Research Article

Estimation of *n*-Octanol/Water Partition Coefficients (logK_{OW}) of Polychlorinated Biphenyls by Using Quantum Chemical Descriptors and Partial Least Squares

Fang-Li Zhang,[1] Xing-Jian Yang,[1] Xiu-Ling Xue,[2] Xue-Qin Tao,[3] Gui-Ning Lu,[1,4] and Zhi Dang[1,5]

[1] *School of Environment and Energy, South China University of Technology, Guangzhou 510006, China*

[2] *College of Chemical Engineering, Huaqiao University, Xiamen 361021, China*

[3] *School of Environmental Science and Engineering, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China*

[4] *Guangdong Provincial Key Laboratory of Atmospheric Environment and Pollution Control, Guangzhou 510006, China*

[5] *The Key Lab of Pollution Control and Ecosystem Restoration in Industry Clusters, Ministry of Education, Guangzhou 510006, China*

Correspondence should be addressed to Gui-Ning Lu; lgn36@qq.com

Received 15 May 2013; Revised 29 July 2013; Accepted 22 August 2013

Academic Editor: Enrico Veschetti

The *n*-octanol/water partition coefficient (log K_{OW}) is a useful parameter for the assessment of the environmental fate and impact of xenobiotic trace contaminants. Quantitative structure-activity relationship (QSAR) model for log K_{OW} of polychlorinated biphenyls (PCBs) was analyzed by using the density functional theory at B3LYP/6-31G(d) level and the partial least squares (PLS) method with an optimizing procedure. A PLS model with reasonably good coefficient ($R^2 = 0.992$) and cross-validation test ($Q^2_{cum} = 0.988$) values was obtained. All the predicted values are within the range of ±0.3 log unit from the observed values. The log K_{OW} values of 7 PCBs in the test set predicted by the model are very close to those observed, indicating that this model has high fitting precision and good predictability. The PLS analysis showed that PCBs with larger electronic spatial extent and lower molecular total energy values tend to be more hydrophobic and lipophilic.

1. Introduction

Polychlorinated biphenyls (PCBs) are highly stable industrial chemical products that have been used as industrial fluids, flame retardants, diluents, hydraulic fluids, and dielectric fluids. Due to careless disposal practices and chemical stability, PCBs are some of the most prevalent pollutants in the total environment. Generally, a sum of 209 theoretical PCBs exist, and approximately 150 PCBs are found in the environment [1].

Despite the fact that the production and the use of PCBs have long been banned, they are still considered environmental contaminants as they are among the most widespread and persistent man-made products in the ecosystem. It is well known that these chemicals are toxic and lipophilic and

tend to be bioaccumulated. Some interfere with mammalian and avian reproduction, and others disturb human's germ cell development or are promoters of cancer. Some PCBs are included in the increasing list of environmental pollutants that are able to mimic estrogen action and are termed "environmental estrogens" [2].

PCBs are categorized as persistent organic pollutants (POPs) [3]. Individual PCB congeners exhibit different physicochemical properties and biological activities that result in different environmental distributions and toxicity profiles. The variable composition of PCB residues in environmental matrices and their different mechanisms of toxicity complicate the development of scientifically based regulations for the risk assessment [4]. The *n*-octanol/water distribution coefficient (K_{OW}) is an important physicochemical

property suitable for the characterization of the lipophilicity of the compounds such as PCBs. K_{OW} has been shown to correlate with aquatic solubility, toxicity, and bioaccumulation [5–7], and as such it is a useful parameter for the assessment of the environmental fate and impact of xenobiotic trace contaminants [8]. During the last 50 years, K_{OW} has proven to be a valuable tool in many areas of natural sciences: chemistry, biochemistry, biology, pharmacology, environmental sciences, and so forth. During that period, many hundreds or even thousands of scientific publications have been made in which the successful application of K_{OW} in correlation with many physical, chemical, or biological properties has been demonstrated for a large variety of organic chemicals. Unfortunately, many correlations have been also published in which the processes studied have nothing in common with the process of partition between water and n-octanol phases. It must be emphasized that in order to correlate, two properties do not have to be related; both of them may simply be related to a third property. The shake flask is the classical method for measuring K_{OW}, and this procedure will give reproducible and accurate results for chemicals with log K_{OW}. Accurate results can also be obtained for more hydrophobic chemicals, but this will need a very precise and skillful handling and the use of very pure n-octanol. The problem with this procedure is that the complete phase separation is difficult to achieve, and this may lead to the underestimation of K_{OW}, particularly for the more hydrophobic chemicals [9].

It is impractical to measure the K_{OW} of all the PCBs directly in the laboratory because of the extremely low water solubility of some high-chlorinated PCBs. The lack of complete, reliable, and comparable data has led to the development of different K_{OW} estimation methods. With the advent of inexpensive and rapid computation, there has been a remarkable growth in the area of quantitative structure-activity relationships (QSAR), which correlate the properties of pollutants with relevant properties and molecular descriptors [11]. A large number of calculation methods have been presently developed for estimation of the partition coefficients with varying success and applicability. These methods are cable of predicting environmental behaviors not only for common environmental pollutants but also for those chemicals that have not yet been synthesized or those that cannot be examined experimentally due to their extremely hazardous nature. According to the descriptors used, these methods can be classified into two groups: empirical and theoretical methods [12]. Hansen et al. [1] used the degree of chlorination in the orthoposition of the PCB molecule as predictor variable together with either the calculated molecular total surface area (TSA) or the gas chromatographic relative retention times (RRT) to predict the log K_{OW} and soil sorption coefficients (log K_{OC}) for all 209 PCB congeners which were based on 53 and 48 experimental data points, respectively. Also, reporting the relationship between K_{OW} and K_{OC}, which is promising for K_{OW}, could proceed with K_{OC}. However, TSA and RRT are both empirical molecular descriptors. Firstly, the accurate determination of empirical molecular descriptors may be difficult and expensive in terms of cost and time or even impossible for some PCBs which might have been not synthesized or purified. Meanwhile, the

experimental errors may be introduced, especially for those congeners difficult to be separated and identified by chromatography. Furthermore, the empirical molecular descriptors often reflect complex and multiple physical interactions. Hence, the interpretation of the respective property could be difficult and ambiguous [13]. Many attempts have been made to model the physicochemical properties and bioactivities of PCBs by utilizing QSAR techniques [14–21]. For instance, Qin et al. [17] reported the correlation of bioconcentration factors and molecular electronegativity distance vector (MEDV) of PCBs based on molecular structure. Tuppurainen and Ruuskanen [18] introduced a new QSAR descriptor electronic eigenvalue (EEVA) based on molecular orbital energies to predict the Ah receptor binding affinity of PCBs, polychlorinated dibenzo-p-dioxins, and dibenzofurans (PCDD/Fs). Daren [19] reported QSPR studies of K_{OW}, S_W, Y_W, and K_{OC} of all 209 PCBs and biphenyl by the combination of genetic algorithms and PLS analysis.

Modern theoretical method in quantum chemistry such as density functional theory (DFT) with high calculation precision should have its advantages in estimating the properties of environmental toxics such as PCBs, PCDD/Fs [13, 22]. The aim of the present study is to explore QSAR model relating to log K_{OW} of 27 PCBs by using partial least squares (PLS) analysis with optimizing procedure, based on reported log K_{OW} data and quantum chemical descriptors computed by DFT contained in Gaussian 03, in an attempt to develop reliable and predictive QSAR model and identify the primary molecular factors governing the K_{OW} of PCBs.

2. Materials and Methods

2.1. Target PCBs. A total of 27 PCBs, whose K_{OW} and/or log K_{OW} were previously published, were chosen to constitute the training set and the test set in this work. The training set consists of 20 of these 27 PCBs selected randomly, and the remaining 7 PCBs constitute the test set. Their chemical abstracts service numbers (CAS no.) and reported log K_{OW} are listed in Table 1.

2.2. Calculation and Selection of the Descriptors. The molecular modeling system HyperChem (Release 7.0, Hypercube Inc. 2002) was used to construct and view all molecular structures. Molecular geometry was optimized, and quantum chemical descriptors were computed using the B3LYP hybrid functional of DFT in conjunction with 6-31G(d), a split-valence basis set with polarization function. The B3LYP calculations were performed using the quantum chemical computation software Gaussian 03 [23]. All calculations were run on an Intel Core i5-2410M CPU/2.30 GHz computer equipped with 2.00 GB of internal memory and Microsoft Windows 7 professional operating system.

According to the chemometrics theory, it is suggested that a QSAR model should include as many relevant descriptors as possible to increase the probability of a good characterization for a class of compounds [24]. In this study, 18 independent quantum chemical variables were selected for developing QSAR models. The descriptors cover the dihedral angle of

TABLE 1: *n*-Octanol/water partition coefficients of the PCBs studied.

No.[a]	Compounds	CAS no.	Observed log K_{OW} [b]	Predicted log K_{OW}		
				Reference [10]	Model **I**	Model **II**
1	Biphenyl	92-52-4	4.10	4.14	3.96	3.87
2	2-Chlorobiphenyl	2051-60-7	4.56	4.64	4.56	4.56
3	3-Chlorobiphenyl	2051-61-8	4.72	4.63	4.77	4.77
4	2,2′-Dichlorobiphenyl	13029-08-8	5.02	5.20	5.03	5.05
5	2,4-Dichlorobiphenyl	33284-50-3	5.15	5.16	5.05	5.06
6	2,2′,5-Trichlorobiphenyl	37680-65-2	5.64	5.73	5.57	5.59
7	2,4,4′-Trichlorobiphenyl	7012-37-5	5.74	5.73	5.82	5.80
8	2,4,5-Trichlorobiphenyl	15862-07-4	5.77	5.76	5.57	5.60
9	2,3′,4,5-Tetrachlorobiphenyl	73557-53-8	6.39	6.30	6.58	6.58
10	3,3′,4,4′-Tetrachlorobiphenyl	32598-13-3	6.52	6.39	6.68	6.71
11	2,2′,3,4,5′-Pentachlorobiphenyl	38380-02-8	6.85	6.84	6.85	6.91
12	2,2′,4,5,5′-Pentachlorobiphenyl	37680-73-2	6.85	6.80	6.85	6.86
13	2,3,4,5,6-Pentachlorobiphenyl	18259-05-7	6.85	6.86	6.85	6.88
14	2,2′,3,3′,4,4′-Hexachlorobiphenyl	38380-07-3	7.44	7.43	7.45	7.53
15	2,2′,4,4′,5,5′-Hexachlorobiphenyl	35065-27-1	7.44	7.37	7.48	7.43
16	2,2′,4,4′,6,6′-Hexachlorobiphenyl	33979-03-2	7.12	7.39	7.41	7.37
17	2,2′,3,3′,4,4′,5,5′-Octachlorobiphenyl	35694-08-7	8.68	8.60	8.60	8.59
18	2,2′,3,3′,5,5′,6,6′-Octachlorobiphenyl	2136-99-4	8.42	8.50	8.58	8.54
19	2,2′,3,3′,4,4′,5,5′,6-Nonachlorobiphenyl	40186-72-9	9.14	9.08	8.90	8.90
20	2,2′,3,3′,4,4′,5,5′,6,6′-Decachlorobiphenyl	2051-24-3	9.60	9.72	9.34	9.30
21	4-Chlorobiphenyl	2051-62-9	4.69	4.69	4.50	4.51
22	2,5-Dichlorobiphenyl	34883-39-1	5.18	5.16	5.15	5.11
23	4,4′-Dichlorobiphenyl	2050-68-2	5.28	5.29	5.34	5.25
24	3,4,4′-Trichlorobiphenyl	38444-90-5	5.90	5.77	5.98	5.97
25	2,2′,5,5′-Tetrachlorobiphenyl	35693-99-3	6.26	6.29	6.39	6.35
26	2,3′,4′,5-Tetrachlorobiphenyl	32598-11-1	6.39	6.31	6.53	6.57
27	2,2′,3,4′,5,5′,6-Heptachlorobiphenyl	52663-68-0	7.93	7.95	8.04	8.05

[a]Compounds no. 1–20 constitute the training set and compounds No. 21–27 constitute the test set; [b]from [10].

the two benzene ring plans (DA), the eigenvalue of the highest occupied molecular orbital (E_{HOMO}), the eigenvalue of the lowest unoccupied molecular orbital (E_{LUMO}), the Mulliken atomic charges on 12 carbon atoms from C1 to C12 (Q_{C1}–Q_{C12}, Figure 1 shows the numbering of selected atoms on the molecular structure), the electronic spatial extent (Re), the dipole moment (μ), and molecular total energy (TE). All the above descriptors were obtained directly in the output files of Gaussian 03 calculation through a single full optimizing process for the molecular structure: B3LYP/6-31G(d) FOPT. The values of these quantum chemical descriptors are listed in Table 2. The units of dihedral angle, atomic charge, electronic spatial extent, dipole moment, and energy are as follows: degree, atomic charge unit, atom unit, Debye, and Hartree, respectively.

2.3. Modeling Method and Evaluation Indexes.

Since a big number of descriptors were selected in this study, intercorrelation of independent variables (multicollinearity) might become a technical problem. It was especially difficult when the number of independent variables was equal to or greater than the number of compounds selected in the
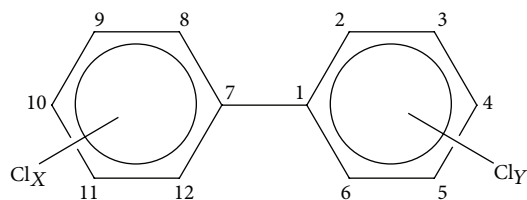

FIGURE 1: Numbering of selected atoms in the molecular structure. X and Y are integers between 0 and 5, inclusive. The positions 7, 8, 9, 10, 11, and 12 are corresponding to $1'$, $2'$, $3'$, $4'$, $5'$, and $6'$, respectively.

training set. The common regression analysis frequently used in QSAR studies became ineffective because overfitting readily occurred. To overcome these problems, we used the PLS regression, a well-known multivariate method which gives a stepwise solution for a regression model. It extracts principal component-like latent variables from original independent variables (predictor variables) and dependent variables (response variables), respectively [24]. This method can analyze data which are strongly collinear, and more importantly, it is a remedy to overfitting. PLS finds

TABLE 2: Quantum chemical descriptors for the PCBs studied.

No.	DA	E_{HOMO}	E_{LUMO}	Q_{C1}	Q_{C2}	Q_{C3}	Q_{C4}	Q_{C5}	Q_{C6}	Q_{C7}	Q_{C8}	Q_{C9}	Q_{C10}	Q_{C11}	Q_{C12}	Re	μ	TE
1	0.00	−0.21730	−0.03300	0.10126	−0.17937	−0.13395	−0.12487	−0.13395	−0.17937	0.10126	−0.17937	−0.13395	−0.12487	−0.13395	−0.17937	2336.596	0.0000	−463.3027
2	56.13	−0.23323	−0.02747	0.07290	−0.11564	−0.11564	−0.12189	−0.13101	−0.16089	0.06717	−0.14793	−0.13548	−0.12442	−0.13374	−0.16461	2764.356	1.6832	−922.8973
3	0.00	−0.22671	−0.04365	0.10349	−0.18053	−0.06979	−0.12783	−0.13333	−0.17359	0.10035	−0.17756	−0.13421	−0.12302	−0.13374	−0.17846	3399.230	2.1220	−922.8989
4	110.08	−0.24310	−0.02654	0.05227	−0.10265	−0.13685	−0.11926	−0.13195	−0.13566	0.05227	−0.10265	−0.13685	−0.11926	−0.13195	−0.13566	3242.151	2.0016	−1382.4899
5	55.59	−0.23637	−0.03739	0.07579	−0.11304	−0.13680	−0.05946	−0.13118	−0.16007	0.06673	−0.14837	−0.13517	−0.12363	−0.13361	−0.16490	4150.644	2.1687	−1382.4921
6	81.21	−0.24564	−0.03220	0.04787	−0.08812	−0.13561	−0.11931	−0.06799	−0.14376	0.04642	−0.09074	−0.13702	−0.11796	−0.13062	−0.14315	4428.552	1.8571	−1842.0844
7	55.36	−0.23937	−0.04578	0.07547	−0.11313	−0.13653	−0.05901	−0.13102	−0.15988	0.06733	−0.14482	−0.13587	−0.06134	−0.13435	−0.16175	6126.716	1.4297	−1842.0878
8	56.16	−0.24156	−0.04537	0.07940	−0.10817	−0.13619	−0.06867	−0.07657	−0.16238	0.06469	−0.14763	−0.13494	−0.12251	−0.13332	−0.16366	5070.323	2.2621	−1842.0820
9	126.87	−0.24760	−0.05378	0.07904	−0.10835	−0.13577	−0.06785	−0.07710	−0.16166	0.06579	−0.16413	−0.07033	−0.12363	−0.13364	−0.14285	6700.831	1.9485	−2301.6772
10	0.00	−0.23831	−0.06630	0.10963	−0.17838	−0.07907	−0.07446	−0.13347	−0.16970	0.10963	−0.17838	−0.07907	−0.07446	−0.13347	−0.16970	8012.630	2.6716	−2301.6757
11	83.60	−0.25298	−0.04434	0.05232	−0.09797	−0.09264	−0.06419	−0.12804	−0.13433	0.04160	−0.08793	−0.13512	−0.11822	−0.06856	−0.14180	7181.162	2.6402	−2761.2634
12	80.64	−0.25360	−0.04744	0.04992	−0.08328	−0.13748	−0.06497	−0.07717	−0.14251	0.04568	−0.08834	−0.13512	−0.11755	−0.06869	−0.14277	7317.571	1.5120	−2761.2680
13	89.95	−0.25762	−0.04957	0.06408	−0.09995	−0.08788	−0.07241	−0.08788	−0.09995	0.02910	−0.13369	−0.13182	−0.12112	−0.13182	−0.13369	6348.796	2.1628	−2761.2491
14	85.65	−0.26235	−0.04741	0.04784	−0.09803	−0.09259	−0.06444	−0.12802	−0.13388	0.04784	−0.09803	−0.09259	−0.06444	−0.12802	−0.13388	8880.956	3.1952	−3220.8481
15	81.14	−0.25785	−0.05272	0.04819	−0.08337	−0.13735	−0.06481	−0.07739	−0.14173	0.04819	−0.08337	−0.13735	−0.06481	−0.07739	−0.14173	9135.369	0.0278	−3220.8572
16	90.00	−0.26706	−0.04525	0.04104	−0.07511	−0.13762	−0.05110	−0.13762	−0.07511	0.04104	−0.07511	−0.13762	−0.05110	−0.13762	−0.07511	7602.549	0.0000	−3220.8597
17	85.88	−0.26513	−0.05927	0.04910	−0.09391	−0.09064	−0.07515	−0.07354	−0.13392	0.04910	−0.09391	−0.09064	−0.07515	−0.07354	−0.13392	11582.656	1.1648	−4140.0242
18	89.98	−0.26008	−0.05592	0.03906	−0.08305	−0.07831	−0.11437	−0.07831	−0.08305	0.03906	−0.08305	−0.07831	−0.11437	−0.07831	−0.08305	9319.431	0.0000	−4140.0277
19	90.83	−0.26526	−0.06331	0.05601	−0.09023	−0.08906	−0.06972	−0.08907	−0.09023	0.03596	−0.08464	−0.09111	−0.07435	−0.07447	−0.11978	12083.969	1.0110	−4599.6086
20	90.01	−0.26670	−0.06373	0.04327	−0.07929	−0.08962	−0.06896	−0.08962	−0.07929	0.04327	−0.07929	−0.08962	−0.06896	−0.08962	−0.07929	12579.593	0.0000	−5059.1933
21	0.00	−0.22180	−0.04264	0.10353	−0.17652	−0.13541	−0.06248	−0.13541	−0.17652	0.10345	−0.17931	−0.13367	−0.12379	−0.13367	−0.17931	3744.110	2.1391	−922.8990
22	55.89	−0.23832	−0.03813	0.07487	−0.11298	−0.13394	−0.12217	−0.06823	−0.16185	0.06624	−0.14755	−0.13515	−0.12312	−0.13355	−0.16419	3923.000	0.7279	−1382.4922
23	0.00	−0.22591	−0.05162	0.10551	−0.17639	−0.13505	−0.06209	−0.13505	−0.17639	0.10551	−0.17639	−0.13505	−0.06209	−0.13505	−0.17639	5691.897	0.0000	−1382.4950
24	0.00	−0.23229	−0.05925	0.11241	−0.18061	−0.07848	−0.07511	−0.13380	−0.17067	0.10325	−0.17454	−0.13518	−0.06122	−0.13481	−0.17556	6796.892	1.5815	−1842.0855
25	82.68	−0.25010	−0.03958	0.04559	−0.08791	−0.13527	−0.11819	−0.06836	−0.14172	0.04559	−0.08791	−0.13527	−0.11819	−0.06836	−0.14172	5819.703	0.2171	−2301.6786
26	55.56	−0.24687	−0.05342	0.07462	−0.11306	−0.13333	−0.12008	−0.06896	−0.16122	0.07015	−0.14507	−0.08147	−0.07164	−0.13241	−0.15566	6827.221	2.3961	−2301.6779
27	90.59	−0.25929	−0.05534	0.05332	−0.09406	−0.07793	−0.11522	−0.07793	−0.09406	0.03168	−0.07455	−0.13741	−0.06394	−0.07807	−0.12514	9224.954	1.0323	−3680.4424

TABLE 3: Fitting results for PLS model **I** and model **II**.

Model	Y	X	h	R^2_X	$R^2_{X(cum)}$	R^2_Y	$R^2_{Y(cum)}$	Eig	Q^2	Q^2_{cum}
I	$\log K_{OW}$	$X_1{}^a$	1	0.519	0.519	0.928	0.928	9.35	0.922	0.922
			2	0.172	0.691	0.065	0.992	3.09	0.841	0.988
II	$\log K_{OW}$	$X_2{}^b$	1	0.542	0.542	0.928	0.928	9.21	0.921	0.921
			2	0.181	0.723	0.063	0.992	3.08	0.830	0.987

$^a X_1$ containing all 18 independent variables; $^b X_2$: the variable μ was eliminated from matrix **X** for the model.

the relationship between a matrix **Y** (containing dependent variables, often only one for QSAR studies) and a matrix **X** (containing predictor variables) by reducing the dimension of the matrices while concurrently maximizing the relationship between them.

SIMCA-P (Version 10.5, Umetrics AB, 2004) software was employed to perform the PLS analysis. The default values given by the software were used as the initial conditions for computation. According to the user's guide to SIMCA-P [25], the criterion used to determine the model dimensionality, the number of significant PLS components, is cross-validation (CV). With CV, the tested PLS component is thought significant if the fraction of the total variation of the dependent variables estimated by a component, Q^2, is higher than a significance limit (0.0975) for the whole dataset. The model is thought to have a good predicting ability when the cumulative Q^2 for the extracted components, Q^2_{cum}, is more than 0.500 [25]. Model adequacy was assessed based primarily on the number of PLS principal components (h), Q^2_{cum}, the squared correlation coefficient between observed values and fitted values (R^2), the standard error (SE) of the estimate, the variance ratio of variance analysis (F), and the significance level (p).

3. Results and Discussion

3.1. Modeling and Optimizing. For the 20 PCBs contained in the training set, PLS analysis was performed with $\log K_{OW}$ as the dependent variable and the 18 independent variables as the predictor variables, yielding QSAR model **I**. For this model, we have $R^2 = 0.992$, SE $= 0.145$, $F = 2.12 \times 10^3$, and $p = 4.00 \times 10^{-20}$. The fitting results for this model are listed in Table 3. In Table 3, R^2_X and R^2_Y stand for the fraction of the sum of squares of all the X's and Y's explained by the current component, $R^2_{X(cum)}$ and $R^2_{Y(cum)}$ stand for the fraction of the cumulative variance of all the X's and Y's explained by all extracted components, respectively, and Eig stands for eigenvalue, which denotes the importance of the PLS principal component.

Variable importance in the projection (VIP) is a parameter that shows the importance of a variable in a model in the assistant analysis of PLS modeling. According to the manual of SIMCA-P, terms with large VIP (>1.0) are the most relevant for explaining dependent variables [25]. To obtain an optimal model, the following PLS analysis procedure was adopted. A PLS model with all the predictor variables was first calculated and then the variable with the lowest VIP was eliminated, and a new PLS regression was performed, yielding a new PLS

model. This procedure was repeated until an optimal model was obtained. The optimal model was selected with respect to Q^2_{cum}, R^2, SE, F, and p. According to the statistics and metrics theories, a QSAR model with larger values of Q^2_{cum}, R^2 and F and smaller values of SE and p tends to be more stable and reliable than in the opposite case.

The above described PLS analysis procedure with $\log K_{OW}$ as dependent variable and the 18 independent variables as the initial predictor variables, for the 20 PCBs contained in the training set, resulted in model **II**. For this model, we have $R^2 = 0.992$, SE $= 0.151$, $F = 1.97 \times 10^3$, and $p = 7.62 \times 10^{-20}$. The fitting results for this model are shown in Table 3. As shown in the table, based on the estimated indexes, model **I** is better than model **II**, indicating all that the selected 18 independent variables are necessary for the PLS modeling. In model **I**, 2 PLS principal components were selected in the model, which explained 69.1% of the variance of the predictor variables and 99.2% of the variance of the dependent variable. Based on the estimated indexes employed in this study, this model is statistically significant.

3.2. Analysis and Discussion. Based on the optimal model **I**, the predicted $\log K_{OW}$ for the 20 PCBs contained in the training set are as listed in Table 1. A comparison of the observed and predicted $\log K_{OW}$ is shown in Figure 2. As can be seen in Table 1 and Figure 2, the predicted values are within the range of ±0.3 log unit from the observed values. The observed $\log K_{OW}$ values are close to those predicted by the optimal model, and the correlation between observed and predicted $\log K_{OW}$ is significant. For the PCBs studied, the differences between observed and predicted $\log K_{OW}$ are very small, so the predicted values are acceptable. Note that the cross-validated Q^2_{cum} value (=0.988) of the optimal model is much larger than 0.500; it would seem that this model is stable and has good prediction ability and can be used to predict the K_{OW} of PCBs effectively.

When **X** and **Y** are unscaled and uncentered, the unscaled coefficients of the independent variables and the constant, transformed from the PLS results, are listed in Table 4. The signs of the coefficients of the independent variables in the model enable one to see the direction in which each affects the K_{OW} of PCBs. Based on the unscaled coefficients, a QSAR equation like that obtained from multiple linear regression can be created for the optimal model, and the predicted K_{OW} for other PCBs can be estimated. Based on model **I**, predictions of K_{OW} for the 7 PAHs contained in the test set were generated and listed in Table 1, and the comparison of observed and predicted $\log K_{OW}$ is shown in Figure 2.

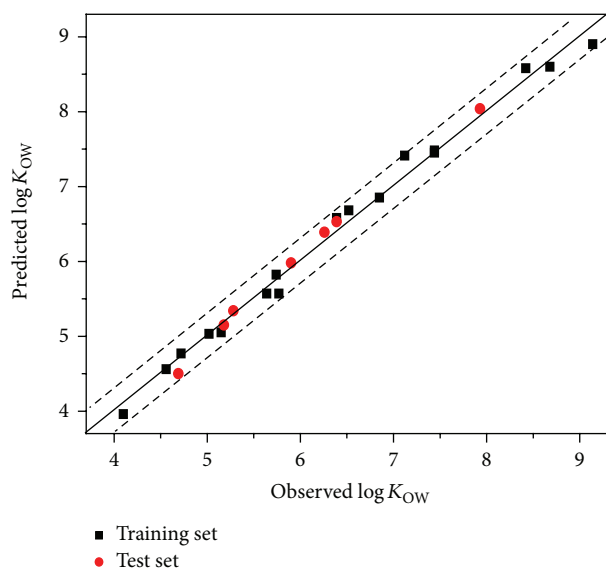FIGURE 2: Plot of observed values versus those predicted by the optimal model **I**.

TABLE 4: The unscaled coefficients and VIPs of the optimal model **I**.

Variables	Unscaled coefficients	VIPs
TE	-1.645×10^{-4}	1.412
Re	0.747×10^{-4}	1.404
E_{HOMO}	-10.390	1.274
E_{LUMO}	-20.295	1.220
Q_{C6}	3.418	1.082
Q_{C12}	3.366	1.074
Q_{C8}	1.924	1.009
Q_{C11}	5.212	0.992
Q_{C1}	-1.999	0.947
Q_{C9}	7.412	0.940
Q_{C7}	-1.975	0.926
Q_{C2}	1.087	0.922
Q_{C10}	4.711	0.834
DA	5.726×10^{-4}	0.794
Q_{C3}	6.512	0.756
Q_{C5}	3.434	0.716
Q_{C4}	3.535	0.690
μ	-0.047	0.473
Constant	6.995	/

The correlation coefficient between observed and predicted $\log K_{OW}$ in the test set was as high as 0.995, larger than that in the training set. As can be seen in Table 1 and Figure 2, the predicted $\log K_{OW}$ for the test set are very close to those observed. This indicates that the PLS model has good predictive ability.

The VIP values for the optimal model were calculated and listed in Table 4. We can find that the VIP values of TE, Re, E_{HOMO}, and E_{LUMO} are all greater than 1.0, indicating that they are important in governing the K_{OW} values of PCBs. TE and Re are the two most important variables since their VIP values are larger than 1.40, which is greater than the VIP values of the other variables. It is expected that $\log K_{OW}$ increases with the increase of molecular size. Considering the following examples: 4-chlorobiphenyl ($\log K_{OW}$ = 4.69, molecular weight (MW) = 188.5), 4, 4′-dichlorobiphenyl ($\log K_{OW}$ = 5.28, MW = 223.0), 2, 4, 4′-trichlorobiphenyl ($\log K_{OW}$ = 5.74, MW = 257.5), 3, 3′, 4, 4′-tetrachlorobiphenyl ($\log K_{OW}$ = 6.52, MW = 292.0), which differ between themselves by the number of chlorine substituents, the increase in $\log K_{OW}$ as a function of the number of chlorine substituents and molecular sizes is clearly seen. When H atom is replaced by Cl atom on the biphenyl, K_{OW} value will be increased [26]. So, the more chlorine atoms on the biphenyls molecular, the greater the K_{OW} value. When isomers share the same molecular weight, molecular sizes differ from the arrangements of chlorine atoms, which can be expressed by molecular volume and surface area, leading to the difference in $\log K_{OW}$. 2-Chlorobiphenyl ($\log K_{OW}$ = 4.56), 3-chlorobiphenyl ($\log K_{OW}$ = 4.72), and 4-chlorobiphenyl ($\log K_{OW}$ = 4.69) are isomers (MW = 188.5), but they have different n-octanol/water partition coefficients.

The variety of molecular size could be described by the quantum chemical descriptors of TE and Re in this work. Firstly, for the whole compounds set in this study, the decrease of TE values indicates that the molecule contains more chlorine atoms, which might lead to the increase of molecular volume. Zou et al. [27] found that molecular volume and molecular surface area were highly correlated for PCBs sharing similar structures. So, the decrease of TE values might lead to the increase of molecular surface area. Molecular volume and molecular surface area are measures of molecular size. In general, the bigger the molecular size, the stronger the effect of the intermolecular dispersion force. On the other hand, the molecular volume is a measure of absorbed energy of the molecule when it forms the hole in a solvent. Due to stronger polar of water molecules contributing to stronger binding energy, the larger solute molecules need to absorb more energy to destroy the binding between water molecules when distributing in the aqueous phase, resulting in tending to be allocated in weakly polar solvents. It means that the decrease of TE values might lead to larger $\log K_{OW}$ which is consistent with the study of Zou et al. [27]. Secondly, Re is the expectation value of the operator ρ in the formula of $\int r^2 \rho(r) d^3 r$. It is occasionally seen in the literature as a measure of molecular volume. Although in many cases Re correlates poorly with molecular volume, the correlation for PCBs sharing similar structures is quite significant. A PCB molecule with large electronic spatial extent will have large molecular volume. Therefore, it can be concluded that PCBs molecules with lower molecular total energy values and larger electronic spatial extent tend to be more hydrophobic and lipophilic, leading to larger $\log K_{OW}$ values. This implies that the decrease of the TE value and the increase of the Re value lead to an increase in $\log K_{OW}$ value.

The value of $\log K_{OW}$ is in inverse ratio to E_{HOMO}, and E_{LUMO}; that is, large value of E_{HOMO} and E_{LUMO} would result in small value of $\log K_{OW}$ for PCBs. This is because E_{HOMO}

represents the proton acceptance ability in forming hydrogen bond, while E_{LUMO} represents the proton donation ability in the formation of hydrogen bond. Therefore, the compounds with large values of E_{HOMO} and E_{LUMO} tend to donate or accept protons easily. In this case, hydrogen bond can be easily formed between PCBs and H_2O molecules, and PCBs are able to enter water phase, resulting in low $\log K_{OW}$ value [13].

3.3. Comparison with the Results from Other Models. In [13], there are 3 variables (E_{HOMO}, E_{LUMO}, and α) obtained from B3LYP/6-31G(d) level in the model, and the squared correlation coefficient is $R^2 = 0.9484$, less than that of the optimal model in the present study. It was found that there are high correlations relating the number of Cl substitution position (N) to α and E_{HOMO}. And, taking N as theoretical descriptors to correlate the three-variable models for predicting $\log K_{OW}$ of PCBs, the achieved new model is having $R^2 = 0.9497$ and exhibiting optimum stability, convenience for use, and better predictive power than that from AM1 method [28]. In contrast, the squared correlation coefficient $R^2 = 0.992$ obtained by PLS in the present study is obviously larger than that reported.

Taking molar weight (M) and melting point (T_m, °C) as independent variables, Wang et al. [10] had developed the binary linear regression equation

$$\log K_{OW} = 0.0155M + 0.00107 \left(T_m - 25\right) + 1.71 \quad (n = 27, r = 0.997), \tag{1}$$

whose predicted $\log K_{OW}$ are also listed in Table 1, which seems superior to the optimal model **I** in the present study obtained solely on quantum chemical descriptors. However, the variable melting point in the model equation is an experimental physicochemical parameter. It is impractical to determine the melting points of all 209 PCBs directly in the laboratory which makes the model become of limited applicability. Therefore, our optimal model based only on quantum chemical descriptors has wider applicability than that based on known physicochemical properties. The overall high quality of the model obtained in this study indicates that it will find application in the estimation of K_{OW} of PCBs for which experimental values are lacking.

4. Conclusions

In this work, based on some quantum chemical descriptors computed by DFT at the B3LYP/6-31G(d) level, by the use of PLS analysis, QSAR models were obtained for logarithmic n-octanol/water partition coefficients of PCBs. The squared correlation coefficient of the optimal model is 0.992. The optimal model has high fitting precision and good predictability, so it can be applied in the estimation of the n-octanol/water partition behavior of PCBs. It can generally be concluded that PCBs with larger electronic spatial extent and lower molecular total energy values tend to be more hydrophobic and lipophilic.

References

[1] B. G. Hansen, A. B. Paya-Perez, M. Rahman, and B. R. Larsen, "QSARs for K_{OW} and K_{OC} of PCB congeners: a critical examination of data, assumptions and statistical approaches," *Chemosphere*, vol. 39, no. 13, pp. 2209–2228, 1999.

[2] S.-S. Liu, Y. Liu, D.-Q. Yin, X.-D. Wang, and L.-S. Wang, "Prediction of chromatographic relative retention time of polychlorinated biphenyls from the molecular electronegativity distance vector," *Journal of Separation Science*, vol. 29, no. 2, pp. 296–301, 2006.

[3] United Nations Environment Programme, "Stockholm Convention on Persistent Organic Pollutants," 2001, http://www.pops.int/documents/convtext/convtext_en.pdf.

[4] J. P. Giesy and K. Kannan, "Dioxin-like and non-dioxin-like toxic effects of polychlorinated biphenyls (PCBs): implications for risk assessment," *Critical Reviews in Toxicology*, vol. 28, no. 6, pp. 511–569, 1998.

[5] C. T. Chiou, V. H. Freed, D. W. Schmedding, and R. L. Kohnert, "Partition coefficient and bioaccumulation of selected organic chemicals," *Environmental Science and Technology*, vol. 11, no. 5, pp. 475–478, 1977.

[6] H. Konemann, "Structure-activity relationships and additivity in fish toxicities of environmental pollutants," *Ecotoxicology and Environmental Safety*, vol. 4, no. 4, pp. 415–421, 1980.

[7] G. Patil, "Correlation of aqueous solubility and octanol-water partition coefficient based on molecular structure," *Chemosphere*, vol. 22, no. 8, pp. 723–738, 1991.

[8] D. Mackay, A. Bobra, W. Y. Shiu, and S. H. Yalkowsky, "Relationships between aqueous solubility and octanol-water partition coefficients," *Chemosphere*, vol. 9, no. 11, pp. 701–711, 1980.

[9] A. Sabljic, H. Gusten, J. Hermens, and A. Opperhuizen, "Modeling octanol/water partition coefficients by molecular topology: chlorinated benzenes and biphenyls," *Environmental Science and Technology*, vol. 27, no. 7, pp. 1394–1402, 1993.

[10] L. S. Wang, Y. H. Zhao, and H. Gao, "Predicting aqueous solubility and octanol/water partition coefficients of organic chemicals from molar volume," *Environmental Chemistry*, vol. 11, pp. 55–70, 1992.

[11] G.-N. Lu, Z. Dang, X.-Q. Tao, and X.-Y. Yi, "Estimation of water solubility of polycyclic aromatic hydrocarbons using quantum chemical descriptors and partial least squares," *QSAR and Combinatorial Science*, vol. 27, no. 5, pp. 618–626, 2008.

[12] G.-N. Lu, X.-Q. Tao, Z. Dang, X.-Y. Yi, and C. Yang, "Estimation of n-octanol/water partition coefficients of polycyclic aromatic hydrocarbons by quantum chemical descriptors," *Central European Journal of Chemistry*, vol. 6, no. 2, pp. 310–318, 2008.

[13] W. Zhou, Z. Zhai, Z. Wang, and L. Wang, "Estimation of *n*-octanol/water partition coefficients (K_{OW}) of all PCB congeners by density functional theory," *Journal of Molecular Structure*, vol. 755, no. 1–3, pp. 137–145, 2005.

[14] F. Verweij, K. Booij, K. Satumalay, N. Van Der Molen, and R. Van Der Oost, "Assessment of bioavailable PAH, PCB and OCP concentrations in water, using semipermeable membrane devices (SPMDs), sediments and caged carp," *Chemosphere*, vol. 54, no. 11, pp. 1675–1689, 2004.

[15] L. Jäntschi, S. Bolboacă, and R. Sestraş, "Meta-heuristics on quantitative structure -activity relationships: study on polychlorinated biphenyls," *Journal of Molecular Modeling*, vol. 16, pp. 377–386, 2010.

[16] P. Ruiz, O. Faroon, C. J. Moudgal, H. Hansen, C. T. De Rosa, and M. Mumtaz, "Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure-activity relationship (QSAR)," *Toxicology Letters*, vol. 181, no. 1, pp. 53–65, 2008.

[17] L.-T. Qin, S.-S. Liu, H.-L. Liu, and H.-L. Ge, "A new predictive model for the bioconcentration factors of polychlorinated biphenyls (PCBs) based on the molecular electronegativity distance vector (MEDV)," *Chemosphere*, vol. 70, no. 9, pp. 1577–1587, 2008.

[18] K. Tuppurainen and J. Ruuskanen, "Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital energies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-*p*-dioxins (PCDDs) and dibenzofurans (PCDFs)," *Chemosphere*, vol. 41, no. 6, pp. 843–848, 2000.

[19] Z. Daren, "QSPR studies of PCBs by the combination of genetic algorithms and PLS analysis," *Computers and Chemistry*, vol. 25, no. 2, pp. 197–204, 2001.

[20] E. B. de Melo, "A new quantitative structure-property relationship model to predict bioconcentration factors of polychlorinated biphenyls (PCBs) in fishes using E-state index and topological descriptors," *Ecotoxicology and Environmental Safety*, vol. 75, no. 1, pp. 213–222, 2012.

[21] J. Y. Long, H. B. Yi, X. K. Liu, and Y. F. Wang, "Quantitative structure-property relationship for polychlorinated biphenyls: toxicity and structure by density functional theory," *Acta Chimica Sinica*, vol. 70, pp. 949–960, 2012.

[22] G.-N. Lu, X.-Q. Tao, Z. H. I. Dang, W. Huang, and Z. Li, "Quantitative structureproperty relationships on dissolvability of pcdd/fs using quantum chemical descriptors and partial least squares," *Journal of Theoretical and Computational Chemistry*, vol. 9, no. 1, pp. 9–22, 2010.

[23] M. J. Frisch, G. W. Trucks, H. B. Schlegel et al., *Gaussian 03, Revision B. 01*, Gaussian, Pittsburgh, Pa, USA, 2003.

[24] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 109–130, 2001.

[25] A. B. Umetrics, *USer's Guide to SIMCA-P, SIMCA-P+, Version 10. 0*, Umeå, Sweden, 2005.

[26] B. Z. Zhao, Y. Y. Li, X. R. Hao et al., "Theoretical calculation of the octanol/water partition coefficient for polychlorobiphenyls," *Journal of Northeast Normal University*, vol. 36, no. 2, pp. 41–44, 2004.

[27] J.-W. Zou, Y.-J. Jiang, G.-X. Hu, M. Zeng, S.-L. Zhuang, and Q.-S. Yu, "QSPR/QSAR studies on the physicochemical properties and biological activities of polychlorinated biphenyls," *Acta Physico*, vol. 21, no. 3, pp. 267–272, 2005.

[28] X.-Y. Han, Z.-Y. Wang, Z.-C. Zhai, and L.-S. Wang, "Estimation of n-octanol/water partition coefficients (K_{OW}) of all PCB congeners by *ab initio* and a Cl substitution position method," *QSAR and Combinatorial Science*, vol. 25, no. 4, pp. 333–341, 2006.