VERSITA

Cent. Eur. J. Chem. • 6(2) • 2008 • 310–318
DOI: 10.2478/s11532-008-0010-y

Central European Journal of **Chemistry**

# Estimation of *n*-octanol/water partition coefficients of polycyclic aromatic hydrocarbons by quantum chemical descriptors

**Research Article**

Gui-Ning Lu[1,3], Xue-Qin Tao[2], Zhi Dang[1*], Xiao-Yun Yi[1], Chen Yang[1]

[1] *School of Environmental Science and Engineering, South China University of Technology, Guangzhou Higher Education Mega Center, Guangzhou 510006, P.R. China*

[2] *Department of Environmental Science and Engineering, Zhongkai University of Agriculture and Technology, Guangzhou 510225, P.R. China*

[3] *Department of Environmental Sciences, School of Environmental and Biological Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA*

**Received 3 October 2007; Accepted 14 January 2008**

**Abstract:** Quantitative structure-property relationship (QSPR) modeling is a powerful approach for predicting environmental behavior of organic pollutants with their structure descriptors. This study reports an optimal QSPR model for estimating logarithmic *n*-octanol/water partition coefficients (log $K_{OW}$) of polycyclic aromatic hydrocarbons (PAHs). Quantum chemical descriptors computed with density functional theory at B3LYP/6-31G(d) level and partial least squares (PLS) analysis with optimizing procedure were used for generating QSPR models for log $K_{OW}$ of PAHs. The squared correlation coefficient ($R^2$) of the optimal model was 0.990, and the results of cross-validation test ($Q^2_{cum}$=0.976) showed this optimal model had high fitting precision and good predictability. The log $K_{OW}$ values predicted by the optimal model are very close to those observed. The PLS analysis indicated that PAHs with larger electronic spatial extent and lower total energy values tend to be more hydrophobic and lipophilic.

**Keywords:** *PAH • QSPR • Quantum chemical descriptors*



## 1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) constitute a large and diverse class of organic compounds consisting of two or more fused aromatic rings in various structural configurations generated from both natural and anthropogenic processes [1,2]. The low water solubility of PAHs result in high persistence of these chemicals at contaminated sites. They have been detected in the atmosphere, water, soil, sediment and food [3-7].

The environmental fate of PAHs has become a major issue in recent years [8,9], since many PAHs such as benzo[*a*]pyrene, chrysene and benz[*a*]anthracene are mutagens and carcinogens [10] and are considered to be primary pollutants by many countries. Understanding the distribution of PAHs among environmental phases is crucial to their risk assessments and remediation of contaminated sites. It is well established that the fate of PAHs in the environment is primarily controlled by their physicochemical properties [11], such as the *n*-octanol/water partition coefficients ($K_{OW}$), which estimates the solubility in both aqueous and organic phases (in general *n*-octanol is used). By the definition, $K_{OW}$ is inversely proportional to aqueous solubility.

* *E-mail: chzdang@scut.edu.cn*

When $K_{OW}$>1, compounds are lipophilic or hydrophobic, and hydrophilic while $K_{OW}$<1 [12]. Since values of $K_{OW}$ may vary by several orders of magnitude, it is usually expressed in the logarithmic form.

The $K_{OW}$ may significantly influence the chemical and biological transformation or degradation of chemicals, so it is essential for understanding the transport mechanism and distribution of compounds into the environment, for example, the mechanism that involves drug absorption by transport through a biological membrane, or the process involving the deposition of a pollutant into bodies of water [13]. Thus the measurement or accurate estimation of $K_{OW}$ is of critical importance for evaluating the fate and potential exposure of chemicals in the environment, and consequently, for the whole process of environmental risk assessment. In general, compounds with higher values of $K_{OW}$ tend to be less mobile than those with lower values in soil-water systems [14]. However, the accurate determination of $K_{OW}$ may be difficult and expensive in terms of cost and time, or even impossible for some compounds, which might not have been synthesized or purified. Also experimental errors may be introduced, especially for those congeners difficult to be separated and identified by chromatography. Moreover, it is impractical to measure $K_{OW}$ of all PAHs directly in the laboratory because there are so many PAHs that have been found in the environment.

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-property relationships (QSPR), which correlate the properties of pollutants with relevant properties and molecular descriptors [15]. A large number of calculation methods have been presently developed for estimation of the partition coefficients with varying success and applicability. According to the descriptors used, these methods can be classified into two groups: empirical and theoretical methods [16]. Chu and Chan [17] reported the relationships between soil sorption coefficients ($K_{OC}$), water solubility ($S$), and $K_{OW}$ of a diverse collection of pollutants, use whole word not abbreviation aliphatics, aromatics, pesticides, herbicides and PAHs. Such property correlations are designed to estimate properties of environmental interest from other known physicochemical properties which work reasonably well. It is however limited by the unavailability of the latter properties for the majority of chemicals of environmental concern [18]. Various studies have shown that parameters such as *n*-octanol/air partition coefficients ($K_{OA}$), $S$, $K_{OW}$ and $K_{OC}$ are correlated to some molecular descriptors, which can be calculated directly just from chemical structures without the input of any other experimental data [13,16,19-26]. Predictive models based on non-experimental molecular descriptors can provide cost effective and rapid estimates of partitioning behavior of contaminants. Topological, geometrical and quantum chemical indexes comprise a set of descriptors, which were useful in the prediction of properties of structurally similar molecules [27-30]. Basak and Mills [31] developed predictive models solely on topological and geometrical descriptors for $S$, $K_{OW}$ and $K_{OC}$ of 136 chemicals including 19 PAHs.

Quantum chemical descriptors can be easily obtained by computation to clearly describe specific molecular properties for structurally related compounds, and can also provide insight into the environmental behavior of chemicals not yet synthesized or those that cannot be examined experimentally due to their extremely hazardous nature. Hence, the development of QSPR models in which quantum chemical descriptors are used is of great importance [32,33]. Rapid advancement of modern computational capability and development of fast algorithms allow the high precision method to be expeditiously applied in current QSPR studies [16, 25,26,34-38], several of which are about partitioning properties of environmental pollutants [16,25,26]. Modern theoretical method in quantum chemistry with high calculation precision was proved having its advantages in estimating properties of environmental concern [16,35]. However, few QSPR studies on partitioning behavior of PAHs using quantum chemical descriptors have appeared so far.

The aim of this work is to develop a new reliable and predictive QSPR model for log $K_{OW}$ of PAHs using partial least squares (PLS) analysis with optimizing procedure, based on reported $K_{OW}$ and/or log $K_{OW}$ data and quantum chemical descriptors computed by density functional theory (DFT) contained in Gaussian 03 [39].

## 2. Materials and Methods

### 2.1. Target PAHs

A total of 24 PAHs containing 2 to7 fused rings, whose $K_{OW}$ and/or log $K_{OW}$ data were previously published [14, 17,40], were chosen to constitute the training and test set in this work. The training set consists of 18 of these 24 PAHs selected randomly, and the rest of the 6 PAHs constitute the test set. Their chemical abstracts service numbers (CAS No.) and reported log $K_{OW}$ are listed in Table 1 and molecular structures are given in Fig. 1. The compound numbers in Figure 1 correspond to those in Table 1. As shown in Fig. 1, the training set of this work consists of non-substituted five- and six-membered ring PAHs, as well as, alkyl-substituted PAHs.

**Table 1.** *n*-Octanol/water partition coefficients of the PAHs studied.

| No.[a] | Compounds | CAS No. | log $K_{OW}$ | | SE [b] | Diff. [c] |
|---|---|---|---|---|---|---|
| | | | Observed | Predicted | | |
| 1 | Naphthalene | 91-20-3 | 3.37 [d] | 3.49 | 0.070 | –0.12 |
| 2 | Anthracene | 120-12-7 | 4.54 [d] | 4.66 | 0.062 | –0.12 |
| 3 | Phenanthrene | 85-01-8 | 4.57 [d] | 4.46 | 0.046 | 0.11 |
| 4 | Chrysene | 218-01-9 | 5.86 [d] | 5.71 | 0.038 | 0.15 |
| 5 | Benz[*a*]anthracene | 56-55-3 | 5.91 [d] | 5.85 | 0.035 | 0.06 |
| 6 | Benzo[*a*]pyrene | 50-32-8 | 6.04 [d] | 6.19 | 0.047 | –0.15 |
| 7 | Acenaphthene | 83-32-9 | 3.92 [d] | 3.97 | 0.069 | –0.05 |
| 8 | Fluorene | 86-73-7 | 4.18 [d] | 4.28 | 0.054 | –0.09 |
| 9 | Fluoranthene | 206-44-0 | 5.22 [d] | 5.18 | 0.042 | 0.04 |
| 10 | Benzo[a]fluorene | 238-84-6 | 5.40 [d] | 5.52 | 0.051 | –0.12 |
| 11 | Triphenylene | 217-59-4 | 5.49 [d] | 5.44 | 0.063 | 0.05 |
| 12 | Perylene | 198-55-0 | 6.25 [d] | 6.13 | 0.059 | 0.12 |
| 13 | 1-Methylnaphthalene | 90-12-0 | 3.87 [e] | 3.84 | 0.046 | 0.03 |
| 14 | 2-Methylnaphthalene | 91-57-6 | 3.86 [e] | 3.78 | 0.046 | 0.08 |
| 15 | 9-Methylanthracene | 779-02-2 | 5.07 [e] | 5.00 | 0.061 | 0.07 |
| 16 | Benzo[*b*]fluorene | 243-17-4 | 5.75 [d] | 5.64 | 0.054 | 0.11 |
| 17 | Benzo[*g,h,i*]perylene | 191-24-2 | 6.50 [d] | 6.52 | 0.047 | –0.02 |
| 18 | Coronene | 191-07-1 | 6.75 [d] | 6.89 | 0.065 | –0.14 |
| 19 | Acenaphthylene | 208-96-8 | 4.00 [d] | 4.24 | 0.067 | –0.24 |
| 20 | 1-Ethylnaphthalene | 1127-76-0 | 4.39 [e] | 4.13 | 0.043 | 0.26 |
| 21 | Pyrene | 129-00-0 | 4.88 [e, f] | 4.90 | 0.047 | –0.02 |
| 22 | Naphthacene | 92-24-0 | 5.90 [e] | 6.10 | 0.072 | –0.20 |
| 23 | Benzo[*b*]fluoranthene | 205-99-2 | 6.06 [f] | 6.37 | 0.052 | –0.31 |
| 24 | Indeno[1,2,3-c,d]pyrene | 193-39-5 | 6.50 [f] | 6.89 | 0.056 | –0.39 |

[a] *Compounds No. 1~18 constitute the training set and compounds No. 19~24 constitute the test set;*
[b] *SE represents standard error for the predicted log $K_{OW}$ values;*
[c] *Diff. = log $K_{OW}$ (observed) – log $K_{OW}$ (predicted);*
[d] *From ref. [40];*
[e] *From ref. [14];*
[f] *From ref. [17].*

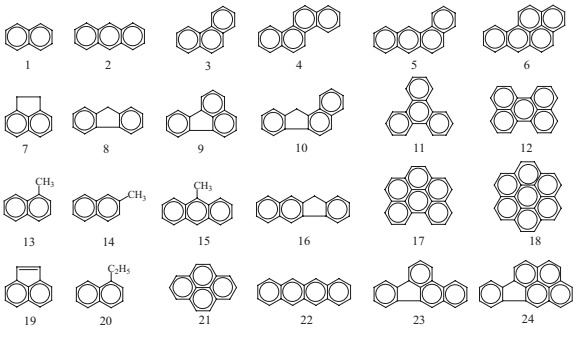


**Figure 1.** Molecular structures of the PAHs studied.

## 2.2. Calculation and selection of 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 [41,42]. The B3LYP calculations were performed using the quantum chemical computation software Gaussian 03 [39]. All calculations were run on an Intel Pentium D/2.66 GHz computer equipped with 1024 megabytes of internal memory and Microsoft Windows XP professional operating system.

According to the chemometrics theory, it is suggested that a QSPR model should include as many relevant descriptors as possible to increase the probability of a good characterization for a class of compounds [43]. In this study, 11 independent variables including 8 quantum chemical descriptors were selected for developing QSPR models. The 8 descriptors cover the eigenvalue of the highest occupied molecular orbital ($E_{HOMO}$), the eigenvalue of the lowest unoccupied molecular orbital ($E_{LUMO}$), molecular total energy (*TE*), dipole moment (μ), the most negative Mulliken atomic charges on a carbon

**Table 2.** Quantum chemical descriptors for the PAHs studied.

| No. | $E_{HOMO}$ | $E_{LUMO}$ | *TE* | μ | $Q_C^-$ | $Q_H^+$ | $L_{CC}$ | $R_e$ |
|---|---|---|---|---|---|---|---|---|
| 1 | –0.21267 | –0.03527 | –385.8927289 | 0.0000 | –0.190922 | 0.129521 | 1.4345 | 1291.7137 |
| 2 | –0.19201 | –0.06005 | –539.5305216 | 0.0000 | –0.297755 | 0.130208 | 1.4460 | 2861.0192 |
| 3 | –0.21056 | –0.03655 | –539.5386564 | 0.0420 | –0.206640 | 0.133247 | 1.4577 | 2623.7434 |
| 4 | –0.20248 | –0.04660 | –693.1820248 | 0.0000 | –0.206250 | 0.134154 | 1.4529 | 4779.3414 |
| 5 | –0.19556 | –0.05696 | –693.1789643 | 0.0657 | –0.316335 | 0.134410 | 1.4655 | 4932.2234 |
| 6 | –0.18748 | –0.06385 | –769.4137837 | 0.0454 | –0.336304 | 0.134338 | 1.4431 | 5260.4605 |
| 7 | –0.20094 | –0.02781 | –463.3149436 | 0.8187 | –0.361712 | 0.159215 | 1.5688 | 1746.1112 |
| 8 | –0.21150 | –0.02622 | –501.4232016 | 0.4821 | –0.420260 | 0.172179 | 1.5161 | 2352.1686 |
| 9 | –0.21206 | –0.06438 | –615.7502058 | 0.3291 | –0.220472 | 0.133868 | 1.4759 | 3194.2266 |
| 10 | –0.20048 | –0.04179 | –655.0668764 | 0.4401 | –0.421373 | 0.173231 | 1.5153 | 4339.4356 |
| 11 | –0.21488 | –0.03422 | –693.1810887 | 0.0000 | –0.202328 | 0.134104 | 1.4669 | 4119.3318 |
| 12 | –0.18198 | –0.06959 | –769.4061188 | 0.0000 | –0.226768 | 0.132838 | 1.4762 | 4833.1441 |
| 13 | –0.20853 | –0.03389 | –425.2096910 | 0.2957 | –0.535209 | 0.162462 | 1.5104 | 1613.7939 |
| 14 | –0.20816 | –0.03418 | –425.2102103 | 0.4221 | –0.531849 | 0.164105 | 1.4336 | 1778.0688 |
| 15 | –0.18787 | –0.05886 | –578.8421060 | 0.3236 | –0.540751 | 0.166254 | 1.5139 | 3158.3301 |
| 16 | –0.20280 | –0.04408 | –655.0675111 | 0.5856 | –0.419917 | 0.172820 | 1.5175 | 4587.3211 |
| 17 | –0.19120 | –0.06102 | –845.6544426 | 0.0669 | –0.228600 | 0.133409 | 1.4694 | 5447.8122 |
| 18 | –0.20031 | –0.05192 | –921.8978878 | 0.0000 | –0.208795 | 0.130048 | 1.4275 | 6177.9855 |
| 19 | –0.21336 | –0.06933 | –462.0881949 | 0.3351 | –0.221442 | 0.134607 | 1.4726 | 1653.0001 |
| 20 | –0.20804 | –0.03313 | –464.5209840 | 0.4208 | –0.453608 | 0.153446 | 1.5313 | 2122.3544 |
| 21 | –0.19575 | –0.05440 | –615.7731341 | 0.0000 | –0.225803 | 0.130157 | 1.4376 | 2991.0909 |
| 22 | –0.17843 | –0.07635 | –693.1658119 | 0.0000 | –0.298815 | 0.130654 | 1.4529 | 5445.1396 |
| 23 | –0.21027 | –0.06309 | –769.3987687 | 0.3823 | –0.305885 | 0.134380 | 1.4750 | 5423.8821 |
| 24 | –0.19624 | –0.07356 | –845.6292983 | 0.6075 | –0.314450 | 0.134606 | 1.4760 | 6167.5033 |

atom ($Q_C^-$), the most positive Mulliken atomic charges on a hydrogen atom ($Q_H^+$), the largest bond length between two carbon atoms ($L_{CC}$), and the electronic spatial extent ($R_e$). 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 energy, dipole moment, atomic charge, bond length and extent are: hartree, debye, atomic charge unit, angstrom and atom unit, respectively. In addition, 3 combinations of frontier molecular orbital eigenvalues, $E_{LUMO}–E_{HOMO}$, $(E_{LUMO}–E_{HOMO})^2$ and $E_{LUMO}+E_{HOMO}$ were also selected as independent variables. $E_{LUMO}–E_{HOMO}$ and $E_{LUMO}+E_{HOMO}$ can be related to PAH absolute hardness and electronegativity, respectively [44]. $E_{LUMO}–E_{HOMO}$ and $(E_{LUMO}–E_{HOMO})^2$ were proven to be significant in QSAR studies on photoinduced toxicity of PAHs and QSPR studies on direct photolysis rate constants of PAHs [45,46]. And $E_{LUMO}+E_{HOMO}$ was shown to be significant in direct photolysis quantum yield QSPRs studies of substituted aromatic halides [47].

## 2.3. Modeling method and evaluation indexes

Since a large number of descriptors were selected in this study, intercorrelation of independent variables (multicollinearity) might become a technical problem. To overcome this problem, the PLS regression, a methodology that makes use of all available descriptors as opposed to subset regression and is useful when the descriptors are strongly collinear [43], was used. PLS could find the relationship between a matrix ***Y*** (containing dependent variables, often only one for QSPR studies) and a matrix ***X*** (containing predictor variables) by reducing the dimension of the matrices while concurrently maximizing the relationship between them [48].

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 [49], the criterion used to determine the model dimensionality, *ie,* the number of significant PLS components (*h*), is 7-fold cross validation (CV). With CV, observations are excluded from the model; the response values for the excluded observations are

predicted by the model and compared with the actual values. This procedure is repeated several times until every observation has been excluded once and only once. For every component, the fraction of the total variation of the dependent variables ($Q^2$) and the cumulative $Q^2$ for the extracted components ($Q^2_{cum}$) are computed using the following equations:

$$Q^2 = 1.0 - \frac{PRESS}{SS} \quad (1)$$

$$Q^2_{cum} = 1.0 - \Pi\left(\frac{PRESS}{SS}\right)_n \qquad (n = 1,.....h) \quad (2)$$

where *PRESS* is the prediction error sum of squares when the observations were excluded, and *SS* is the residual sum of squares of the previous component. The tested PLS component is thought significant if $Q^2$ is larger than a significance limit (0.0975) for the whole data set. The model is thought to have a good predictability when $Q^2_{cum}$ is larger than 0.5 [49]. Model adequacy was assessed based primarily on *h*, $Q^2_{cum}$, the squared correlation coefficient between observed values and fitted values ($R^2$), the standard error of the estimate (*SE*), the variance ratio (*F*), and the significance level (*p*).

# 3. Results and Discussion

## 3.1. Modeling and optimizing

For the 18 PAHs contained in the training set, PLS analysis was performed with log $K_{OW}$ as the dependent variable and the 11 independent variables as the predictor variables, yielding QSPR model **I**. For this model, we have $R^2$=0.987, *SE*=0.120, $F$=1.20×10$^3$, and $p$=2.22×10$^{-16}$. 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 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 cumulative variance of all the ***X***'s and ***Y***'s explained by all extracted components respectively, and *Eig* stands for eigenvalue that 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. Previous studies found that all independent variables were not necessary for PLS modeling [29,30,50-52]. 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 11 independent variables as the initial predictor variables, for the 18 PAHs contained in the training set, resulted in model **II** as an optimal one. For this optimal model, we have $R^2$=0.990, *SE*=0.106, $F$=1.56×10$^3$, and $p$<1.00×10$^{-16}$. The fitting results for the optimal model are shown in Table 3. As shown in the table, 3 PLS principal components were selected in the optimal model, which explained 88.5% of the variance of the predictor variables and 99.0% of the variance of the dependent variable. Based on the estimate indexes employed in this study, this optimal model is statistically significant.

## 3.2. Analysis and Discussion

The predicted log $K_{OW}$ calculated from the optimal model for the 18 PAHs contained in the training set were listed in Table 1 and the comparison of observed and predicted log $K_{OW}$ was shown in Fig. 2. 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 PAHs under study, the differences between observed and predicted log $K_{OW}$

**Table 3.** Fitting results for PLS model **I** and **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 | 0.538 | 0.538 | 0.759 | 0.759 | 5.92 | 0.712 | 0.712 |
| | | | 2 | 0.138 | 0.676 | 0.202 | 0.961 | 1.52 | 0.760 | 0.931 |
| | | | 3 | 0.190 | 0.867 | 0.026 | 0.987 | 2.09 | 0.598 | 0.972 |
| **II** | log $K_{OW}$ | $X_2$ [b] | 1 | 0.568 | 0.568 | 0.761 | 0.761 | 5.68 | 0.718 | 0.718 |
| | | | 2 | 0.127 | 0.695 | 0.212 | 0.973 | 1.27 | 0.813 | 0.947 |
| | | | 3 | 0.190 | 0.885 | 0.016 | 0.990 | 1.90 | 0.545 | 0.976 |

[a] $X_1$ *containing all 11 independent variables;*
[b] $X_2$ *containing 10 independent variables, not including* $Q_C^-$.

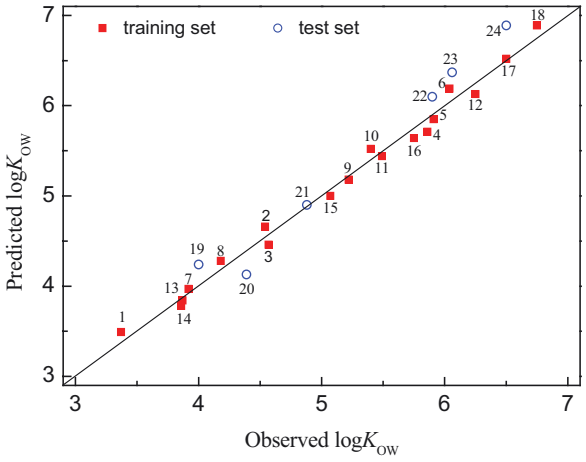


**Figure 2.** Plot of observed values vs. those predicted by the optimal model.

are very small, so the predicted values are acceptable. Note that the cross-validated $Q^2_{cum}$ value (=0.976) of the optimal model is not only remarkably larger than 0.500, but also larger than that of model **I** without optimizing; thus this optimal model is stable and is a good predictor of $K_{OW}$ for PAHs.

When ***X*** and ***Y*** are unscaled and uncentered, the unscaled coefficients of the independent variables and constant transformed from the PLS results are listed in Table 4. From the positive or negative signs of the coefficients of the independent variables, one can evaluate the effect of each independent variable on the $K_{OW}$ of PAHs. Based on the unscaled coefficients, a QSPR equation like that obtained from multiple linear regressions can be created for the optimal model and predicted $K_{OW}$ for other PAHs can be calculated. Based on model **II**, predictions of $K_{OW}$ for the 6 PAHs containing in the test set were generated and listed in Table 1 and the comparison of observed and predicted log $K_{OW}$ was shown in Fig. 2. The squared correlation coefficient between observed and predicted log $K_{OW}$ in the test set was as high as 0.976, close to that in the training set. As can be seen in Table 1 and Fig. 2, the predicted log $K_{OW}$ values for the test set are very close to those observed. This indicated that the optimal model has stable and good predictability.

The *VIP* values for the optimal model were calculated and listed in Table 4. We can find that the *VIP* values of $R_e$, *TE*, and $E_{LUMO}$ are all greater than 1.0, indicating they are important in governing the $K_{OW}$ values of PAHs. In addition, the *VIP* values of $E_{LUMO}$–$E_{HOMO}$ and $(E_{LUMO}–E_{HOMO})^2$ are very close to 1.0. $R_e$ and *TE* are the two most important variables since their *VIP* values are larger than 1.50, which is greater than the *VIP* values of the other variables. As shown in Table 5, the absolute value of correlation coefficient between $R_e$ and *TE* is as high

**Table 4.** The unscaled coefficients and *VIPs* of the optimal model.

| Variables | Unscaled Coefficients | *VIPs* |
|---|---|---|
| $R_e$ | $3.278\times10^{-4}$ | 1.543 |
| *TE* | $-3.169\times10^{-3}$ | 1.517 |
| $E_{LUMO}$ | –2.193 | 1.002 |
| $(E_{LUMO}–E_{HOMO})^2$ | –3.941 | 0.989 |
| $E_{LUMO}–E_{HOMO}$ | $-7.247\times10^{-1}$ | 0.982 |
| $E_{HOMO}$ | $-6.433\times10^{-1}$ | 0.814 |
| μ | $-3.767\times10^{-2}$ | 0.688 |
| $Q_{H^+}$ | 1.612 | 0.680 |
| $E_{LUMO}+E_{HOMO}$ | –5.630 | 0.667 |
| $L_{CC}$ | 1.425 | 0.574 |
| Constant | –1.766 | |

as 0.976, and $E_{LUMO}$, $E_{LUMO}–E_{HOMO}$ and $(E_{LUMO}–E_{HOMO})^2$ are highly intercorrelated for the PAHs under study, while the correlations between the former two variables and the latter three variables are inconspicuous. Thus these five variables can be divided into two groups by their cross-correlations.

It is expected that log $K_{OW}$ increases with the increasing of molecular size. Considering the examples illustrated previously (see structures in Fig. 1): naphthalene (log $K_{OW}$=3.37, molecular weight (*MW*)=128.18), anthracene (log $K_{OW}$=4.54, *MW*=178.24) and naphthacene (log $K_{OW}$=5.90, *MW*=228.30), which differ among themselves by the number of fused rings arranged linearly, the increase in log $K_{OW}$ as a function of the number of rings and molecular sizes is clearly seen. When isomers share the same molecular weight, molecular sizes differ from the arrangements of atoms, which can be expressed by molecular volume and surface area, leading to the difference in log $K_{OW}$. For example, chrysene (log $K_{OW}$=5.86), benz[*a*]anthracene (log $K_{OW}$=5.91), triphenylene (log $K_{OW}$=5.49) and naphthacene (log $K_{OW}$=5.90) are isomers (*MW*=228.30, structures shown in Fig. 1), but they have different *n*-octanol/water partition coefficients.

The variety of molecular size could be described by the quantum chemical descriptors of $R_e$ and *TE* in this work. Firstly, $R_e$ is the expectation value of the operator ρ in the formula of $\int r^2\rho(r)d^3r$. It is occasionally seen in the literature as a measure of molecular volume. Although in many cases $R_e$ correlates poorly with molecular volume, the correlation for PAHs sharing similar structures is quite significant. A PAH molecule with large electronic spatial extent will have large molecular volume [30]. Secondly, for the whole compounds set in this study, the decrease of *TE* values indicates more carbon and hydrogen atoms containing in the molecule, which might lead to the increase of molecular volume. Lu et al. [30] found that *TE* correlated with molecular volume

**Table 5.** Cross-correlation coefficients for some important variables (N=24).

| | $R_e$ | *TE* | $E_{LUMO}$ | $E_{LUMO}$–$E_{HOMO}$ | $(E_{LUMO}–E_{HOMO})^2$ |
|---|---|---|---|---|---|
| $R_e$ | 1.000 | | | | |
| *TE* | –0.976 | 1.000 | | | |
| $E_{LUMO}$ | –0.615 | 0.627 | 1.000 | | |
| $E_{LUMO}$–$E_{HOMO}$ | –0.613 | 0.615 | 0.955 | 1.000 | |
| $(E_{LUMO}–E_{HOMO})^2$ | –0.626 | 0.624 | 0.958 | 0.998 | 1.000 |

significantly for PAHs sharing similar structures. When the carbon and hydrogen atoms of different PAHs were equal, *TE* values differed from isomers. As shown in Table 4, the coefficient sign of $R_e$ and *TE* are positive and negative, respectively. This implies that the increase of the $R_e$ value and the decrease of the *TE* value lead to an increase in log $K_{OW}$ value. So it can be concluded that PAHs molecules with larger electronic spatial extent and lower molecular total energy values tend to be more hydrophobic and lipophilic, leading to larger log $K_{OW}$ values.

The gap between $E_{LUMO}$ and $E_{HOMO}$, $E_{LUMO}$–$E_{HOMO}$, which defines the energy necessary to excite an electron from the highest occupied and the lowest unoccupied molecular orbital, turned out to be a useful descriptor [12]. $E_{LUMO}$–$E_{HOMO}$ is related to absolute hardness [53], defined as half the absolute value of $E_{LUMO}$–$E_{HOMO}$, which is regarded as a measure of energy stabilization in chemical systems: Chemical structures tend to be more stable at large values of the $E_{LUMO}$–$E_{HOMO}$ gap [54]. As can be seen in Table 4, $E_{LUMO}$, $E_{LUMO}$–$E_{HOMO}$ and $(E_{LUMO}–E_{HOMO})^2$ are shown to be significant besides $R_e$ and *TE* in this study and all of their coefficient signs are negative. As $E_{LUMO}$ is always larger than $E_{HOMO}$ (Table 2), so the value of $E_{LUMO}$–$E_{HOMO}$ is positive and the $(E_{LUMO}–E_{HOMO})^2$ value decreases along with the decrease of $E_{LUMO}$–$E_{HOMO}$ value. Thus the decrease of $E_{LUMO}$ and $E_{LUMO}$–$E_{HOMO}$ values leads to the increase in log $K_{OW}$ values.

### 3.3. Comparison with the results from other models

So far as in the literatures, there have been several QSPR studies on log $K_{OW}$ of PAHs (e.g. ref. [12,13,31]). The best squared correlation coefficient of the QSPR model for log $K_{OW}$ of PAHs developed solely on calculated descriptors (molecular weight and volume) was 0.976, less than that of the optimal model in the present study obtained solely on quantum chemical descriptors. In addition, only non-substituted PAHs containing 2~7 fused rings with five and six carbon atoms were included in that work [13]. The highest *R* checked in literature is 0.996, generated from PLS analysis on 5 combinatorial descriptors including electron affinity, edge-connectivity, surface area, enthalpy of formation and retention index for a data set only consisting of six-membered ring PAHs [12]. It seems superior to the optimal model in this study. However, as stated earlier, such models based on known physicochemical properties were of limited applicability. Therefore, our optimal model based only on quantum chemical descriptors has wider applicability than those based on known physicochemical properties and/or topological descriptors. This work demonstrated again that modern high precision quantum chemical method can be an effective means in QSPR study of environmental contaminants, and a series of similar study in our laboratory is under way. The overall high quality of the obtained optimal model in this study indicates that it will find application in the estimation of *n*-octanol/water partition coefficients of PAHs having no known experimental values.

## 4. Conclusions

In this study, based on some quantum chemical descriptors computed by DFT at the B3LYP/6-31G(d) level, by the use of PLS analysis, QSPR models were obtained for logarithmic *n*-octanol/water partition coefficients of PAHs. The squared correlation coefficient of the optimal model is 0.990. 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 PAHs. It can generally be concluded that PAHs with larger electronic spatial extent and lower molecular total energy values tend to be more hydrophobic and lipophilic.

## Acknowledgements

The work was supported by the National Natural Science Foundation of China (No. 40730741), the Guangdong Provincial Natural Science Foundation (No. 05103552), the State Scholarship Fund of China Scholarship Council (No. 2007103564) and the Doctorate Foundation of South China University of Technology.

## References


[1] N.T. Edwards, J. Environ. Qual. 12, 427 (1983)
[2] C.E. Cerniglia, Biodegradation 3, 351 (1992)
[3] K.C. Jones, J.A. Stratford, K.S. Waterhouse, E.T. Furlong, W. Giger, R.A. Hites, C. Schaffner, A.E. Johnston, Environ. Sci. Technol. 23, 95 (1989)
[4] D.J. Freeman, F.C.R. Cattell, Environ. Sci. Technol. 24, 1581 (1990)
[5] W. Lijinsky, Mutat. Res. 259, 251 (1991)
[6] J.P. Meador, J.E. Stein, W.L. Reichert, U. Varanasi, Rev. Environ.Contam. Toxicol. 143, 79 (1995)
[7] A. Stella, M.T. Piccardo, R. Coradeghini, A. Redaelli, S. Lanteri, C. Armanino, F. Valerio, Anal. Chim. Acta 461, 201 (2002)
[8] S.K. Samanta, O.V. Singh, R.K. Jain, Trends Biotech. 20, 243 (2002)
[9] X.Q. Tao, G.N. Lu, Z. Dang, C. Yang, X.Y. Yi, Process Biochem. 42, 401 (2007)
[10] J. Jacob, Pure Appl. Chem. 68, 301 (1996)
[11] D. Mackay, D. Callcot, Partitioning and physical properties of PAHs. In: A.H. Neilson, (Ed.), The handbook of environmental chemistry, Vol. 3, Part J. PAHs and related compounds. (Springer, Berlin, 325, 1998)
[12] M.M.C. Ferreira, Chemosphere, 44, 125 (2001)
[13] F.A.D. Ribeiro, M.M.C. Ferreira, J. Mol. Struc. Theochem 663, 109 (2003)
[14] A. Sabljić, H. Güsten, H. Verhaar, J. Hermens, Chemosphere 31, 4489 (1995)
[15] M.T.D. Cronin, D.J. Livingstone, Predicting chemical toxicity and fate. (CRC Press LLC, Baca Raton, Florida, 2004)
[16] W. Zhou, Z.C. Zhai, Z.Y. Wang, L.S. Wang, J. Mol. Struc. Theochem 755, 137 (2005)
[17] W. Chu, K.H. Chan, Sci. Total Environ. 248, 1 (2000)
[18] C.M. Auer, M. Zeeman, J.V. Nabholz, R.G.. Clements, SAR QSAR Environ. Res. 2, 29 (1994)
[19] S.C. Basak, B.D. Gute, G.D. Grunwald, J. Chem. Inf. Comp. Sci. 36, 1054 (1996)
[20] J. Huuskonen, J. Chem. Inf. Comp. Sci. 40, 773 (2000)
[21] J.W. Chen, X.Y. Xue, K.W. Schramm, M. Quan, F.L. Yang, A. Kettrup, Comput. Biol. Chem. 27, 165 (2003)
[22] M.T. Sacan, M. Ozkul, S.S. Erdem, SAR QSAR Environ. Res. 16, 443 (2005)
[23] H.X. Zhao, Q. Zhang, J.P. Chen, X.Y. Xue, X.M. Liang, Chemosphere 59, 1421(2005)
[24] J.W. Zou, Y.J. Jiang, G.X. Hu, M. Zeng, S.L. Zhuang, Q.S.Yu, Acta Phys. Chim. Sin. 21, 267 (2005)
[25] G.Y. Yang, X.C. Zhang, Z.Y. Wang, H.X. Liu, X.H. Ju, J. Mol. Struc. Theochem 766, 25 (2006)
[26] X.C. Zhang, J. Yu, Z.Y. Wang, H.X. Liu, Chinese J. Struc. Chem. 25, 823 (2006)
[27] P. Gramatica, M. Corradi, V. Consonni, Chemosphere 41, 763 (2000)
[28] S.C. Basak, D. Mills, D.M. Hawkins, H.A. El-Masri, SAR QSAR Environ. Res. 13, 649 (2002)
[29] G.N. Lu, Z.Dang, X.Q. Tao, X.P. Chen, X.Y. Yi, C. Yang, QSAR Comb. Sci. 26, 182 (2007)
[30] G.N. Lu, Z. Dang, X.Q. Tao, C. Yang, X.Y. Yi, Sci. Total Environ. 373, 289 (2007)
[31] S.C. Basak, D. Mills, Arkivoc (ii), 60 (2005)
[32] P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, Chem. Phys. Lett. 323, 59 (2000)
[33] S. Trohalaki, R. Pachter, SAR QSAR Environ. Res. 14, 131 (2003)
[34] S. Trohalaki, E. Gifford, R. Pachter, Comput. Chem. 24, 421 (2000)
[35] G.N. Lu, Z. Dang, X.Q. Tao, P.A. Peng, D.C. Zhang, J. Theor. Comp. Chem. 4, 811 (2005)
[36] Z.Y. Wang, Z.C. Zhai, L.S. Wang, QSAR Comb. Sci. 24, 211 (2005)
[37] J.W. Gao, X.Y. Wang, X.B. Li, X.L. Yu, H.L. Wang, J. Mol. Model. 12, 513 (2006)
[38] J.W. Gao, X.Y. Wang, X.L. Yu, X.B. Li, H.L. J. Mol. Model. 12, 521 (2006)
[39] M. J. Frisch et al., Gaussian 03, Revision B.01, Gaussian, Inc. Pittsburgh PA, 2003.
[40] D. Mackay, W.Y. Shiu, K.C. Ma, Illustrated handbook of physical-chemical properties and environmental fate for organic chemicals, vol. 3, (Lewis, London, 1998)
[41] A.D. Becke, J. Chem. Phys. 98, 5648 (1993)
[42] W. J. Hehre, R. Ditchfield, J. A. Pople, J. Chem. Phys. 56, 2257 (1972)
[43] S. Wold, M. Sjöström, L. Eriksson, Chemometr. Intell. Lab. Syst. 58, 109 (2001)
[44] R.G. Pearson, P. Natl. Acad. Sci. USA 83, 8440 (1986)
[45] G..D. Veith, O.G. Mekenyan, G.T. Ankley, D.J. Call, Chemosphere 30, 2129 (1995)
[46] J.W. Chen, L.R. Kong, C.M. Zhu, Q.G. Huang, L.S. Wang, Chemosphere 33, 1143 (1996)
[47] J.W. Chen, W.J.G.M. Peijnenburg, X. Quan, Y.Z. Zhao, D.M. Xue, F.L. Yang, Chemosphere 37, 1169 (1998)
[48] J.W. Chen, W.J.G.M. Peijnenburg, X. Quan, S. Chen, D. Martens, K.W. Schramm, A. Kettrup, Environ. Pollut. 114, 137 (2001)
[49] Umetrics. User's Guide to SIMCA-P, SIMCA-P+, Version 10.0. (Umetrics AB, Umeå, Sweden, 2002)

[50] J.W. Chen, P. Yang, S. Chen, X. Quan, X. Yuan, K.W. Schramm, A. Kettrup, SAR QSAR Environ. Res. 14, 97 (2003)
[51] J.F. Niu, L.P. Huang, J.W. Chen, G. Yu, K.W. Schramm, Chemosphere 58, 917 (2005)
[52] G.N. Lu, Z. Dang, X.Q. Tao, C. Yang, X.Y. Yi, QSAR Comb. Sci. DOI: 10.1002/qsar.200710014
[53] J.C. Faucon, R. Bureau, J. Faisant, F. Briens, S. Rault, Chemosphere 38, 3261 (1999)
[54] O.G. Mekenyan, G.T. Ankley, G.D. Veith, D.J. Call, SAR QSAR Environ. Res. 4, 139 (1995)