 | 57.14 |
| logVP | 2713 | 0.917 | 0.965 | 2034 | 0.923 | 0.928 | 679 | 0.900 | 1.071 |

0.946. Linear methods MLR and PLSR performed nearly as well as nonlinear methods RF and SVR. MLR is a simple regression method that does not require time-consuming parameter optimization in order to achieve good performance. If this constraint were a concern, the MLR model would be particularly suitable for the prediction of logVP.

Although these models yielded accurate predictions for most chemicals, some predictions deviated considerably from the experimental values. As indicated in Figure S5B, the model performed poorly for chemicals with low vapor pressure (logVP < −5.0 log units). It is difficult to accurately measure vapor pressure for chemicals with very low volatility and experimental errors tend to be larger. The lack of high quality experimental data may be a driving factor for the failure of the models at extreme values.

**Comparison to the Prediction from EPI Suite.** Many software programs are available for predicting physicochemical properties.[44] Among them, the EPI Suite,[45] developed by Syracuse Research Corporation (SRC), is a standalone, reproducible, EPA-endorsed, and branded product with a long history of usage. It is widely used by governmental regulatory agencies in the United States, Canada, and Europe to predict physicochemical properties of environmental chemicals. The EPI Suite models were developed using fragment descriptors and multiple linear regression method.[41] To compare our models with EPI Suite predictions, the KOWWIN (version 1.67), WATERNT (version 1.01), BCFBAF (version 1.20), and MPBPNT (version 1.43) modules were employed to estimate logP, logS, and logBCF, as well as MP, BP, and logVP, respectively. For the six property sets, not all the information about the training and test chemicals is available from EPI Suite original data. We broke the entire data set into training and test sets as described in Table 1. As shown in Table 5, the entire, training and test sets have very similar correlation statistics between experimental data and estimated values for all six properties. When compared with the test sets using the coefficient of determination $R^2$, our logS model is inferior to EPI Suite's (0.939 vs 0.955), and our other models exhibit better predictions for logP (0.935 vs 0.904), BP (0.965 vs 0.937), logVP (0.946 vs 0.900), and logBCF (0.885 vs 0.813). Our estimation for melting point is substantially superior to EPI Suite's, with an $R^2$ of 0.826 compared to 0.638 from EPI Suite. It should be noted that some test chemicals in our models served as training chemicals in EPI Suite, producing potentially artificially inflated correlations for EPI Suite. Overall, our models are simpler and more effective with only binary fingerprint bits as descriptors, and therefore represent an improvement in property prediction.

*Analysis of Applicability Domain.* The QSPR models were developed using training sets and thus, their applicability to external chemicals depends on the structural similarity between the external test chemicals and the training chemicals. The models would be expected to provide more reliable predictions for chemicals that fall in the AD, as defined earlier by the three distance measures (leverage, kNN, and distance to centroid). In this study, a test chemical is deemed to be completely outside the AD only if the thresholds from all three distance measures are exceeded. If only one or two thresholds are exceeded, the chemical is considered to be potentially outside the AD. Table 6 and Table S6 summarize the number of test chemicals outside the AD for each regression model identified by individual distance measures and their combinations. These tables also show $R^2$ and RMSE for the test sets after removing the chemicals considered outside the AD. In all cases, the number of test chemicals completely outside the AD was very small (0 to 9), and their removal did not have a significant influence on the model statistics. Using kNN and distance from centroid measures, the model statistics were improved remarkably with kNN exhibiting better performance than the distance from centroid. Taking the logP model as an example, $R^2$ increased from 0.935 to 0.940 while RMSE decreased from 0.451 to 0.434 after removing 121 test chemicals which were identified outside the domain by five nearest neighbors. kNN is a more appropriate measure of distance between training and test chemicals because the descriptors employed in this study are binary variables. Leverage measure did not considerably impact the model statistics. This is in part because the number of chemicals outside the AD identified by leverage was small. It is also important to consider that not all chemicals outside the

Table 6. Applicability Domain (AD) of LogP and LogS Models: Test Set Evaluation[a]

| | | | | experimental vs predicted test chemicals inside AD | |
|---|---|---|---|---|---|
| property | measure | chemicals outside AD | chemicals inside AD | $R^2$ | RMSE |
| logP | leverage (I) | 10 | 2827 | 0.935 | 0.448 |
| | distance from centroid (II) | 136 | 2701 | 0.936 | 0.443 |
| | distance by kNN (III) | 121 | 2716 | 0.940 | 0.434 |
| | I and II and III | 3 | 2834 | 0.935 | 0.450 |
| | I or II or III | 247 | 2590 | 0.938 | 0.439 |
| logS | leverage (I) | 7 | 496 | 0.939 | 0.541 |
| | distance from centroid (II) | 28 | 475 | 0.941 | 0.538 |
| | distance by kNN (III) | 22 | 481 | 0.943 | 0.531 |
| | I and II and III | 0 | 503 | 0.939 | 0.542 |
| | I or II or III | 43 | 460 | 0.941 | 0.534 |

[a]The models were built by SVR using 600 FP bits + MW for logP and 350 FP bits + MW + logP for logS.



DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

AD are always wrongly predicted and not all chemicals inside the AD are correctly predicted, which is in agreement with the literature.[63] This phenomenon can be observed in the plots of leverage versus standardized residuals (Figure 5 and Figure S6),

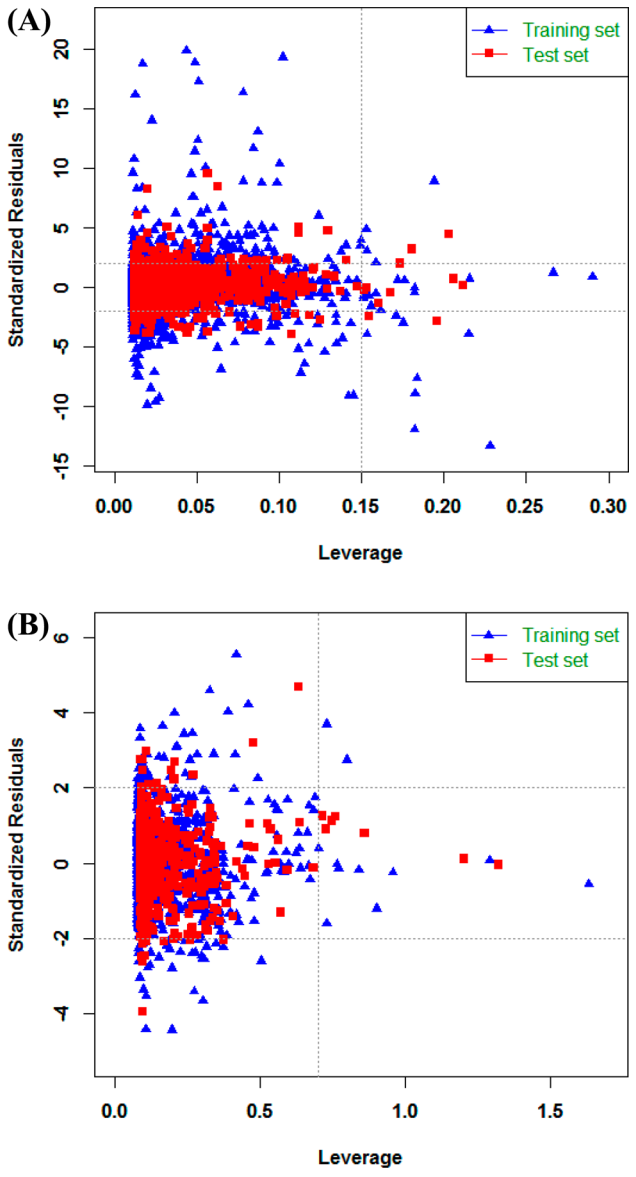



**Figure 5.** Plots of leverage versus standardized residuals for logP (A) and logS (B) models' training and test sets. The models were built by SVR using 350 and 600 fingerprint bits for logS and logP, respectively. Vertical dashed line marks AD threshold based on the leverage value. Horizon dashed lines define a region where predictions were within two standardized residuals.

which show many test chemicals outside the AD were correctly predicted and some of the test chemicals inside the AD were not accurately predicted. Tables S7−S12 list experimental and predicted values as well as molecular structures for top ten chemicals with largest residuals for the six properties.

*Adherence to OECD (Q)SAR Validation Principles.* OECD has defined five validation principles to facilitate the consideration of a (Q)SAR model for regulatory purposes. To encourage the acceptance and application of the models presented here, we have provided the following information in conjunction with the OECD validation principles.

(1) A defined end point. We have built models for six well-defined physicochemical properties (logP, logS, logBCF, BP, MP, and logVP).
(2) An unambiguous algorithm. We have evaluated four unambiguous machine learning algorithms (MLR, PLSR, RF, and SVR). To increase transparency, the R code associated with these modeling approaches is made available in the Supporting Information.
(3) A defined domain of applicability. We have evaluated the AD using three different distance measures.
(4) Appropriate measure of goodness-of-fit, robustness, and predictivity. We have evaluated our models based on cross-validation and external validation sets, iteratively varied the feature sets to examine robustness, and compared them to currently available gold standard physicochemical property prediction software.
(5) A mechanistic interpretation, if possible. Here it is not practical to make an interpretation linking each and every selected fingerprint bit to the modeled end points. However, we assume that the statistically selected fingerprint bits represent fragments that are relevant to the studied end points. The physicochemical properties being predicted by the QSPR models presented here are critical inputs to environmental fate and transport and toxicity prediction models.

Accurate prediction of physicochemical properties using validated QSPR models will ensure a much clearer understanding of chemical hazard and exposure potential. The QSAR Model Reporting Format (QMRF) is available in the Supporting Information.

## CONCLUSIONS

In the present study, QSPR models were implemented to predict six physicochemical properties from binary molecular fingerprints on the basis of large and structurally diverse sets of environmental chemicals using four distinctly different modeling approaches. Satisfactory predictive performance was achieved using optimal subsets of fingerprint bits and optimized regression parameters, and the estimated values correlated very well with experimental values. Both linear and nonlinear approaches provided accurate property predictions with similar regression statistics.

Since not all fingerprint bits contain useful or unique information related to the physicochemical properties, it was crucial to select the most relevant variables to construct the model. On one hand, an insufficient number of fingerprint bits led to underfitted and statistically unstable models and potential lack of coverage in new chemistry space. On the other hand, use of excessive numbers of fingerprint bits produced overfitted models from the training data, resulting in poor predictions for the test sets. Feature selection using GA was found to substantially improve the predictive ability of MLR models, influence PLSR and SVR models only slightly, and exert no effect on RF models.

The four regression approaches exhibited different modeling characteristics. MLR is a simple linear regression method and does not require a time-consuming parameter optimization procedure. PLSR yields relatively reasonable accuracy of predictions and performs well even in the presence of noise variables. RF is robust against overfitting since the algorithm constructed an ensemble of regression trees to model the data, and we found it predicted the training and test sets equally well.


DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

Table 7. Regression Statistics of Best Performing SVR Models for Each Property

| | | 10-fold cross-validation | | | | | |
|---|---|---|---|---|---|---|---|
| | | $Q^2$ | | RMSEcv | | | |
| property | variables | mean | interval | mean | interval | $R^2$ for test set | RMSE for test set |
| logP | 600 bits + MW | 0.932 | 0.926−0.935 | 0.478 | 0.428−0.558 | 0.935 | 0.451 |
| logS | 350 bits + MW + logP | 0.928 | 0.921−0.936 | 0.580 | 0.487−0.666 | 0.939 | 0.542 |
| logBCF | 200 bits + MW + logP | 0.863 | 0.851−0.877 | 0.465 | 0.361−0.525 | 0.885 | 0.444 |
| BP | 400 bits + MW | 0.955 | 0.941−0.968 | 17.98 | 14.30−20.64 | 0.965 | 15.63 |
| MP | 500 bits + MW + BP | 0.824 | 0.812−0.831 | 41.41 | 38.37−45.29 | 0.826 | 39.14 |
| logVP | 350 bits + MW + BP | 0.929 | 0.915−0.939 | 0.975 | 0.806−1.080 | 0.946 | 0.810 |

Nevertheless, the RF model did not perform as well as the other models when binary fingerprints were used as descriptors.

SVR, a complex and nonlinear modeling technique, was shown to be superior to the other three approaches for modeling these properties. SVR coupled with GA in selecting the most significant descriptors achieved excellent results and predicted all the six properties accurately in terms of the coefficient of determination and RMSE for both 10-fold cross-validation and the external test sets. Table 7 summarizes the best performing models and shows that the statistics between cross-validation and test sets are close to each other, confirming the stability, robustness, and reliability of these regression models. The correlation between experimental and predicted values for the test sets was found to vary over a range of $R^2$ for different properties. The BP model had the highest test $R^2$ of 0.965, followed by logVP ($R^2$ = 0.946), logS ($R^2$ = 0.939), and logP ($R^2$ = 0.935). MP and logBCF model predictions were less accurate, with test $R^2$ = 0.826 and 0.885, respectively.

There are numerous programs available for the prediction of physicochemical properties, where the QSPR models were built based on training sets from a few to several thousands of chemicals with a broad structural diversity and coverage of chemical space as well as a variety of specific chemical classes.[44] These models have considerable variations in performance. Taking logP as an example, the RMSE values range from 0.41 to 1.98 log unit whereas our model achieved an RMSE of 0.45 log unit. When compared to EPI Suite predictions, our methods demonstrate better accuracy for a wider range of chemicals of interest.

Taken together, the results of this study demonstrate that the combination of careful data curation, binary molecular fingerprints, and machine learning approaches provides a rapid and efficient way to estimate physicochemical properties of environmental chemicals. We have thus developed QSPR models that are highly stable and reliable. The models conform with most of the validation principles put forth by the OECD and therefore have broad applicability for property estimation of many classes of compounds.

The models are freely available via github (https://github.com/zang1123/Physicochemical-Property-Prediction/) and can be used by researchers and regulators to make predictions on new chemical sets, improve toxicity models, and inform hazard/risk characterization. In the near future, the models developed here will be used to predict data for the six reported properties for the CompTox dashboard web site content of >720 000 chemicals and made available to the public. These predictions will be identifiable with NICEATM Models as the source.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00625.

QSAR Model Reporting Formats. Examples of R code: feature selection and regression analysis. Figure S1: Data distribution of logBCF, BP, MP and logVP. Figures S2−S5: Relationship between model complexity and prediction errors as well as the plots of estimated values versus experimental data for logBCF, BP, MP, and logVP, respectively. Figure S6: Plots of leverage versus standardized residuals for logBCF, BP, MP, and logVP models. Table S1: Chemical product classes for training and test sets. Tables S2−S5: Regression statistics for logBCF, BP, MP, and logVP, respectively. Table S6: Applicability domains for logBCF, BP, MP, and logVP. Tables S7−S12: Chemicals with large prediction residuals for the six properties (PDF)

Chemical names, CAS registry number and SMILES as well as experimentally measured and estimated property values of the training and test sets (XLSX)

## ■ AUTHOR INFORMATION


### Corresponding Author

*Mailing address: 530 Davis Drive, Morrisville, NC 27560. E-mail: nicole.kleinstreuer@nih.gov.

### ORCID

Qingda Zang: 0000-0003-1543-8307

Nicole C. Kleinstreuer: 0000-0002-7914-3682


### Notes

The views expressed in this article are those of the authors and do not necessarily reflect the views of policies of the U.S. Environmental Protection Agency and National Institute of Environmental Health Sciences. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We would like to express our deep appreciation to Dr. Ann Richard (EPA National Center for Computational Toxicology) and Dr. Shannon Bell (Integrated Laboratory Systems, Inc.) for their constructive suggestions for this manuscript. We also thank Ms. Catherine Sprankle (Integrated Laboratory Systems, Inc.) for editorial review. This project was funded in whole or in part with federal funds from the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH) under contract HHSN273201500010C to Integrated Laboratory Systems in support of NICEATM.



DOI: 10.1021/acs.jcim.6b00625
J. Chem. Inf. Model. 2017, 57, 36−49

## ABBREVIATIONS

AD, applicability domain; BCF, bioconcentration factor; BP, boiling point; CASRN, chemical abstracts service registry number; CV, cross-validation; EPA, the U.S. Environmental Protection Agency; EPI, Estimation Program Interface; GA, genetic algorithm; HTS, high-throughput screening; ICATM, International Cooperation on Alternative Test Methods; kNN, *k*-nearest neighbors; logP, octanol−water partition coefficient; logS, water solubility; MLR, multiple linear regression; MP, melting point; MW, molecular weight; NICEATM, NTP Interagency Center for the Evaluation of Alternative Toxicological Methods; NTP, National Toxicology Program; OECD, Organisation of Economic Cooperation and Development; OOB, out-of-bag; OPERA, open sar application; PC, principal component; PLSR, partial least-squares regression; QMRF, QSAR Model Reporting Format; QSAR, quantitative structure−activity relationship; QSPR, quantitative structure−property relationship; $R^2$, coefficient determination; RBF, radial basis function; REACH, Registration, Evaluation, Authorization and Restriction of Chemicals; RF, random forest; RMSE, root mean squared error; SEP, standard error of prediction; SMARTS, smiles arbitrary target specification; SMILES, Simplified Molecular Identification and Line Entry System; SRC, Syracuse Research Corporation; SVR, support vector regression; VP, vapor pressure

## REFERENCES


(1) U.S. EPA, Office of Pollution Prevention and Toxics (OPPT) Chemical Reviews and Tools Case Study. http://www.who.int/ifcs/documents/forums/forum5/precaution/epa_en.pdf (accessed December 16, 2016).

(2) Chemicals under the Toxic Substances Control Act (TSCA). https://www.epa.gov/chemicals-under-tsca (accessed December 16, 2016).

(3) Egeghy, P. P.; Judson, R. S.; Gangwal, S.; Mosher, S.; Smith, D.; Vail, J.; Cohen-Hubal, E. A. The exposure Data Landscape for Manufactured Chemicals. *Sci. Total Environ.* **2012**, *414* (1), 159−166.

(4) Judson, R. S.; Martin, M. T.; Egeghy, P. P.; Gangwal, S.; Reif, D. M.; Kothiya, P.; Wolf, M. A.; Cathey, T.; Transue, T. R.; Smith, D.; et al. Aggregating Data for Computational Toxicology Applications: The U.S. Environmental Protection Agency (EPA) Aggregated Computational Toxicology Resource (ACToR) System. *Int. J. Mol. Sci.* **2012**, *13* (2), 1805−1831.

(5) Judson, R. S.; Richard, A. M.; Dix, D. J.; Houck, K. A.; Elloumi, F.; Martin, M. T.; Cathey, T.; Transue, T. R.; Spencer, R.; Wolf, M. A. ACToR − Aggregated Computational Toxicology Resource. *Toxicol. Appl. Pharmacol.* **2008**, *233* (1), 7−13.

(6) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124* (7), 1023−1033.

(7) Cohen-Hubal, E. A.; Richard, A. M.; Aylward, L.; Edwards, S. W.; Gallagher, J.; Goldsmith, J. M.; Isukapalli, S.; Tornero-Velez, R.; Weber, E. J.; Kavlock, R. J. Advancing Exposure Characterization for Chemical Evaluation and Risk Assessment. *J. Toxicol. Environ. Health, Part B* **2010**, *13* (2−4), 299−313.

(8) Knudsen, T. B.; Houck, K. A.; Sipes, N.; Singh, A. V.; Judson, R. S.; Martin, M. T.; Weissman, A.; Kleinstreuer, N.; Mortensen, H. M.; Reif, D. M.; et al. Activity Profiles of 309 ToxCast Chemicals Evaluated across 292 Biochemical Targets. *Toxicology* **2011**, *282* (1−2), 1−15.

(9) Judson, R. S.; Richard, A. M.; Dix, D. J.; Houck, K. A.; Martin, M. T.; Kavlock, R. J.; Dellarco, V.; Henry, T.; Holderman, T.; Sayre, P.; et al. The Toxicity Data Landscape for Environmental Chemicals. *Environ. Health Perspect.* **2009**, *117* (5), 685−695.

(10) Kavlock, R. J.; Dix, D. J. Computational Toxicology as Implemented by the U.S. EPA: Providing High Throughput Decision Support Tools for Screening and Assessing Chemical Exposure, Hazard and Risk. *J. Toxicol. Environ. Health, Part B* **2010**, *13* (2−4), 197−217.

(11) Wetmore, B. A.; Wambaugh, J. F.; Ferguson, S. S.; Sochaski, M. A.; Rotroff, D. M.; Freeman, K.; Clewell, H. J., III; Dix, D. J.; Andersen, M. E.; Houck, K. A.; et al. Integration of Dosimetry, Exposure and High-Throughput Screening Data in Chemical Toxicity Assessment. *Toxicol. Sci.* **2012**, *125* (1), 157−174.

(12) Judson, R. S.; Kavlock, R. J.; Setzer, R. W.; Cohen-Hubal, E. A.; Martin, M. T.; Knudsen, T. B.; Houck, K. A.; Thomas, R. S.; Wetmore, B. A.; Dix, D. J. Estimating Toxicity-Related Biological Pathway Altering Doses for High-Throughput Chemical Risk Assessment. *Chem. Res. Toxicol.* **2011**, *24* (4), 451−462.

(13) Martin, M. T.; Dix, D. J.; Judson, R. S.; Kavlock, R. J.; Reif, D. M.; Richard, A. M.; Rotroff, D. M.; Romanov, S.; Medvedev, A.; Poltoratskaya, N.; et al. Impact of Environmental Chemicals on Key Transcription Regulators and Correlation to Toxicity End Points within EPA's ToxCast Program. *Chem. Res. Toxicol.* **2010**, *23* (3), 578−590.

(14) Judson, R. S.; Houck, K. A.; Kavlock, R. J.; Knudsen, T. B.; Martin, M. T.; Mortensen, H. M.; Reif, D. M.; Rotroff, D. M.; Shah, I. A.; Richard, A. M.; et al. In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: the ToxCast Project. *Environ. Health Perspect.* **2010**, *118* (4), 485−492.

(15) Dix, D. J.; Houck, K. A.; Martin, M. T.; Richard, A. M.; Setzer, R. W.; Kavlock, R. J. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. *Toxicol. Sci.* **2007**, *95* (1), 5−12.

(16) Sipes, N. S.; Martin, M. T.; Kothiya, P.; Reif, D. M.; Judson, R. S.; Richard, A. M.; Houck, K. A.; Dix, D. J.; Kavlock, R. J.; Knudsen, T. B. Profiling 976 ToxCast Chemicals across 331 Enzymatic and Receptor Signaling Assays. *Chem. Res. Toxicol.* **2013**, *26* (6), 878−895.

(17) Browne, P.; Judson, R. S.; Casey, W. M.; Kleinstreuer, N. C.; Thomas, R. S. Screening Chemicals for Estrogen Receptor Bioactivity Using a Computational Model. *Environ. Sci. Technol.* **2015**, *49* (14), 8804−8814.

(18) Kleinstreuer, N. C.; Yang, J.; Berg, E. L.; Knudsen, T. B.; Richard, A. M.; Martin, M. T.; Reif, D. M.; Judson, R. S.; Polokoff, M.; Dix, D. J.; et al. Phenotypic Screening of the ToxCast Chemical Library to Classify Toxic and Therapeutic Mechanisms. *Nat. Biotechnol.* **2014**, *32* (6), 583−591.

(19) Hermens, J. L.; de Bruijn, J. H.; Brooke, D. N. The Octanol-Water Partition Coefficient: Strengths and Limitations. *Environ. Toxicol. Chem.* **2013**, *32* (4), 732−733.

(20) Wang, J.; Hou, T. Recent Advances on Aqueous Solubility Prediction. *Comb. Chem. High Throughput Screening* **2011**, *14* (5), 328−338.

(21) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. *In Silico* Prediction of Aqueous Solubility: the Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49* (11), 2572−2587.

(22) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2009**, *49* (1), 1−5.

(23) Gissi, A.; Gadaleta, D.; Floris, M.; Olla, S.; Carotti, A.; Novellino, E.; Benfenati, E.; Nicolotti, O. An Alternative QSAR-Based Approach for Predicting the Bioconcentration Factor for Regulatory Purpose. *ALTEX* **2014**, *31* (1), 23−36.

(24) Rudén, C.; Hansson, S. O. Registration, Evaluation, and Authorization of Chemicals (REACH) Is but the First Step - How Far Will It Take Us? Six Further Steps to Improve the European Chemicals Legislation. *Environ. Health Perspect.* **2010**, *118* (1), 6−10.

(25) Schoeters, G. The REACH Perspective: toward a New Concept of Toxicity Testing. *J. Toxicol. Environ. Health, Part B* **2010**, *13* (2−4), 232−241.

(26) Winder, C.; Azzi, R.; Wagner, D. The Development of the Globally Harmonized System (GHS) of Classification and Labelling of Hazardous Chemicals. *J. Hazard. Mater.* **2005**, *125* (1−3), 29−44.




DOI: 10.1021/acs.jcim.6b00625
*J. Chem. Inf. Model.* 2017, 57, 36−49

(27) Kujawski, J.; Popielarska, H.; Myka, A.; Drabińska, B.; Bernard, M. K. The LogP Parameter as a Molecular Descriptor in the Computer-Aided Drug Design - an Overview. *Comput. Methods Sci. Technol.* **2012**, *18* (2), 81−88.
(28) Zang, Q.; Rotroff, D. M.; Judson, R. S. Binary Classification of a Large Collection of Environmental Chemicals from Estrogen Receptor Assays by Quantitative Structure-Activity Relationship and Machine Learning Methods. *J. Chem. Inf. Model.* **2013**, *53* (12), 3244−3261.
(29) Vinggaard, A. M.; Niemelä, J.; Wedebye, E. B.; Jensen, G. E. Screening of 397 Chemicals and Development of a Quantitative Structure-Activity Relationship Model for Androgen Receptor Antagonism. *Chem. Res. Toxicol.* **2008**, *21* (4), 813−823.
(30) Scholz, S.; Sela, E.; Blaha, L.; Braunbeck, T.; Galay-Burgos, M.; García-Franco, M.; Guinea, J.; Klüver, N.; Schirmer, K.; Tanneberger, K.; et al. A European Perspective on Alternatives to Animal Testing for Environmental Hazard Identification and Risk Assessment. *Regul. Toxicol. Pharmacol.* **2013**, *67* (3), 506−530.
(31) Burden, N.; Sewell, F.; Chapman, K. Testing Chemical Safety: What Is Needed to Ensure the Widespread Application of Nonanimal Approaches? *PLoS Biol.* **2015**, *13* (5), e1002156.
(32) Bhhatarai, B.; Teetz, W.; Liu, T.; Öberg, T.; Jeliazkova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E.; et al. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Mol. Inf.* **2011**, *30* (2−3), 189−204.
(33) Zhang, J.; Liu, Z.; Liu, W. QSPR Study for Prediction of Boiling Points of 2475 Organic Compounds Using Stochastic Gradient Boosting. *J. Chemom.* **2014**, *28* (3), 161−167.
(34) Liang, C.; Gallagher, D. A. QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (2), 321−324.
(35) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult to Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and LogP. *J. Chem. Inf. Model.* **2008**, *48* (1), 220−232.
(36) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1208−1217.
(37) Lombardo, A.; Roncaglioni, A.; Boriani, E.; Milan, C.; Benfenati, E. Assessment and Validation of the CAESAR Predictive Model for Bioconcentration Factor (BCF) in Fish. *Chem. Cent. J.* **2010**, *4* (Suppl 1), S1.
(38) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 773−777.
(39) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45* (1), 30−38.
(40) Cheng, T.; Zhao, Y.; Li, X.; Lin, F.; Xu, Y.; Zhang, X.; Li, Y.; Wang, R.; Lai, L. Computation of Octanol-Water Partition Coefficients by Guiding an Additive Model with Knowledge. *J. Chem. Inf. Model.* **2007**, *47* (6), 2140−2148.
(41) Meylan, W. M.; Howard, P. H. Estimating LogP with Atom/Fragments and Water Solubility with LogP. *Perspect. Drug Discovery Des.* **2000**, *19* (1), 67−84.
(42) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 266−275.
(43) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 439−445.
(44) Dearden, J. C. Prediction of Physicochemical Properties. In *Computational Toxicology*. Methods in Molecular Biology; Reisfeld, B., Mayeno, A. N., Eds.; Humana Press: New York, USA, 2012; Vol. *929*, pp 93−138.
(45) EPI Suite−Estimation Program Interface. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface (accessed December 16, 2016).
(46) The influence of Data Curation on QSAR Modeling−Examining Issues of Quality versus Quantity of Data. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=311418 (accessed December 16, 2016).
(47) Mansouri, K.; Grulke, C. M.; Richard, A. M.; Judson, R. S.; Williams, A. J. An Automated Curation Procedure for Addressing Chemical Errors and Inconsistencies in Public Datasets Used in QSAR Modeling. *SAR QSAR Environ. Res.* **2016**, *27* (11), 939−965.
(48) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(49) OECD Principles for the Validation of (Q)SARs. https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/background/oecd-principles (accessed December 16, 2016).
(50) EPI Suite Data. http://esc.syrres.com/interkow/EPiSuiteData.htm (accessed December 16, 2016).
(51) Yap, C. W. PaDEL-Descriptor: an Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466−1474.
(52) PaDEL-Descriptor. http://www.yapcwsoft.com/dd/padeldescriptor/ (accessed December 16, 2016).
(53) Judson, R. S. Genetic Algorithms and Their Use in Chemistry. In *Reviews in Computational Chemistry*, first ed.; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publisher, Inc.: New York, USA, 1997; pp 1−73.
(54) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 1077−1084.
(55) Zang, Q.; Keire, D. A.; Wood, R. D.; Buhse, L. F.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; Trehy, M. L.; Welsh, W. J. Determination of Galactosamine Impurities in Heparin Samples by Multivariate Regression Analysis of Their $^{1}$H NMR Spectra. *Anal. Bioanal. Chem.* **2011**, *399* (2), 635−649.
(56) Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC Press: Boca Raton, FL, USA, 2009.
(57) Cao, S. S.; Xu, Q. S.; Liang, Y. Z.; Chen, X.; Li, H. D. Prediction of Aqueous Solubility of Druglike Organic Compounds Using Partial Least Squares, Back-Propagation Network and Support Vector Machine. *J. Chemom.* **2010**, *24* (9), 584−595.
(58) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47* (1), 150−158.
(59) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1855−1859.
(60) Zang, Q.; Keire, D. A.; Buhse, L. F.; Wood, R. D.; Mital, D. P.; Haque, S.; Srinivasan, S.; Moore, C. M. V.; Nasr, M.; Al-Hakim, A.; et al. Identification of Heparin Samples That Contain Impurities or Contaminants by Chemometric Pattern Recognition Analysis of Proton NMR Spectral Data. *Anal. Bioanal. Chem.* **2011**, *401* (3), 939−955.
(61) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22* (1), 69−77.
(62) Veerasamy, R.; Rajak, H.; Jain, A.; Sivadasan, S.; Varghese, C. V.; Agrawa, R. K. Validation of QSAR Models−Strategies and Importance. *Int. J. Drug Des. Discovery* **2011**, *2* (3), 511−519.
(63) 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* (5), 4791−4810.
(64) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Assessing the Validity of QSARs for Ready Biodegradability of Chemicals: an Applicability Domain Perspective. *Curr. Comput.-Aided Drug Des.* **2014**, *10* (2), 137−147.
(65) Sahigara, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Defining a Novel-K-Nearest Neighbours Approach to Assess the Applicability


DOI: 10.1021/acs.jcim.6b00625
*J. Chem. Inf. Model.* 2017, 57, 36−49

Domain of a QSAR Model for Reliable Predictions. *J. Cheminf.* **2013**, *5* (1), 27.
(66) R Development Core Team R: *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011; http://www.R-project.org/ (accessed December 16, 2016).
(67) Tetko, I. V.; Lowe, D. M.; Williams, A. J. The Development of Models to Predict Melting and Pyrolysis Point Data Associated with Several Hundred Thousand Compounds Mined from PATENTS. *J. Cheminf.* **2016**, *8* (1), 2.



DOI: 10.1021/acs.jcim.6b00625
*J. Chem. Inf. Model.* 2017, 57, 36−49