Research article

Journal of CHEMOMETRICS

Received: 24 November 2014, Revised: 29 March 2015, Accepted: 31 March 2015, Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/cem.2718

In silico evaluation of $logD_{7.4}$ and comparison with other prediction methods

Jian-Bing Wang[a,b], Dong-Sheng Cao[a]*, Min-Feng Zhu[c], Yong-Huan Yun[b], Nan Xiao[c] and Yi-Zeng Liang[b]*

Lipophilicity, evaluated by either *n*-octanol/water partition coefficient or *n*-octanol/buffer solution distribution coefficient, is of high importance in pharmacology, toxicology, and medicinal chemistry. A quantitative structure–property relationship study was carried out to predict distribution coefficients at pH 7.4 ($logD_{7.4}$) of a large data set consisting of 1130 organic compounds. Partial least squares and support vector machine (SVM) regressions were employed to build prediction models with 30 molecular descriptors selected by genetic algorithm. The obtained results demonstrated that the SVM model is more reliable and has a better prediction performance than the partial least squares model. The square correlation coefficients of fitting, cross validation, and prediction are 0.92, 0.90, and 0.89, respectively. The corresponding root mean square errors are 0.52, 0.59, and 0.56, respectively. The robustness, reliability, and generalization ability of the model were assessed by *Y*-randomization test and applicability domain. When compared with $logD_{7.4}$ values calculated by five existing methods from Discovery Studio and ChemAxon, our SVM model shows superiority over them. The results indicated that our model could give a reliable and robust prediction of $logD_{7.4}$. Copyright © 2015 John Wiley & Sons, Ltd.
Supporting information may be found in the online version of this paper

Keywords: lipophilicity; distribution coefficient; $logD_{7.4}$; quantitative structure–property relationship (QSPR); support vector machine (SVM); genetic algorithm (GA)

1. INTRODUCTION

To exert a therapeutic effect, one drug must enter the blood circulation and then reach the site of action. Thus, an eligible drug usually needs to keep a balance between lipophilicity and hydrophilicity to dissolve in the body fluid and penetrate the biofilm effectively. Therefore, it is very important to evaluate the lipophilicity of candidate compounds in drug research and development process. Additionally, several studies have reported the effect of lipophilicity on biological activities and transport properties [1–7], indicating the importance to evaluate the lipophilicity of new drugs or pro-drugs.

The lipophilicity of a compound can be quantitatively characterized by the partition coefficient (its logarithm form is denoted as logP) or the distribution coefficient (its logarithm form is denoted as logD) if ionized molecular species are present [4,5]. Partition coefficient [8] is the equilibrium concentration ratio of the solute between two immiscible solvents (e.g., *n*-octanol and water). Although it is a major descriptor in many quantitative structure–activity relationship (QSAR)/quantitative structure–property relationship (QSPR) equations [8–11] and a crucial part of Lipinski's rule of five [12,13], logP only refers to the neutral form of the compound and is independent of the ionization under physiological conditions. However, it is estimated that 95% of all drugs are ionizable [14,15]. Thus, the distribution coefficient, which takes account of ionization, may be a more reliable measurement for the lipophilicity at physiological pH [14–17]. Distribution coefficient, also known as pH-dependent distribution coefficient, is the ratio of the sum of the equilibrium concentrations of all forms of the compound (i.e., the total sum of ionized and unionized) between two phases. There are several experimental methods to measure the logD value [18] such as the shake-flask method, the slow stirring method, the filter probe method, some chromatography methods, and pH metric techniques. However, these experimental procedures are costly and time-consuming and require substantial quantities of the compound being synthesized. Hence, it is necessary to establish a reliable prediction model to accurately determine logD values without the need for experiments, especially for new or even virtual compounds.

Currently, there are three main thoughts to estimate logD *in silico*. (i) logD is calculated from logP and dissociation constant (pKa), assuming that only the neutral species exist in the nonaqueous phase [19,20]. However, the presence of both the ionized and unionized species that necessitate a dissociative

* *Correspondence to: D. S. Cao, School of Pharmaceutical Sciences, Central South University, Changsha 410008, China.*
E-mail: oriental-cds@163.com

Y. Z. Liang, College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China.
E-mail: yizeng_liang@263.net

a *J.-B. Wang, D.-S. Cao*
School of Pharmaceutical Sciences, Central South University, Changsha 410008, China

b *J.-B. Wang, Y.-H. Yun, Y.-Z. Liang*
College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China

c *M.-F. Zhu, N. Xiao*
School of Mathematics and Statistics, Central South University, Changsha 410083, China

equilibrium in the non-aqueous phase may bring a big error. (ii) The fragment-based methods divide a molecule into various fragments (either at the molecular level or at the atomic level) and then aggregate all the contributions of these individual fragments to obtain the predicted logD value by a linear model [18]. However, the nonlinear relationship between logD and the fragment values is usually neglected. (iii) The molecular property-based methods utilize numeric characteristics of the entire molecule to predict logD values. These characteristics are usually known as molecular descriptors and normally calculated from the topological structure of the molecule. Combined with different modeling methods, the model can be linear or nonlinear. Nowadays, some of the predictive methods are available from commercial and academic resources, such as the ALOGPS [21,22], ChemAxon [23,24], Discovery Studio, and so on [18]. ALOGPS adjusts logP prediction with a library of measured logD data to make a logD prediction. It is sometimes difficult to accurately calculate logD values for new molecules. Four methods from ChemAxon software belong to the fragment-based methods [23] and thereby rely on the quality of defined fragments to a large extent. Although these approaches have been successfully applied to the logD prediction for some molecules, they are usually not global in their success and require some modification when applied to new data or a larger chemical library [25]. Additionally, to our knowledge, it has rarely been performed by modeling logD directly except two studies by Pierre Bruneau with Nathan R. McElroy [25] and Cerep [26].

In the present study, our aim focuses on the accurate *in silico* prediction of $logD_{7.4}$ with a large and diverse data set collected from the internet. Support vector machine (SVM), as a popular machine learning algorithm, was used to establish the predicted model for $logD_{7.4}$. After descriptors of each molecule were calculated and selected, a predicted model was built with the basic assumption that compounds with similar chemical structures have similar properties [27,28]. Model validation and evaluation approaches, applicability domain analysis, and comparison with several available methods from two pieces of software (ChemAxon and Discovery Studio) were used to assess the robustness and reliability of our model. The selected descriptors provide some hints for mechanism of action related to $logD_{7.4}$. The results demonstrate that the model built by SVM has a good predictive ability and could give some guidelines for improving undesirable $logD_{7.4}$ in rational drug design.

2. MATERIALS AND METHODS

2.1. Data collection

Experimental $logD_{7.4}$ values were collected from two resources. One is the ChEMBL database (https://www.ebi.ac.uk/chembl/) including 1451 logD values. The other one is the online chemical database (https://ochem.eu/). This database includes 367 logD values. As the logD data in two databases were collected from different literature sources, we only extracted those logD values under the homogeneous experimental conditions, that is, pH = 7.4, temperature is 25 °C, and the organic solvent is *n*-octanol. Only those molecules with reliable logD values were considered, and those molecules with empty or indeterminate logD values were removed. If there were two or more entries for one molecule, the arithmetic mean value of these values was adopted. The data set was filtered to remove compounds with logD values greater than 10 or less than −10 because of their potential unreliability. One record was reserved for the conformational or optical isomers, because the subsequent analysis did not consider the stereochemistry. Solvent or saline ions adhering to the molecule were removed automatically by OpenBabel (http://openbabel.org/wiki/Get_Open_Babel). The SMILES structures of these compounds were checked one by one to ensure that they were correct. After a series of pretreatments, 1130 molecules and their $logD_{7.4}$ values were finally collected. Their SMILES structures and experimental $logD_{7.4}$ values can be found in the Supporting Information (see S1).

2.2. Descriptor calculation and pruning

The SMILES structures of all 1130 molecules were imported into the Molecular Operating Environment software (version 2011.10) to calculate two-dimensional descriptors, resulting in 188 descriptors. All descriptors were firstly checked to ensure that each descriptor value is available for each molecule. Before further descriptor selection, two descriptor pre-selection steps were performed to eliminate some uninformative descriptors: (1) remove descriptors whose variance is zero or near zero; and (2) if the correlation of two descriptors is larger than 0.95, one of them was removed. Finally, 121 molecular descriptors were obtained to represent each compound, and these molecular descriptors were used as inputs for further variable selection and the QSPR model construction. These descriptors were listed in the Supporting Information (see S2).

According to the Organization for Economic Cooperation and Development (OECD) principles, the QSPR models should be checked by both internal and external statistical validation to ensure both reliability and predictive ability of the derived model. Herein, all molecules were divided into two parts, namely the training set and the test set, using the Kennard–Stone method [29,30] to guarantee that the test samples could map the measured region of the input variables space completely. Thus, we obtained a training set of 904 molecules (80% of the data set) and a test set of 226 molecules (20% of the data set). The training set was used to construct the prediction model, and the test set was used for further assessing the performance of the model.

2.3. Support vector machine algorithm

Support vector machine developed by Vapnik and coworkers is based on the structural risk minimization principle from statistical learning theory. Although developed for classification problems, SVM can also be applied to the case of regression. Detailed descriptions of SVM can be easily found in several excellent books and literature [31–36]. As an excellent machine learning algorithm, SVM can be used to solve linear and nonlinear regression problems with good prediction performance [11,34] and has been successfully used to solve many QSAR regression problems in previous studies [37–40]. For linear regression cases, given the training data $D = \{(X_i, , y_i)\}_{i=1}^{N}$ (x_i is the input vector representing some molecular descriptors and y_i is the output vector representing the experimental bioactivity values), SVM approximates the function in the following way:

$$f(\mathbf{x}_i) = \mathbf{w}^t\mathbf{x}_i + b \tag{1}$$

where **w** is a vector of weights, and *b* is the constant coefficient. These can be estimated by minimizing the regularized risk loss $R(C)$:

$$R(C) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N} L(y_i - f(\mathbf{x}_i), \varepsilon) \quad (2)$$

where C is a predefined regularizing parameter. $L(y - f(\mathbf{x}), \varepsilon)$ is a ε-intensive loss function defined as

$$L(y - f(\mathbf{x}), \varepsilon) = \begin{cases} |y - f(\mathbf{x})| - \varepsilon, & \text{if } |y - f(\mathbf{x})| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Formula 2 can further be expressed in the following form with a slack variable ξ introduced:

$$\begin{aligned} \text{minimize} \ : \ & R(C) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ & w^t\mathbf{x}_i + b - y_i \leq \varepsilon + \xi_i \\ \text{subject to} \ : \ & y_i - (w^t\mathbf{x}_i + b) \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0 \quad i = 1, 2, \ldots, N \end{aligned} \quad (4)$$

With the help of the Lagrange multiplier method and the quadratic programming algorithm, the minimum problem can be solved as

$$f(\mathbf{x}) = \sum_{i=1}^{N}(a_i^* - a_i)(\mathbf{x}_i^t\mathbf{x}_i) + b \quad (5)$$

$$b = y_j - \sum_{i=1}^{N}(a_i^* - a_i)(\mathbf{x}_i^t\mathbf{x}_j) + \varepsilon \quad (6)$$

where a_i^* and a_i are the optimized Lagrange multipliers. For nonlinear regression cases, SVM projects the input feature vectors into a high-dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. Especially, the Gaussian kernel, which has been extensively used in different studies with good performance, can be represented as

$$K(\mathbf{x}_i\mathbf{x}_j) = \exp\left(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/2\delta^2\right) \quad (7)$$

Thus, a linear SVM is applied to this high-dimensional feature space, and the solutions are given by

$$f(\mathbf{x}) = \sum_{i=1}^{N}(a_i^* - a_i)K(\mathbf{x}_i, \mathbf{x}) + b \quad (8)$$

$$b = y_j - \sum_{i=1}^{N}(a_i^* - a_i)K(\mathbf{x}_i, , \mathbf{x}_j) + \varepsilon \quad (9)$$

The previous SVM algorithm is called ε-SVM. In ε-SVM, some parameters (e.g., the regularization parameter C, ε-insensitive loss function, and the type and parameters of kernel function) are important parameters that need to be further optimized. The regularization parameter C is an important parameter because of its possible effects on both trained and predicted results, because it controls the tradeoff between maximizing the margin and minimizing the training error. Usually, C is an unknown parameter before modeling. If C is too small, an insufficient stress will be placed on fitting the training data. If C is too large, the algorithm will overfit the training data. Therefore, C should be optimized together with the kernel functions. The parameter ε is also an important parameter, which depends on the type of noise present in the data, and it is usually unknown. There is a practical consideration of the number of resulting support vectors, even if enough knowledge of the noise is given to select an optimal ε. It prevents the entire training set from meeting boundary conditions, so we have to optimize ε and seek for the optimal value. In this work, the Gaussian kernel function was used to model the nonlinear relationship, and grid search was used to obtain the best combination of parameters.

2.4. Model validation

To ensure that the derived model from the training set has a good generalization ability, fivefold cross validation [41–43] and an external test set were used for the validation purpose. For fivefold cross validation, the training set was split into five roughly equal-sized parts firstly. Then the model was built with four parts of the data and the prediction error of the other one part was calculated. The process was repeated five times so that every part could be used as a validation set. Four commonly used parameters in regression problems were employed to evaluate the model performance [43], including the square correlation coefficients of fitting (R^2), the square correlation coefficients of cross validation (Q^2), the root mean squared error of fitting ($RMSE_F$), and the root mean squared error of cross validation ($RMSE_{CV}$). These regression statistics are defined as follows:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2} \quad (10)$$

$$RMSE_F = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (11)$$

$$Q^2 = 1 - \frac{\sum\left(\hat{y}_{(v)i} - y_i\right)^2}{\sum(y_i - \overline{y})^2} \quad (12)$$

$$RMSE_{CV} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_{(v)i}\right)^2} \quad (13)$$

where y_i is the experimental value of the ith sample in the training set; $\hat{y}_i$ is the predicted value of the ith sample in the training set; $\overline{y}$ is the mean value of all experimental values in the training set; $\hat{y}_{(v)i}$ is the predicted value of the ith sample for cross validation; and N is the number of samples in the training set. When the external test set was performed, the statistics RT^2 and $RMSE_P$ were calculated in the similar way.

3. RESULTS AND DISCUSSION

3.1. Descriptor selection by genetic algorithm-based multivariate linear regression

Genetic algorithm-based multivariate linear regression (GA-MLR) was used to select the best subset of descriptors, which are the most relevant variables in modeling $logD_{7.4}$. GA, as a relatively high-efficiency evolutionary algorithm, has been widely applied to various research fields [44–47]. Herein, we adopted GA-MLR in the Molecular Operating Environment software (version 2011.10) to search for the optimal descriptor combination, which could give the highest adjusted R^2 (i.e., $adjR^2$). The parameters used in GA-MLR are as follows: population size = 100, operant density = 4, generation = 50000,

mutation probability = 0.5, eugenic factor = 100, and auto termination = 1000. Each constructed MLR model in the evolution process was evaluated by leaving 20% out cross validation.

To study the influence of the fixed number of descriptors on model performance, GA-MLR based on the fixed-length search was employed. That is, we allowed GA-MLR to search the optimal descriptor subset with the fixed number at each turn. The number of descriptors was firstly fixed on 4, and then increased by one descriptor once. The relationship between the fitness ($adjR^2$) and the number of selected descriptors is shown in Figure 1. From this plot, we can see that $adjR^2$ tends to improve with the increasing of the number of descriptors. The curve rises quickly at the beginning and then flattens out. When the number of descriptors reaches 30, the $adjR^2$ is almost invariable. Thus, to balance the model complexity and model performance, we finally selected 30 descriptors to establish the prediction models in the following study. These selected descriptors could be found in the Supporting Information (see S3).

3.2. Model building and external validation

Thirty descriptors selected by GA-MLR were used as an input to generate the QSPR model by SVM. We used fivefold cross validation to evaluate the model performance, as mentioned in Section 2.4. By means of the grid search, the optimal values of δ, C, and ε are set to 2^{-9}, 64, and 0.25, respectively. Once the parameters were optimized, we could use them to establish the SVM model and perform the subsequent prediction. For the training set, $R^2 = 0.92$, $RMSE_F = 0.51$, and for cross validation, $Q^2 = 0.90$, $RMSE_{CV} = 0.59$. When applying the SVM model to the test set, we obtained $RT^2 = 0.89$, $RMSE_P = 0.56$. As a comparison, partial least squares (PLS) was also employed to establish the QSPR model with the same 30 descriptors. The number of the latent variables was determined through fivefold cross validation. For the training set, $R^2 = 0.87$, $RMSE_F = 0.66$, and for cross validation, $Q^2 = 0.86$, $RMSE_{CV} = 0.68$. When applying the PLS model to the test set, we obtained $RT^2 = 0.83$, $RMSE_P = 0.69$. Figure 2 shows the prediction results by PLS and SVM. The experimental $\log D_{7.4}$ and predicted $\log D_{7.4}$ by PLS and SVM could be found in the Supporting Information (see S1).

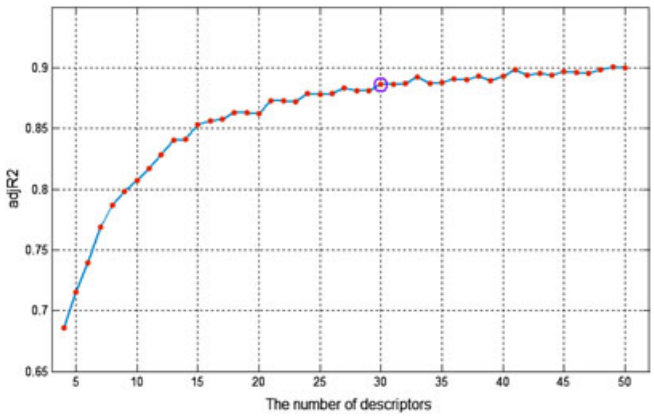

Figure 1. Relationship between fitness ($adjR^2$) and the number of fix descriptors in genetic algorithm-based multivariate linear regression QSAR modeling of the $\log D_{7.4}$ value.

Table I lists the regression statistics of two prediction models. As can be shown, the SVM and PLS models all obtain satisfactory prediction results, indicating that 30 selected molecular descriptors can effectively model $\log D_{7.4}$. It can be seen from Table I that the results from SVM seem better than those from PLS, in terms of different regression statistics. The results indicate that there may be a certain nonlinear relationship between the selected descriptors and $\log D_{7.4}$. Furthermore, for two models, Q^2 is all slightly lower than R^2, indicating that these two models are reliable and have avoided the overfitting effect. When applying them to the test set, we find that RT^2 is also only a litter lower than Q^2. As the compounds of the test set were not used in model building, this phenomenon indicates that our model may be generally applicable. The performance of the SVM model is also comparable with the results that were derived by Bayesian regularized neural networks [25].

Additionally, our model was further evaluated by several stricter criteria provided by Tropsha *et al.* [48–50]. According to the suggestions from Tropsha *et al.*, a QSAR/QSPR model is successful if it satisfies several criteria as follows:

$$Q^2 > 0.5, R^2 > 0.6$$
$$\frac{R^2 - R_0^2}{R^2} < 0.1 \text{ or } \frac{R^2 - R_0'^2}{R^2} < 0.1$$
$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15$$

where Q^2 and R^2 are the square of the correlation coefficient for cross validation and external test set, respectively; R_0^2 and $R_0'^2$ are the determination coefficient of predicted versus experimental values and experimental versus predicted values, respectively; k

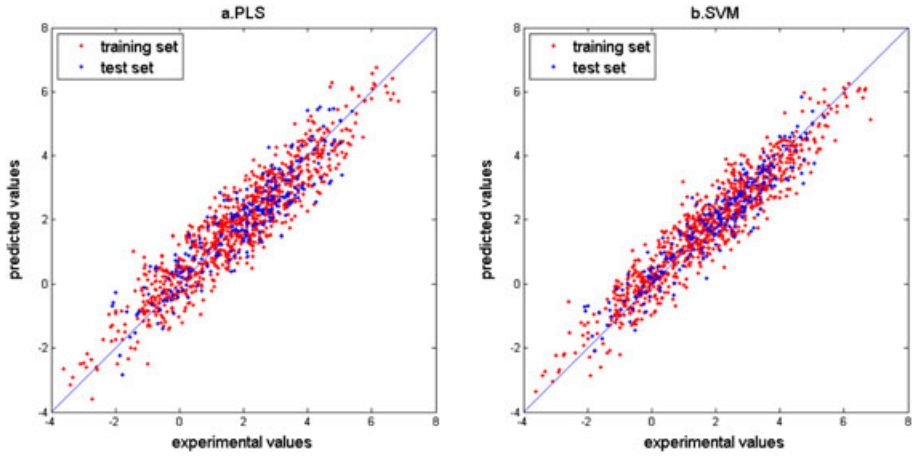

Figure 2. Plot of predicted $\log D_{7.4}$ versus experimental $\log D_{7.4}$ for the training set (red) and the test set (blue) (a. the partial least squares (PLS) model; b. the support vector machine (SVM) model).

Table I. Statistical results of the fitting, fivefold cross validation, and independent test set for PLS and SVM

	Training set ($N = 904$)				Test set ($N = 226$)	
	Fitting		Cross validation		Prediction	
	R_F^2	$RMSE_F$	Q^2	$RMSE_{CV}$	RT^2	$RMSE_P$
PLS	0.87	0.66	0.86	0.68	0.83	0.69
SVM	0.92	0.51	0.90	0.59	0.89	0.56

PLS, partial least squares; SVM, support vector machine; RMSE, root mean squared error.

and k' are the slope of the regression lines through the origin, respectively. According to the aforementioned criteria, our SVM model could be considered acceptable as it satisfies all the following conditions: $Q^2 = 0.90 > 0.5$; $R^2 = 0.89 > 0.6$, $\frac{R^2 - R_0^2}{R^2} = 0.007 < 0.1$, $\frac{R^2 - R_0'^2}{R^2} = 7.764 \times e^{-6} < 0.1$, $0.85 \le k = 0.970 \le 1.15$, and $0.85 \le k' = 0.987 \le 1.15$. In summary, SVM is suitable to predict $logD_{7.4}$ with the selected molecular descriptors. In the latter section, we mainly focus on the predictive results of SVM.

3.3. Y-randomization test

When selecting descriptors that are of relevance to model the property of interest, it is possible to find some descriptors that seem of importance, just-by-chance, given the high dimensionality of feature space from which we are doing such a search by some optimization tools such as GA. To guard against the possibility of having learned such chance models, *Y*-randomization test was advocated to validate the reliability of our QSPR model [43,49,51–53]. In *Y*-randomization test, the $logD_{7.4}$ values were randomly shuffled to change their true order. Thus, although the $logD_{7.4}$ values (and the statistical distribution) stayed the same, their position against the appropriate compound and its descriptors was now altered, thus destroying any meaningful relation that may have existed between independent variables and response values. By these new data, we constructed a large number of QSPR models (e.g., 1000) to obtain metrics like Q^2. These metrics could be compared with those from the true model to obtain some hints about chance correlation. Figure 3 shows the distribution diagram of the Q^2 values of 1000 randomized models and the real model. One can see that Q^2 of randomly shuffled models are located in the range from −0.2 to 0. Compared with the true model ($Q^2 = 0.90$), there is a significant difference between the Q^2 of these shuffled models and the real one. The bad prediction statistics of these shuffled models suggest that our previous model indeed reflects the true relationship between molecular descriptors and $logD_{7.4}$ values rather than from chance correlation.

3.4. Applicability domain evaluation

The applicability domain (AD) allows one to estimate the uncertainty in the prediction of a particular molecule based on how similar it is to the compounds used to build the model [54–60]. This is consistent with the applicability domain criterion in the OECD principles (the third principle of the OECD principles: a defined domain of applicability) [61,62]. Prediction of a molecule in a given model is most likely to be reliable if this molecule falls within the AD; otherwise, its prediction is likely to be unreliable [56,57]. In this study, we used the Williams plot to evaluate the AD of our QSPR model. The Williams plot provides leverage values plotted against the prediction errors. The leverage value (h) measures the distance from the centroid of the training set and could be calculated for a given dataset **X** by obtaining the leverage matrix (**H**) as follows: [54,57]

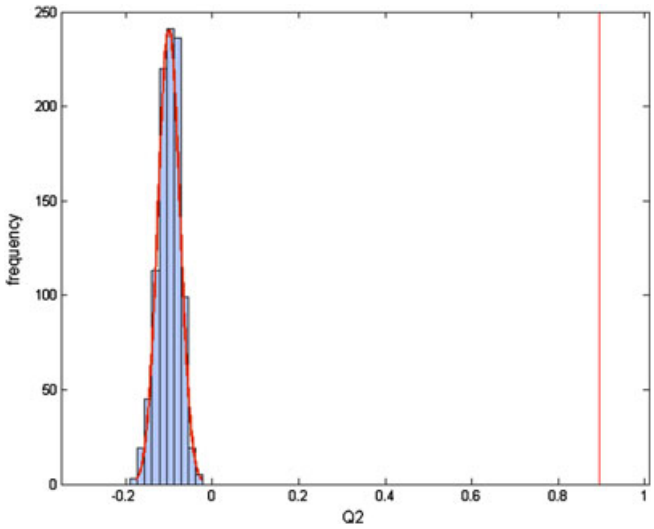

Figure 3. The distribution of Q^2 of randomized models compared with the real model in the *Y*-randomization test. (The red vertical line on the right represents the Q^2 of the true model, and the distribution on the left side represents the distribution of Q^2s of models after randomization).

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

where **X** is the descriptor matrix; $\mathbf{X}^T$ is its transpose matrix; and $(\mathbf{X}^T\mathbf{X})^{-1}$ is the inverse of $(\mathbf{X}^T\mathbf{X})$. The diagonal elements in the **H** matrix represent the leverage values (h) for the molecules in the dataset. The warning leverage, h^*, was fixed at $3p/n$ ($h^* = 0.100$) in this study, where p is the number of descriptors and n is the number of training samples. A query molecule with leverage higher than h^* may be associated with unreliable predictions. Such molecules are believed outside the descriptor space and thus will be considered outside the AD. Figure 4 shows the Williams plot based on leverage values and prediction errors. In Figure 4, the horizontal line divides the leverage value axis into two parts, and the points above the line have a greater leverage than h^*(i.e., $3p/n = 0.100$); while the two vertical lines divide the prediction error axis into three parts and verify the presence of compounds with prediction errors greater than three standard deviation units ($\pm 3\sigma$, i.e., ±1.794) in the training set.

With this Williams plot, the AD of our QSPR model could be defined. It can be seen from Figure 4 that a great majority of compounds in the training and test set fall within the AD. In this domain, the chemical with a low leverage often has a low prediction error, indicating these compounds are likely to be well predicted by the SVM model. However, there are still a few compounds locating outside the AD, indicating predictions of these compounds are likely to be unreliable. Additionally, we can find that some compounds with high leverages have low prediction errors, while some compounds with low leverages possess high prediction errors. This could be explained by the

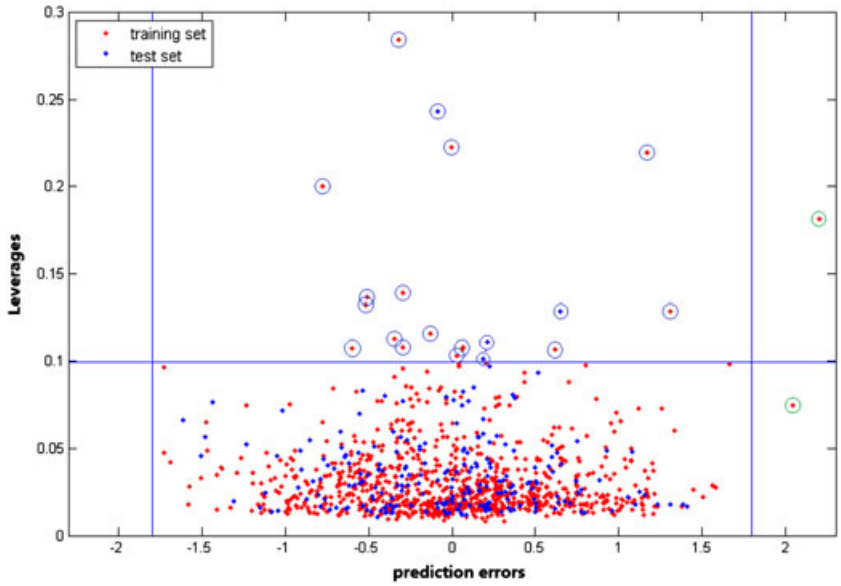

Figure 4. Williams plot of leverages versus prediction errors. The horizontal line represents the warning leverage value ($h^* = 3p/n \approx 0.100$), and the vertical lines indicate the place of ±3 standard deviation units.

fact that the defined AD only considers interpolation by simply excluding all samples in the extremities and including all those surrounded by training samples. As we can see from Figure 4, some compounds, marked by circles, are identified as outliers. Two molecules (A-log$D_{7.4}$-24 and O-412) have large prediction errors and thereby were diagnosed as *y*-direction outliers. One (A-log$D_{7.4}$-24) is a peptide, whose sequence is Ac-Tyr-Gly-Gly-Gln-NH_2. The other one (O-412) is rifampin, a bactericidal antibiotic drug, which is often used in the treatment of tuberculosis, *Enterococcus* infection, and so on. Figure 5 shows the structures of two *y*-direction outliers. After removal of these two outliers marked by green circles, Q^2 increased by 0.03 and $RMSE_{CV}$ decreased by 0.01. Furthermore, the compounds marked by blue circles have relatively high leverage values but low prediction errors. We identified them as **X** direction outliers. These **X** direction outliers are far away from the main body of the training set. However, they do not have big prediction errors and thereby do not damage the prediction performance. In summary, the AD defined by the Williams plot is reasonable. We can use it to evaluate the reliability of our future predictions.

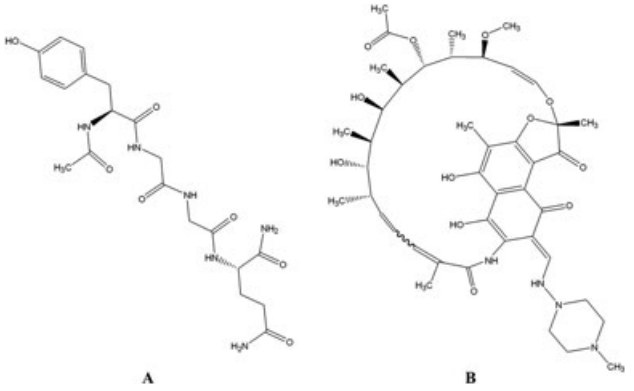

Figure 5. The structure of two compounds that diagnosed as *y*-direction outliers (A. A-log$D_{7.4}$-24,apeptide, whose sequence is Ac-Tyr-Gly-Gly-Gln-NH_2; B. O-412, rifampin).

3.5. Comparison with other methods

To further evaluate the prediction performance of our model, our result was compared with several existing methods from ChemAxon (version 5.4.1.1) and Discovery Studio (version 2.5, DS for short). In ChemAxon, four methods can be used to predict logD, including VG, KLOP, PHYS, and Weight (the average of the previous three methods). Log$D_{7.4}$ is calculated from calculated pKa and logP values, which are evaluated by the additive model. The difference in the first three methods is the diverse fragment set used for modeling. In DS, log$D_{7.4}$ is calculated as a descriptor for further molecular simulation. The SMILES structures of all molecules were imported into the ChemAxon and DS software to calculate log$D_{7.4}$. The calculation results of five methods are listed in the Supporting Information (see S1). The regression statistics of these methods are summarized in Table II.

As we can see from Table II, R^2 values of five methods from two pieces of software are 0.63, 0.67, 0.64, 0.62, and 0.71, respectively. Among these prediction methods, R^2 follows the decreasing order: SVM > Marvin_PHYS > Marvin_weight > Marvin_VG > DS > Marvin_KLOP. Similarly, *RMSE* and mean absolute error follow the increasing order: SVM < Marvin_PHYS < Marvin_weight< DS < Marvin_VG < Marvin_KLOP. Clearly, the SVM method obtained the best

Table II. Comparison of prediction statistics for different modeling methods

	R^2	*R*	*RMSE*	*MAE*
SVM	0.90	0.95	0.59	0.47
DS	0.63	0.79	1.19	0.90
Marvin_weight	0.67	0.82	1.19	0.86
Marvin_VG	0.64	0.80	1.26	0.93
Marvin_KLOP	0.62	0.78	1.32	0.95
Marvin_PHYS	0.71	0.84	1.10	0.80

SVM, support vector machine; RMSE, root mean squared error; MAE, mean absolute error; DS, Discovery Studio.

prediction performance ($R^2 = 0.90$, $RMSE = 0.59$, mean absolute error = 0.47). Figure 6 shows the predictive results from six methods including our proposed SVM model. One can see that SVM model has the best correlation between the predicted and experimental $logD_{7.4}$ values. To further observe the difference between these methods, we plotted the diagram of correlation matrix depicting the patterns of relations between the experimental and all calculated $logD_{7.4}$ values (Figure 7). In Figure 7, the diagonal element is the experimental method (Exp) or several predicted methods (SVM, DS, M_W, M_VG, M_K, and M_P). Each row or column displays the pair correlations between the method expressed by the diagonal element and the other methods. From Figure 7, one can see that SVM has the best correlation with the experimental $logD_{7.4}$ values. We can also see that four methods from ChemAxon software have very high correlation coefficients (i.e., ≥0.95). This phenomenon can be explained by the fact that these three methods (Marvin_VG, Marvin_KLOP, and Marvin_PHYS) have the same idea to calculate the $logD_{7.4}$ and Marvin_Weight adopted the average of three aforementioned methods as the calculated $logD_{7.4}$ value.

Usually, we consider predictions with absolute deviations <0.50 as good estimates. Predictions with deviations ≥0.50 and <1.0 are considered disputable, while predictions with deviations ≥1 are unacceptable [63–65]. Table III summarizes three different error levels for each method. From Table III, SVM obtains the most reliable result as 60.35% of 1130 samples are well predicted, followed by Marvin_Weight (42.21%), Marvin_PHYS (41.77%), DS (40.09%), Marvin_VG (36.64%), and Marvin_KLOP (35.13%). For predictions with deviations ≥0.50 and <1.0, SVM also obtains satisfactory results. Additionally, one can see that Marvin_Weight has a better estimate than Marvin_PHYS, Marvin_VG, and Marvin_KLOP. This phenomenon confirms that taking the average value of many different methods is an efficient way to give a more reliable prediction, especially when we do not know which one is prevailing.

Through a comprehensive comparison with five existing methods, the SVM model shows the best prediction performance guaranteed by good agreement between the experimental and calculated values and accompanied by good regression statistics. The reason why the SVM model obtains better results could be summarized into three points. Firstly, $logD_{7.4}$ was calculated by pKa and logP in ChemAxon, while pKa and logP were evaluated by additive methods. In consequence, multi-step estimation may bring extra error for $logD_{7.4}$ prediction. Secondly, to some extent, the prediction methods in ChemAxon are a group contribution-based additive model. As a result, the model could be linear and cannot well account for the nonlinear relationship between the $logD_{7.4}$ and molecular descriptors. The third reason is probably the diversity of the data. In this study, we collected a big data set, which was not used previously. Therefore, we can reasonably speculate that there may be some fragments that were not been defined in ChemAxon.

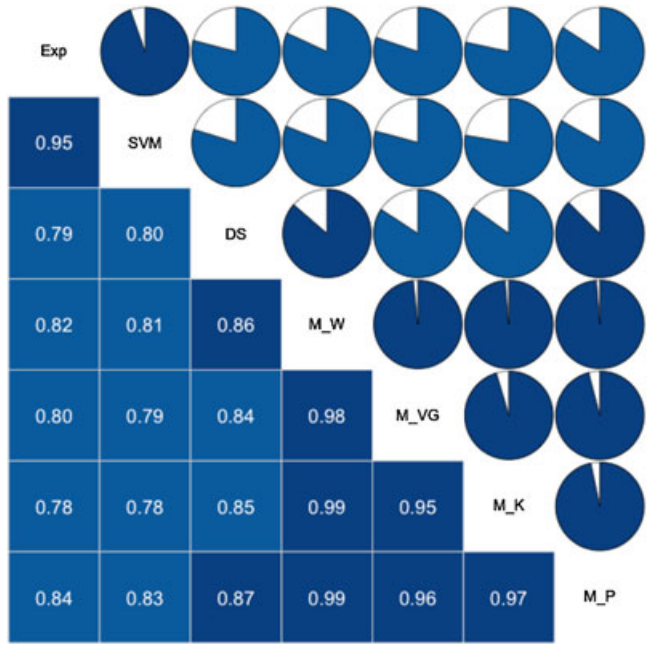

Figure 7. Diagram of correlation matrices of experimental and calculated values of each method. SVM, support vector machine; DS, Discovery Studio.

3.6. Model interpretation

Because many different combinations of descriptors may yield a similar prediction performance, sometimes it is difficult to

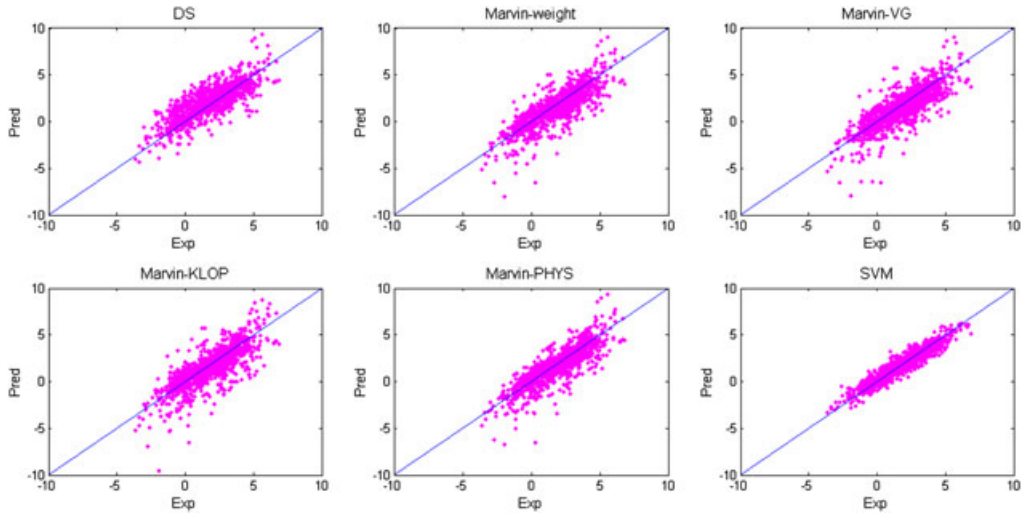

Figure 6. Comparison of the correlationships between experimental and calculated values of each method. DS, Discovery Studio.

Table III. Comparison of the different residual levels of each method

	0–0.5		0.5–1.0		≥1.0	
	Amount	Proportion (%)	Amount	Proportion (%)	Amount	Proportion (%)
SVM	682	60.35	341	30.18	107	9.47
DS	453	40.09	279	24.69	398	35.22
Marvin_weight	477	42.21	302	26.73	351	31.06
Marvin_VG	414	36.64	312	27.61	404	35.75
Marvin_KLOP	397	35.13	337	29.82	396	35.04
Marvin_PHYS	472	41.77	343	30.35	315	27.88

SVM, support vector machine; DS, Discovery Studio.

establish a rich and interpretable model. However, it can still provide some hints for mechanism of action related to $logD_{7.4}$. To well compare and interpret every molecular descriptor, the variable importance of molecular descriptors was computed based on SVM. At each round, one descriptor was removed from the molecular descriptor pool; the remaining 29 molecular descriptors were used to build the model, and a new Q^2 was computed. The difference between the new Q^2 and the preceding one can be seen as a measure of variable importance of the removed molecular descriptor. The process was repeated 30 times, and the importance of every molecular descriptor was obtained and displayed in Figure 8.

From Figure 8, we can see that descriptors vsa_acc, logP(o/w), SMR_VSA6, logS, weinerPol, SlogP_VSA3, TPSA, and vsa_pol are of great importance in predicting $logD_{7.4}$. Generally, these descriptors disclose information about hydrogen bond, polarity, and surface area of the molecule. Vsa_acc, the sum of van der Waals surface areas of pure hydrogen bond acceptors, shows the greatest importance in Figure 8. It contains information about the hydrogen bond of the molecule. As we all know, hydrogen bond plays a significant role in dissolution behavior. Generally speaking, a compound with a strong tendency to form hydrogen bond usually has a large solubility in the aqueous solution and a small solubility in the hydrophobic solvent. Thus, the descriptor Vsa_acc has an essential effect on $logD_{7.4}$. LogP(o/w) is the logarithm of the octanol/water partition coefficient. Several studies reveal that there is an approximate calculation formula between the logP and logD through pKa. For neutral compounds, logD is even numerically equal to logP. It is therefore no doubt that logP has a great influence on $logD_{7.4}$. Several QSAR/QSPR studies have reported that $logP/logD_{7.4}$ was usually used as an indicator to model logS (the logarithm of the aqueous solubility). And the Yalkowski equation also provides the physical basis for relating logP and solubility. There is understanding that they have a high correlationship. TPSA is the total polar surface area calculated from connection table information. Vsa_pol is the sum of van der Waals surface areas of polar atoms. Both of them reveal information regarding the polarity of the molecule. Molecules with a greater polarity are usually believed to accompany with bad dissolving in hydrophobic solvents and a good dissolvability in aqueous solutions. Moreover, SMR_VSA6, weinePol, and SlogP_VSA3 contain information on subdivided

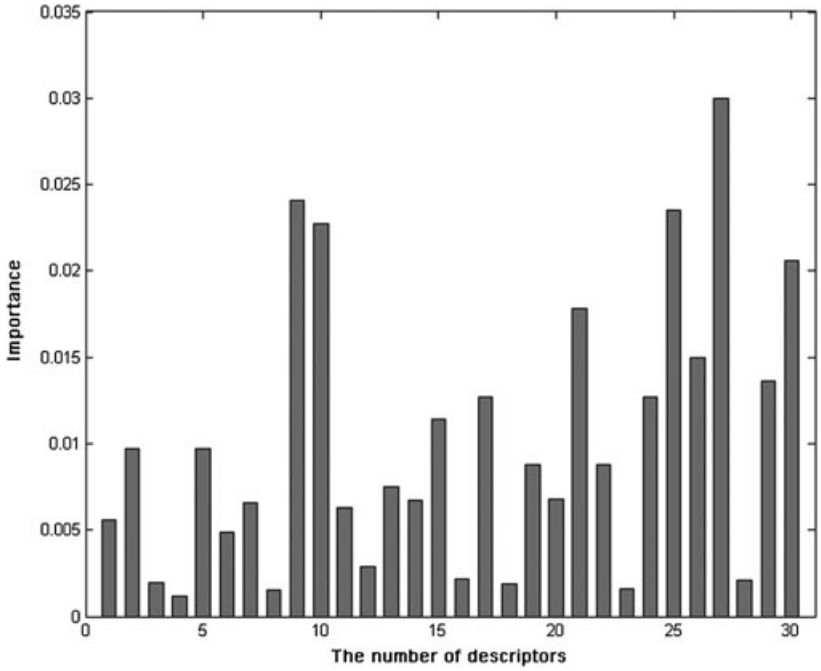

Figure 8. Variable importance of 30 molecular descriptors based on support vector machine.

surface areas of the molecule and were considered relating to molecular shape and size. Thus, these three descriptors have some effects on the logD values of compounds. Based on the analysis of the importance of descriptors used in the SVM model, we cannot only reveal the important factors influencing logD but also provide some guidance to improve the undesirable logD value of molecules to some extent.

4. CONCLUSIONS

As an important parameter in pharmacology, toxicology, and medicinal chemistry, the evaluation of $logD_{7.4}$ is of high importance in the drug discovery process. In the present study, we developed a QSPR model to reliably predict the $logD_{7.4}$ with a big and diverse data set. The comparison between SVM and PLS showed the nonlinear model derived by SVM provides better result with the same molecular descriptors. Furthermore, a series of evaluation steps such as cross validation, *Y*-randomization test, applicability domain, and the external test demonstrate the robustness and reliability of our model, strictly following the spirit of OECD principles. When compared with several calculation methods from ChemAxon and DS, the SVM model also shows superiority over them. The results indicate that the model built by SVM is reliable and has a good predictive ability. Because our proposed QSPR model was developed on the basis of theoretical descriptors calculated only from two-dimensional molecular structures, it could provide a fast, convenient, and accurate way for the evaluation of $logD_{7.4}$ of vast compounds, even for virtual compounds. Thus, this study is necessary and useful in the pharmaceutical industry, because it can save substantial amounts of time, money, and human resources.

REFERENCES

1. Renau TE, Sanchez JP, Shapiro MA, Dever JA, Gracheck SJ, Domagala JM. Effect of lipophilicity at N-1 on activity of fluoroquinolones against mycobacteria. *J. Med. Chem.* 1995; **38**: 2974–2977.
2. AL-Saadi D, Sneader W, Waton N. Relationships between lipophilic character and biological activity of new potential long-acting local anesthetics. *Med. J. Islam. Repub. Iran* 1996; **10**: 53–57.
3. Wils P, Warnery A, Phung-Ba V, Legrain S, Scherman D. High lipophilicity decreases drug transport across intestinal epithelial cells. *J. Pharmacol. Exp. Therapeut.* 1994; **269**: 654–658.
4. Waring MJ. Lipophilicity in drug discovery. *Expet Opin. Drug. Discov.* 2010; **5**: 235–248.
5. Remko M, Boháč A, Kováčiková L. Molecular structure, pKa, lipophilicity, solubility, absorption, polar surface area, and blood brain barrier penetration of some antiangiogenic agents. *J. Struct. Chem.* 2011; **22**: 635–648.
6. Arnott JA, Planey SL. The influence of lipophilicity in drug discovery and design. *Expet Opin. Drug. Discov.* 2012; **7**: 863–875.
7. Testa B, Crivori P, Reist M, Carrupt P-A. The influence of lipophilicity on the pharmacokinetic behavior of drugs: concepts and examples. *Perspect. Drug Discov. Des.* 2000; **19**: 179–211.
8. Leo A, Hansch C, Elkins D. Partition coefficients and their uses. *Chem. Rev.* 1971; **71**: 525–616.
9. NU SRR. Evaluation of the use of partition coefficients and molecular surface properties as predictors of drug absorption: a provisional biopharmaceutical classification of the list of national essential medicines of Pakistan. *Daru* 2011; **19**: 83–99.
10. Souza ES, Zaramello L, Kuhnen CA, Junkes BS, Yunes RA, Heinzen VEF. Estimating the octanol/water partition coefficient for aliphatic organic compounds using semi-empirical electrotopological index. *Int. J. Mol. Sci.* 2011; **12**: 7250–7264.
11. Cao D-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. Chemometr.* 2010; **24**: 584–595.
12. Pollastri MP. Overview on the rule of five. *Curr. Protoc.* 2010; **9**: 2.
13. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 2012; **64**: 4–17.
14. Bhal SK, Kassam K, Peirson IG, Pearl GM. The rule of five revisited: applying log D in place of log P in drug-likeness filters. *Mol. Pharm.* 2007; **4**: 556–560.
15. Comer J, Tam KY. Lipophilicity profiles: theory and measurement. In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical and Computational Strategies*, Testa B, Waterbeemd H, Folkers G, Guy R (eds.). VHCA: Zurich, 2001; 275–304.
16. Ermondi G, Lorenti M, Caron G. Contribution of ionization and lipophilicity to drug binding to albumin: a preliminary step toward biodistribution prediction. *J. Med. Chem.* 2004; **47**: 3949–3961.
17. Zhivkova Z, Doytchinova I. Quantitative structure–clearance relationships of acidic drugs. *Mol. Pharm.* 2013; **10**: 3758–3768.
18. Kah M, Brown CD. Log *D*: lipophilicity for ionisable compounds. *Chemosphere* 2008; **72**: 1401–1408.
19. Lavine BK, Workman J. Chemometrics. *Anal. Chem.* 2002; **74**: 2763–2770.
20. Xing L, Glen RC. Novel methods for the prediction of logP, pKa, and logD. *J. Chem. Inf. Comput. Sci.* 2002; **42**: 796–805.
21. Tetko IV, Bruneau P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J. Pharm. Sci.* 2004; **93**: 3103–3110.
22. Tetko IV, Poda GI. Application of ALOGPS 2.1 to predict log D distribution coefficient for Pfizer proprietary compounds. *J. Med. Chem.* 2004; **47**: 5601–5604.
23. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* 1989; **29**: 163–172.
24. Csizmadia F, Tsantili-Kakoulidou A, Panderi I, Darvas F. Prediction of distribution coefficient from structure. 1. Estimation method. *J. Pharm. Sci.* 1997; **86**: 865–871.
25. Bruneau P, McElroy NR. $logD_{7.4}$ modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *J. Chem. Inf. Model.* 2006; **46**: 1379–1387.
26. Krejsa CM, Horvath D, Rogalski SL, Penzotti JE, Mao B, Barbosa F, Migeon JC. Predicting ADME properties and side effects: the BioPrint approach. *Curr. Opin. Drug Discov. Dev.* 2003; **6**: 470–480.
27. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* 2007; **12**: 225–233.
28. Esposito EX, Hopfinger AJ, Madura JD. Methods for applying the quantitative structure–activity relationship paradigm.In *Chemoinformatics*, Springer: New York, 2004; 131–213.
29. Kennard RW, Stone LA. Computer aided design of experiments. *Technometrics* 1969; **11**: 137–148.
30. Kocjančič R, Zupan J. Modelling of the river flowrate: the influence of the training set selection. *Chemometr. Intell. Lab. Syst.* 2000; **54**: 21–34.
31. Vapnik V. *The Nature of Statistical Learning Theory*, Springer: New York, 2000.
32. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines*, Cambridge University Press Cambridge: United Kingdom, 2000.
33. Scholkopf B, Smola A. *Learning with Kernels*, MIT press Cambridge: Cambridge, 2002.
34. Liang YZ, Xu QS, Li H-D, Cao DS. *Support Vector Machines and Their Application in Chemistry and Biotechnology*, CRC Press: New York, 2011.
35. Vapnik V. *Statistical Learning Theory*, Wiley: New York, 1998.
36. Cao DS, Liang YZ, Xu QS, Hu QN, Zhang LX, Fu GH. Exploring nonlinear relationship in chemical data using kernel-based methods. *Chemometr. Intell. Lab. Syst.* 2011; **107**: 106–115.
37. Li H, Yap C, Ung C, Xue Y, Li Z, Han L, Lin H, Chen YZ. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J. Pharm. Sci.* 2007; **96**: 2838–2860.
38. Caballero J, Fernández L, Garriga M, Abreu JI, Collina S, Fernández M. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. *J. Mol. Graph. Model.* 2007; **26**: 166–178.

39. Zhao C, Zhang H, Zhang X, Liu M, Hu Z, Fan B. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* 2006; **217**: 105–119.
40. Dong X, Jiang C, Hu H, Yan J, Chen J, Hu Y. QSAR study of Akt/protein kinase B (PKB) inhibitors using support vector machine. *Eur. J. Med. Chem.* 2009; **44**: 4090–4097.
41. Clark RD, Fox PC. Statistical variation in progressive scrambling. *J. Comput. Aided Mol. Des.* 2004; **18**: 563–576.
42. Baumann K. Cross-validation as the objective function for variable-selection techniques. *TRAC-Trend Anal. Chem.* 2003; **22**: 395–406.
43. Kiralj R, Ferreira M. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J. Braz. Chem. Soc.* 2009; **20**: 770–787.
44. Yun Y-H, Cao D-S, Tan M-L, Yan J, Ren D-B, Xu Q-S, Yu L, Liang Y-Z. A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration. *Chemometr. Intell. Lab. Syst.* 2014; **130**: 76–83.
45. Leardi R. Genetic algorithms in chemometrics and chemistry: a review. *J. Chemometr.* 2001; **15**: 559–569.
46. Niazi A, Leardi R. Genetic algorithms in chemometrics. *J. Chemometr.* 2012; **26**: 345–351.
47. Goodarzi M, Heyden YV, Funar-Timofei S. Towards better understanding of feature-selection or reduction techniques for quantitative structure–activity relationship models. *TRAC-Trend Anal. Chem.* 2013; **42**: 49–63.
48. Alexander G, Alexander T. Beware of Q2. *J. Mol. Graph. Model.* 2002; **20**: 269–276.
49. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR and Combinatorial Science* 2003; **22**: 69–77.
50. Shahlaei M. Descriptor selection methods in quantitative structure–activity relationship studies: a review study. *Chem. Rev.* 2013; **113**: 8093–8103.
51. Rücker C, Rücker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* 2007; **47**(6): 2345–2357.
52. Fox J-P. Randomized item response theory models. *J. Educ. Behav. Stat.* 2005; **30**: 189–212.
53. Cao D-S, Liu S, Fan L, Liang Y-Z. QSAR analysis of the effects of OATP1B1 transporter by structurally diverse natural products using a particle swarm optimization-combined multiple linear regression approach. *Chemometr. Intell. Lab. Syst.* 2014; **130**: 84–90.
54. Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA. Current status of methods for defining the applicability domain of (quantitative) structure–activity relationships. *ATLA* 2005; **33**: 155–173.
55. Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.* 2008; **26**: 1315–1326.
56. Sahlin U, Jeliazkova N, Öberg T. Applicability domain dependent predictive uncertainty in QSAR regressions. *Mol. Inform.* 2014; **33**: 26–35.
57. 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**: 4791–4810.
58. Cao DS, Liang YZ, Xu QS, Li HD, Chen X. A new strategy of outlier detection for QSAR/QSPR. *J. Comput. Chem.* 2010; **31**: 592–602.
59. Yan J, Huang JH, He M, Lu HB, Yang R, Kong B, Xu QS, Liang YZ. Prediction of retention indices for frequently reported compounds of plant essential oils using multiple linear regression, partial least squares, and support vector machine. *J. Sep. Sci.* 2013; **36**: 2464–2471.
60. Cao DS, Liang YZ, Xu QS, Yun YH, Li HD. Toward better QSAR/QSPR modeling: simultaneous outlier detection and variable selection using distribution of model features. *J. Comput. Aided Mol. Des.* 2011; **25**: 67–80.
61. OECD-QSAR-07---ENV-JM-MONO. 2007; 2.
62. OECD-QSAR-04---ENV-JM-MONO. 2004; 24.
63. Mannhold R, Dross KP, Rekker RF. Drug lipophilicity in QSAR practice: I. A comparison of experimental with calculative approaches. *Quantitative Structure-Activity Relationships* 1990; **9**: 21–28.
64. Vrakas D, Tsantili-Kakoulidou A, Hadjipavlou-Litina D. Exploring the consistency of logP estimation for substituted coumarins. *QSAR and Combinatorial Science* 2003; **22**: 622–629.
65. Chrysanthakopoulos M, Koletsou A, Nicolaou I, Demopoulos VJ, Tsantili-Kakoulidou A. Lipophilicity studies on pyrrolyl-acetic acid derivatives. Experimental versus predicted logP values in relationship with aldose reductase inhibitory activity. *QSAR and Combinatorial Science* 2009; **28**: 551–560.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.