Prediction of Depuration Rate Constants for Polychlorinated Biphenyl Congeners

Xinliang Yu

Hunan Provincial Key Laboratory of Environmental Catalysis & Waste Regeneration, College of Materials and Chemical Engineering, Hunan Institute of Engineering, Xiangtan, Hunan 411104, China
State Key Laboratory of Chemo/Biosensing and Chemometrics, Hunan University, Changsha, Hunan 410082, China

ABSTRACT: The integral equation formalism polarizable continuum model (IEF-PCM) for solvent effects with the default solvent (water) and solvent parameters, together with the density functional theory method at 6-31G(d) level, was used to optimize molecular structures for polychlorinated biphenyl (PCB) congeners. Four molecular descriptors were selected to develop quantitative structure–activity relationship (QSAR) models for the depuration rate constants $k_d$ of 63 PCB congeners in a juvenile rainbow trout (Oncorhynchus mykiss). The optimal multiple linear regression (MLR) model has the correlation coefficient $R$ of 0.933 and the root mean square (rms) error of 0.0681 for the total set of 63 PCB congeners. The support vector regression model has $R$ of 0.953 and rms error of 0.0576 for the total set. Both the MLR and SVM QSAR models in this paper were accurate and acceptable compared with other QSAR models for the depuration rate of PCB congeners reported in references. Thus, applying IEF-PCM and B3LYP/6-31G(d) calculations for molecular descriptor derivation of PCB congeners is successful.

INTRODUCTION

Development of predictive quantitative structure–activity relationship (QSAR) models has been the concern of chemists, pharmacists, and environmental scientists. QSARs play a significant role in evaluating the biological activity, toxicity, and physicochemical property of chemicals that may not be synthetized. One of the main aims of QSAR studies is to develop a mathematical correlation between the activity of interest and molecular descriptors, which reflect the quantitative molecular structure information. A successful QSAR enables the exploration of the chemical reaction mechanism or biochemical process.

Polychlorinated biphenyls (PCBs) are serious environmental pollutants, which have the characteristics of environmental persistence, bioaccumulation and long-distance global migration. These characteristics are closely related to their physical and chemical properties. Animal experiments have shown that PCBs can induce lesions of skin, liver, gastrointestinal system, nervous system, reproductive system, and immune system and even lead to carcinomas. With the increasing attention to environmental and health issues, PCBs have become the focus of research in the field of environmental science and toxicology. There are 209 PCB congeners that possess similar molecular structures. It is unrealistic to determine the physical and chemical properties or biological activities of each PCB because of the costs involved. The QSAR can help to predict the physical and chemical properties or activities of PCBs.

Some researchers have carried out the QSAR studies of bioconcentration factors, octanol/water partition coefficients, and health effects for PCBs. Niu et al. developed a 5-descriptor QSAR model for the photodegradation half-life ($t_{1/2}$) of 22 PCBs in the n-hexane solution under UV irradiation. The ratio ($4.4 = 22/5$) of the samples (22) to descriptors (5) used is too small and lower than the minimum value of $S$.

The research group of Liu developed 2 QSAR models for the depuration rates (the logarithm of the depuration rate constants, log $k_d$) of 34 PCBs. One model was based on 5 descriptors and 26 PCBs in the training set. The other was obtained from the training set of 24 PCBs with 8 descriptors. Both the models were validated with the respective test sets. Although the sample set (34 PCBs) is relatively small compared with the descriptors used in the models, their correlation coefficients $R$ (0.941 and 0.954, respectively) are high.

Liu et al. further used 6 quantum chemical descriptors to build a QSAR model for the depuration rate constants (log $k_d$) of 62 PCBs measured at 285 K in juvenile rainbow trout (Oncorhynchus mykiss). The correlation coefficient $R$ and standard deviation are 0.853 and 0.05, respectively. The statistical results of the models are acceptable. In addition, the descriptors used have definite physical meanings.

Fatemi and Chahi developed QSAR models for the half-lives ($\log t_{1/2}$) for 62 PCBs measured at 281 K in juvenile rainbow trout. The models were based on four descriptors and statistical methods including linear and nonlinear regression.
Table 1. Experimental and Calculated Values of $k_d$ of 63 PCB Congeners

| no. | PCB congener                        | $k_d$(exp.) | $k_d$(MLR) | $k_d$(SVM) |
|-----|-------------------------------------|-------------|------------|------------|
|     | Training Set                        |             |            |            |
| 1   | 2,2',3,3',5'-pentachlorobiphenyl     | 0.3         | 0.36       | 0.36       |
| 2   | 2,2',4,4',5'-pentachlorobiphenyl     | 0.3         | 0.30       | 0.30       |
| 3   | 2,3,3',4,4'-pentachlorobiphenyl      | 0.3         | 0.37       | 0.33       |
| 4   | 2,3,3',4,4',5-pentachlorobiphenyl    | 0.3         | 0.31       | 0.31       |
| 5   | 2,2',3,4,4',5'-hexachlorobiphenyl    | 0.3         | 0.31       | 0.31       |
| 6   | 2,2',3,4,5,5'-hexachlorobiphenyl     | 0.3         | 0.38       | 0.39       |
| 7   | 2,3,3',4,5,6'-pentachlorobiphenyl    | 0.3         | 0.38       | 0.39       |
| 8   | 2,2',4,4',5,5'-hexachlorobiphenyl    | 0.3         | 0.35       | 0.38       |
| 9   | 2,4,4',3-tetrachlorobiphenyl        | 0.3         | 0.35       | 0.31       |
| 10  | 2,2',3,3',4,5,6-heptachlorobiphenyl  | 0.3         | 0.39       | 0.40       |
| 11  | 2,3,4',6-tetrachlorobiphenyl        | 0.4         | 0.44       | 0.42       |
| 12  | 2,2',3,4,5'-pentachlorobiphenyl      | 0.4         | 0.34       | 0.32       |
| 13  | 2,2',3,3',5,5,6'-hexachlorobiphenyl  | 0.4         | 0.38       | 0.39       |
| 14  | 2,2',3,3',4,4',6-heptachlorobiphenyl| 0.4         | 0.41       | 0.41       |
| 15  | 2,2',3,3',4,5,6'-heptachlorobiphenyl| 0.4         | 0.39       | 0.40       |
| 16  | 2,2',3,3',4,5,6-heptachlorobiphenyl  | 0.4         | 0.41       | 0.41       |
| 17  | 2,2',3,4,5,5,6-heptachlorobiphenyl   | 0.4         | 0.38       | 0.39       |
| 18  | 2,2',3,4,5,5,6-heptachlorobiphenyl   | 0.4         | 0.35       | 0.39       |
| 19  | 2,2',3,4,5,5,6-octachlorobiphenyl    | 0.4         | 0.43       | 0.42       |
| 20  | 2,2',3,4,5,5,6-octachlorobiphenyl    | 0.4         | 0.40       | 0.41       |
| 21  | 2,2',3,4,5,5,6'-nonachlorobiphenyl   | 0.4         | 0.47       | 0.45       |
| 22  | 2,3,4',5-tetrachlorobiphenyl        | 0.4         | 0.43       | 0.39       |
| 23  | 2,2',3,4,5,5'-pentachlorobiphenyl    | 0.4         | 0.34       | 0.37       |
| 24  | 2,2',3,3',4,4',5,5'-hexachlorobiphenyl| 0.4         | 0.34       | 0.37       |
| 25  | 2,2',3,3',4,4',5,5,6-octachlorobiphenyl| 0.4         | 0.40       | 0.41       |
| 26  | 2,2',3,3',4,5,5,5,6-octachlorobiphenyl| 0.4         | 0.40       | 0.41       |
| 27  | 2,3,4',4-tetrachlorobiphenyl        | 0.4         | 0.44       | 0.39       |
| 28  | 2,3,4',5-tetrachlorobiphenyl        | 0.4         | 0.43       | 0.40       |
| 29  | 2,2',3,4-tetrachlorobiphenyl        | 0.5         | 0.54       | 0.51       |
| 30  | 2,3,4',6-tetrachlorobiphenyl        | 0.5         | 0.44       | 0.42       |
| 31  | 2,2',3,3',6,6'-hexachlorobiphenyl   | 0.5         | 0.37       | 0.39       |
| 32  | 2,2',3,3',4,4',5,5,6-nonachlorobiphenyl| 0.5         | 0.47       | 0.45       |
| 33  | 2,2',3,3',4,4',5,5,5,6-decachlorobiphenyl| 0.5         | 0.55       | 0.49       |
| 34  | 2,2',4,6'-tetrachlorobiphenyl       | 0.5         | 0.40       | 0.47       |
| 35  | 2,3,4,4',5,5'-pentachlorobiphenyl   | 0.5         | 0.32       | 0.32       |
| 36  | 2,4',6-trichlorobiphenyl            | 0.6         | 0.69       | 0.72       |
| 37  | 2,2',3,3'-tetrachlorobiphenyl       | 0.6         | 0.77       | 0.61       |
| 38  | 2,2',3,6-tetrachlorobiphenyl        | 0.6         | 0.52       | 0.53       |
| 39  | 2,3',4-trichlorobiphenyl            | 0.7         | 0.62       | 0.71       |
| 40  | 2,2',6-trichlorobiphenyl            | 0.8         | 0.77       | 0.79       |
| 41  | 2,3',4'-trichlorobiphenyl           | 1.1         | 0.93       | 0.97       |
| 42  | 2,3'-dichlorobiphenyl               | 1.3         | 1.28       | 1.31       |
|     | Test Set                            |             |            |            |
| 43  | 2,3',4,5-tetrachlorobiphenyl        | 0.3         | 0.44       | 0.40       |
| 44  | 2,2',4,5,5'-pentachlorobiphenyl     | 0.3         | 0.33       | 0.33       |
| 45  | 2,2',3,3',4,4'-hexachlorobiphenyl   | 0.3         | 0.35       | 0.33       |
| 46  | 2,2',3,4,6-pentachlorobiphenyl      | 0.3         | 0.38       | 0.39       |
| 47  | 2,3,3',4,4',6-hexachlorobiphenyl    | 0.3         | 0.37       | 0.39       |
| 48  | 2,3,3',4,5,6-hexachlorobiphenyl     | 0.3         | 0.31       | 0.37       |
| 49  | 2,2',3,3',5'-pentachlorobiphenyl    | 0.4         | 0.30       | 0.30       |
| 50  | 2,2',3,3',4,5,5'-heptachlorobiphenyl| 0.4         | 0.39       | 0.40       |
| 51  | 2,2',3,3',4,5,6'-heptachlorobiphenyl| 0.4         | 0.35       | 0.39       |
| 52  | 2,3,3',4,4',5,5'-heptachlorobiphenyl| 0.4         | 0.38       | 0.39       |
| 53  | 2,3',3',4,4',5,5,6-octachlorobiphenyl| 0.4         | 0.40       | 0.41       |
| 54  | 2,3,3',4,4',5-pentachlorobiphenyl   | 0.4         | 0.34       | 0.37       |
| 55  | 2,2',3,3',4,5,5,6-heptachlorobiphenyl| 0.4         | 0.35       | 0.39       |
| 56  | 2,2',3,3',6-trichlorobiphenyl       | 0.4         | 0.38       | 0.39       |
| 57  | 2,2',3,4,6'-hexachlorobiphenyl      | 0.5         | 0.38       | 0.39       |
| 58  | 2,2',3,3',4,5,5,6'-nonachlorobiphenyl| 0.5         | 0.47       | 0.45       |
techniques. The coefficients of determination $R^2$ from the training set including 43 PCBs are 0.855–0.921. These models are accurate and acceptable.13

Descriptor calculation is a key step toward developing QSAR models. All these QSAR models stated above are based on the descriptors obtained from a single molecule in vacuum and gas phase.5-13 Due to the solvent effect, the geometrical structure, frontier orbital energy, and charge distribution of solute molecules in gas and solution may be different, which affect the dissolution and chemical reactions of solutions.14 In this paper, the integral equation formalism polarizable continuum model (IEF-PCM) was adopted to model the solvent effects in molecular geometry optimization. After that, Dragon 6.015 was used to calculate the molecular descriptors. Multiple linear regression (MLR) and support vector machine (SVM) were used to develop the QSAR models of depuration rate constants for PCB congeners.16

## RESULTS AND DISCUSSION

Correlation analysis of the depuration rate constants ($k_d$) of 42 PCB congeners in the training set (see Table 1) and 284 descriptors calculated with Dragon 6.015 was carried out with the MLR analysis in IBM SPSS Statistical 19. An optimal descriptor subset was obtained, which included four molecular descriptors: LOC, $SpMax_3\_Bh(m)$, $SpMin_8\_Bh(p)$, and $SpMin_5\_Bh(s)$. The MLR equation is as follows

$$k_d = 8.181 - 6.199LOC - 0.845SpMax_3\_Bh(m) + 4.447SpMin_8\_Bh(p) - 0.884SpMin_5\_Bh(s)$$

where $N$ is the number of PCB congeners, $R$ is the correlation coefficient, $rms$ is the root mean square error, and $F$ is the Fischer ratio. Descriptor characteristics and definitions are shown in Tables 2 and 3, respectively. The MLR model was used to predict the depuration rate constants $k_d$ of 21 PCB congeners in the test set. The correlation coefficient $R$ and $rms$ error are 0.922 and 0.0631, respectively. The total set of 63 PCB congeners has $R$ of 0.933 and $rms$ error of 0.0681.

An analysis of the depuration rate constants with respect to 285 descriptors (including the average quadrupole moment (Aqm) converted to $exp(0.08Aqm)$) resulted in the following regression

$$k_d = 0.120 + 7.500JGI + 6.567SpMin_8\_Bh(p) + 0.047RFDF030 + 855.242 \exp(0.08Aqm) \quad R = 0.940$$

$$R^2 = 0.884; \quad se = 0.0725; \quad F = 95.894; \quad N = 42 \quad (2)$$

Similarly, eq 2 was used for the prediction of the test set. The $rms$ errors for the training set, test set, and total set are 0.0681, 0.0703, and 0.0688, respectively. The prediction $rms$ errors from eq 2 are slightly higher than those of eq 1. Therefore, eq 1 was taken as the optimal MLR model in this paper.

The four molecular descriptors, LOC, $SpMax_3\_Bh(m)$, $SpMin_8\_Bh(p)$, and $SpMin_5\_Bh(s)$, were used as input variables to develop SVM models for the depuration rate constants ($k_d$) for 63 PCB congeners (see Table 1) by applying LibSVM package17 on the platform of MATLAB R2014a. The regularization constant $C$ and the Gaussian function parameter $γ$ were tuned with the particle swarm optimization (PSO) algorithm under the following experimental conditions: the cognition learning factor $c_1$ (1.5); the social learning factor $c_2$ (1.7); the number of particles $N_1$ (20); the maximum number of iterations $N_2$ (200); the searching ranges of both parameters $C$ and $γ$ being [1, 100]; the $N$ of the $N$-fold-cross-validation being 5; and $ε$ is the $ε$-insensitive loss function with a default value of 0.001.

The optimization results from the training set show that the optimal SVM model based on four molecular descriptors in the MLR model possesses parameters $C$ of 10.874 and $γ$ of 28.010. The correlation coefficient $R$ and $rms$ error from the training set are 0.957 and 0.0579, respectively. Subsequently, 21 PCB congeners in the test set were used to evaluate the SVM model. The correlation coefficient $R$ and $rms$ error from the test set are 0.940 and 0.0572, respectively. The $R$ and $rms$ error from the total set are 0.953 and 0.0576, respectively. Obviously, the results from the SVM model are more accurate than that from the MLR model (eq 1). Compared with the model12 ($R = 0.853$) that deals with depuration rate constants ($k_d$) at 285 K for 62 PCB congeners, both the MLR ($R = 0.937$) and the SVM model ($R = 0.957$) in this paper show acceptable statistical results.

As can be seen from Table 2, the four descriptors (LOC, $SpMax_3\_Bh(m)$, $SpMin_8\_Bh(p)$, and $SpMin_5\_Bh(s)$) have a sig-value (or $P$-value) less than 0.05, which suggest that all the descriptors are significant and related to depuration rate constants of PCBs. In addition, these descriptors possess a variance inflation factor (VIF) of less than 10, which are acceptable, since the correlation coefficients ($R^2$) between any two descriptors are less than 0.9 according to the definition of $VIF = 1/(1 - R^2)$. Thus, the four descriptors are weakly correlated with each other.

The $t$-test can be used to evaluate the significance of descriptors in MLR equations. A descriptor with a larger absolute value of $t$-test is a more significant descriptor and
makes more contribution to the dependent variable. According to the $t$-test in Table 2, the most significant descriptor appearing in eq 1 is LOC, which is a lopping centric index. It is derived from a molecular graph with H-depletion and obtained by graph pruning, which is a process of gradually removing all terminal vertices. This descriptor can reflect the degree of molecular branching: the higher the value of LOC, the more terminal vertices. This descriptor can be derived from a molecular graph with H-depletion and obtained by graph pruning, which is a process of gradually removing all terminal vertices. It is derived from a molecular graph with H-depletion and obtained by graph pruning, which is a process of gradually removing all terminal vertices.15

The three molecular descriptors, $Sp\text{Max}_3\ Bh(m)$, $Sp\text{Min}_8\ Bh(p)$, and $Sp\text{Min}_5\ Bh(s)$, belong to Burden eigenvalues. They were calculated from molecular diagrams containing hydrogen atoms by means of the Burden matrix, $Bh(w)$. This matrix in nature is a modified adjacency matrix and calculated by the following rules: the diagonal elements are atomic properties scaled by carbon atoms; the off-diagonal elements describing the characteristics pairs of bonded atoms are equal to the square root of the conventional bond order; the elements in the matrix that reflect terminal bonds are augmented by 0.1; all other entries are set to 0.001. For each matrix in the Dragon calculation, there are first eight largest positive eigenvalues $Sp\text{Max}_k\ Bh(w)$ and first eight smallest negative eigenvalues $Sp\text{Min}_k\ Bh(w)$ (absolute values). Here, $k$ is the eigenvalue rank and $w$ is the atomic property such as atomic mass ($m$), atom polarizability ($p$), and intrinsic state ($s$). These molecular descriptors can be used to solve the problem of searching for molecular structural similarity (or diversity).15

Figure 1 shows the relationship between standardized residuals (SR) and leverages and can be used to visualize the applicability domain of the SVM model.14 Predictions only for the samples that fall into this domain are considered reliable. As can be seen from Figure 1, there are three PCBs (no. 42, 2,3’-dichlorobiphenyl; no. 37, 2,2’,3,3’-tetrachlorobiphenyl; and no. 40, 2,2’,6-trichlorobiphenyl) having $h > h^*$ and SR < 3 in the test set, which indicates that the SVM model has good generalizability.

The SVM model was further evaluated with the statistical parameters ($q_{\text{int}}$, $r^*$, $k$, $R^2$, $R^2_0$, and $R^2_\text{adj}$).18–23 The value of the internal correlation coefficient $q_{\text{int}}$ obtained from the training set with leave-one-out cross-validation procedure is 0.690 (>0.5); the value of the correlation coefficient $r^*$ for the test set is 0.883 (>0.6); the slopes $k$ and $k'$ of the regression lines through the origin are 0.983 and 1.003, respectively, within the range of 0.85–1.15; the values of the determination coefficients $R^2$ (0.845) and $R^2_0$ (0.875) are close to those of $r^2$ (0.883) with $(r^2 - R^2_0)/r^2 = 0.044$ and $(r^2 - R^2_0)/r^2 = 0.009$, both less than 0.1; and the value of the correlation coefficient $r^*_0$ (without an intercept for $r^*$) is 0.878 and that of $r^*_0 = r^* \times [1 - (r^2 - r^*_0)^2/2]$ is 0.820 (>0.5). These statistical parameters satisfy the criteria defined for an acceptable QSPR model.19–21 Therefore, the SVM model in this paper is successful.

**CONCLUSIONS**

Four molecular descriptors were successfully used to develop the QSAR models for depuration rate constants ($k_d$) of 63 PCB congeners in juvenile rainbow trout. IEF-PCM and B3LYP/6-31G(d) calculation, with the default solvent (water) and solvent parameters, was used for descriptor derivation. The branching degree (LOC), molecular similarity indices based on atomic mass ($Sp\text{Max}_3\ Bh(m)$), atom polarizability ($Sp\text{Min}_8\ Bh(p)$), and intrinsic state ($Sp\text{Min}_5\ Bh(s)$) can reflect the structure factors affecting the depuration rate constants ($k_d$) of PCBs. Both the MLR and SVM models were tested to be accurate and acceptable with correlation coefficients of 0.933 and 0.953, respectively. It is feasible to develop QSAR models for depuration rate constants ($k_d$) of PCB congeners by applying IEF-PCM and B3LYP/6-31G(d) calculations for molecular descriptor derivation. This result encourages the further application of QSAR to other properties, such as photodegradation half-life and biological activities of PCBs.

**MATERIALS AND METHODS**

The experimental data. The experimental values of depuration rate constants ($k_d$) for 63 PCB congeners in juvenile rainbow trout ($O. mykiss$) were taken from the literature16 and are listed in Table 1. Sixty-three PCB congeners were randomly split into a training set (42 PCBs) and a test set (21 PCBs). The former was used to train QSAR models, and the latter was used to validate the models. In the experiment,16 these fishes were exposed to a high-dose treatment of PCBs and the $k_d$ values measured at 281 K ranged from 0.3 to 1.3.
Because the PCB congeners possess similar molecular structures, the range of the k_{ij} values is relatively narrow. This phenomenon can be found in other properties.\textsuperscript{12,13}  

Therefore, the QSAR models based on PCBs can only be used for the prediction of the properties of PCBs.  

**Structural Descriptor Calculations.** The molecular structures of PCB congeners were sketched by ChemBioDraw Ultra 11.0 and preliminarily optimized by ChemBio3D Ultra 11.0 in ChemOffice 2008 with molecular mechanics (MM2) method with the default convergence criterion. The energy-minimized molecules were further optimized by applying IEF-PCM/B3LYP/6-31G(d) in Gaussian 09 program, with the default solvent (water) and solvent parameters. There are two reasons for carrying out the IEF-PCM and B3LYP/6-31G(d) calculations for the molecular descriptor derivation of PCB congeners. One is that this approach can be used to optimize the molecular structures of solute at a reasonable level, which has been investigated by our previous work.\textsuperscript{14} The other is the fact that this approach can provide more structural information and molecular descriptors than other semiempirical quantum chemical calculations such as the AM1 and PM6. For example, IEF-PCM and B3LYP/6-31G(d) can be used for the calculation of the average quadrupole moment (Aqm), although these descriptors may not be present in models.  

Besides the average quadrupole moment, Dragon 6.0 was used for the molecular descriptor derivation after full geometry optimization. Redundant and nonuseful structure descriptors that are equal to a constant or have a pair of correlation coefficients \(\geq 0.90\) were deleted from the 4885 descriptors calculated. In total, 284 descriptors were obtained with Dragon 6.0.\textsuperscript{15}  

**SVM Principle.** The SVM algorithm has many attractive characteristics and advantages. Unlike the traditional theory, the SVM adopts the principle of structural risk minimization, which makes it possible for the SVM to have a good prediction ability even with fewer training samples. The SVM was originally used in pattern recognition. With the introduction of insensitive loss function, the application of SVM extends to the prediction of properties in nonlinear regression models and the evaluation in a time-series analysis.  

The main idea of the SVM used in the regression prediction is to map the input characteristic parameters into a high-dimensional feature space through a nonlinear mapping function and then carry out a linear regression analysis. The SVM algorithm is based on the following regression\textsuperscript{3,4,24}

\[
f(x) = \sum_{i=1}^{n} \phi(x_i)w + b
\]

where \(n\) is the number of samples used in the training set, \(\phi(x_i)\) is the mapping function, \(x\) is the input vector composed of characteristic parameters reflecting molecular structure, and \(f(x)\) is the output of prediction results. The coefficients \(w\) and \(b\) can be estimated by means of the following minimization

\[
\min_{w,\xi, \xi^*} J(w, \xi, \xi^*) = \frac{1}{2} ||w||^2 + C \sum (\xi_i + \xi_i^*)
\]

subject to

\[
\begin{align*}
\gamma_i - \phi^T(x_i)w - b \leq \epsilon + \xi_i \\
\phi^T(x_i)w + b - \gamma_i \leq \epsilon + \xi_i^*
\end{align*}
\]

where \(C\) is a penalty parameter used for adjusting the training error. \(\epsilon\) is a prescribed parameter of the \(\epsilon\)-insensitive loss function. \(\xi^*\) and \(\xi^*\) are the slack variables. In general, support vector regression is based on the \(\epsilon\)-insensitive loss function to minimize the training error

\[
f(x) - y \leq \epsilon \quad (f(x) - y_l \geq \epsilon)
\]

Thus, eq 3 becomes

\[
f(x) = \sum_{i=1}^{n} (a_i - a_i^*)\phi(x_i)\phi(x) + b
\]

where \(a_i\) and \(a_i^*\) are the Lagrange multipliers that are used to solve the quadratic optimization problem. Equation 9 can be obtained when the kernel function \(k(x_i, y_j)\) is introduced

\[
f(x) = \sum_{i=1}^{n} (a_i - a_i^*)K(x_i, y_j) + b
\]

where \(s\) is equal to the number of samples whose \(a_i^*\) and \(a_i^*\) values are above 0. In the present work, the following Gaussian radial basis function was used

\[
K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)
\]

The kernel width \(\gamma\) affects the prediction performance of the SVM models. In this paper, the SVM parameters \(C\) and \(\gamma\) are tuned with the particle swarm optimization (PSO) algorithm.\textsuperscript{25-28}

**AUTHOR INFORMATION**

**Corresponding Author**  
E-mail: yxl@hnie.edu.cn. Phone: +86 731 58680049. Fax: +86 731 68680125.

**ORCID**  
Xinliang Yu: 0000-0002-0952-3567

**Notes**  
The author declares no competing financial interest.

**ACKNOWLEDGMENTS**

This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (No. 16A047), the Open Project Program of State Key Laboratory of Chemo/Biosensing and Chemometrics (Hunan University) (No. 2016013), and the Open Project Program of Hunan Provincial Key Laboratory of Environmental Catalysis & Waste Regeneration (Hunan Institute of Engineering) (No. 2018KF11).

**REFERENCES**

(1) Negrito, M. C.; Valdez, C.; Sharma, J.; Rosenberg, C.; Selassi, C. R. Growth Inhibition and DNA Damage Induced by X-Phenols in Yeast: A Quantitative Structure–Activity Relationship Study. ACS Omega 2017, 2, 8568–8579.

(2) Yu, X.; Xu, L.; Zhu, Y.; Lu, S.; Dang, L. Correlation between 13C NMR chemical shifts and complete sets of descriptors of natural coumarin derivatives. Chemom. Intell. Lab. Syst. 2019, 184, 167–174.

(3) Speck-Planche, A. Multicellular Target QSAR Model for Simultaneous Prediction and Design of Anti-Pancreatic Cancer Agents. ACS Omega 2019, 4, 3122–3132.

(4) Yu, X.; Yu, R.; Tang, L.; Guo, Q.; Zhang, Y.; Zhou, Y.; Yang, Q.; He, X.; Yang, X.; Wang, K. Recognition of candidate aptamer sequences for human hepatocellular carcinoma in SELEX screening.

(5) This work was supported by the Scientific Research Fund of Hunan Provincial Education Department (No. 16A047), the Open Project Program of State Key Laboratory of Chemo/Biosensing and Chemometrics (Hunan University) (No. 2016013), and the Open Project Program of Hunan Provincial Key Laboratory of Environmental Catalysis & Waste Regeneration (Hunan Institute of Engineering) (No. 2018KF11).
using structure–activity relationships. Chemom. Intell. Lab. Syst. 2014, 136, 10–14.
(5) van der Oost, R.; Beyer, J.; Vermeulen, N. P. E. Fish bioaccumulation and biomarkers in environmental risk assessment: A review. Environ. Toxicol. Pharmacol. 2003, 13, 57–149.
(6) Qin, L. T.; Liu, S. S.; Liu, H. L.; Ge, H. L. A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV). Chemosphere 2008, 70, 1577–1587.
(7) Han, X.-Y.; Wang, Z.-Y.; Zhai, Z.-C.; Wang, L.-S. Estimation of the n-octanol/water Partition Coefficients (Kow) of all PCB Congeners by Ab initio and a CI Substitution Position Method. QSAR Comb. Sci. 2006, 25, 333–341.
(8) Ruiz, P.; Faroon, O.; Moudgal, C. J.; Hansen, H.; De Rosa, C. T.; Muntau, M. Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure–activity relationship (QSAR). Toxicol. Lett. 2008, 181, 53–65.
(9) Niu, J. F.; Yang, Z. F.; Shen, Z. Y.; Wang, L. L. QSRRs for the prediction of photodegradation half-life of PCBs in n-hexane. SAR QSAR Environ. Res. 2006, 17, 173–182.
(10) Xu, M.; Liu, X.; Wang, L.; Wu, D.; Yang, Z.; Cui, B. Quantitative structure-activity relationship for the depuration rate constants of polychlorinated biphenyls in the freshwater mussel Elliptio complanata. J. Environ. Sci. Health, Part B 2009, 44, 278–283.
(11) Wang, L.; Liu, X.; Shan, Z.; Shi, L. Using electrotopological state indices to model the depuration rates of polychlorinated biphenyls in mussels of Elliptio complanata. J. Environ. Sci. 2010, 22, 1544–1550.
(12) Wang, L.; Liu, X. H.; Wu, D.; Xu, M. Z.; Sun, T.; Cui, B. S.; Yang, Z. F. Modelling the depuration rates of polychlorinated biphenyls in Oncorhyncus mykiss with quantum chemical descriptors. SAR QSAR Environ. Res. 2009, 20, 91–101.
(13) Fatemi, M. H.; Ghareh chahi, Z. QSRR-based estimation of the half-lives for polychlorinated biphenyl congeners. SAR QSAR Environ. Res. 2012, 23, 155–168.
(14) Yu, X.; Yu, R. Setschenow constant prediction based on the IEF-PCM Calculations. Ind. Eng. Chem. Res. 2013, 52, 11182–11188.
(15) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. DRAGON Software for the Calculation of Molecular Descriptors, revision 6.0 for Windows; Talete s.r.l.: Milan, 2010.
(16) Buckman, A. H.; Brown, S. B.; Hoekstra, P. F.; Fisk, A. T. Toxicokinetics of three polychlorinated biphenyl technical mixtures in rainbow trout (Oncorhyncus mykiss). Environ. Toxicol. Chem. 2004, 23, 1725–1736.
(17) Chang, C. C.; Lin, C. J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011, 2, No. 27.
(18) Gharaghani, S.; Khayamian, T.; Ebrahimi, M. Molecular dynamics simulation study and molecular docking descriptors in structure-based QSAR on acetylcholinesterase (AChE) inhibitors. SAR QSAR Environ. Res. 2013, 24, 773–794.
(19) 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, 69–77.
(20) Golbraikh, A.; Tropsha, A. Beware of q2! J. Mol. Graphics Modell. 2002, 20, 269–276.
(21) Roy, P. P.; Roy, K. On some aspects of variable selection for partial least squares regression models. QSAR Comb. Sci. 2008, 27, 302–313.
(22) Yu, X.; Yi, B.; Yu, W.; Wang, X. DFT-based quantum theory QSPR studies of molar heat capacity and molar polarization of vinyl polymers. Chem. Pap. 2008, 62, 623–629.
(23) Yu, X.; Yang, H.; Huang, X. Novel Method for Structure–Activity Relationship of Aptamer Sequences for Human Prostate Cancer. ACS Omega 2018, 3, 10002–10007.
(24) Yu, X. L.; Deng, J. Y.; Chen, J. F.; Yang, H. Q. Prediction of 13C NMR Chemical Shifts of Quinolone Derivatives Based on DFT Calculations. J. Struct. Chem. 2019, 60, 772–779.