QSPR models for predicting generator-column-derived octanol/water and octanol/air partition coefficients of polychlorinated biphenyls

Jintao Yuan ^{a, c}, Shuling Yu ^b, Ting Zhang ^c, Xuejie Yuan ^d, Yunyuan Cao ^a, Xingchen Yu ^a, Xuan Yang ^a, Wu Yao ^{a, *}

^a School of Public Health, Zhengzhou University, Zhengzhou, 450001, China

^b Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province, Henan University, Kaifeng, 475004, China

^c Key Laboratory of Environmental Medicine Engineering of Ministry of Education, Southeast University, Nanjing 210009, China

^d Shangqiu Medical College, Shangqiu, Henan Province, 476100, China

* Corresponding author: School of Public Health, Zhengzhou University, 100 Kexuedaodao, Zhengzhou, China, 450001. Tel.: +86-371-67781922; fax: +86-371-67781922. E-mail address: yaowu@zzu.edu.cn.

Abstract

Octanol/water (K_{OW}) and octanol/air (K_{OA}) partition coefficients are two important physicochemical properties of organic substances. In current practice, K_{OW} and K_{OA} values of some polychlorinated biphenyls (PCBs) are measured using generator column method. Quantitative structure–property relationship (QSPR) models can serve as a valuable alternative method of replacing or reducing experimental steps in the determination of K_{OW} and K_{OA} . In this paper, two different methods, i.e., multiple linear regression based on dragon descriptors and hologram quantitative structure–activity relationship, were used to predict generator-column-derived log K_{OW} and log K_{OA} values of PCBs. The predictive ability of the developed models was validated using a test set, and the performances of all generated models were compared with those of three previously reported models. All results indicated that the proposed models were robust and satisfactory and can thus be used as alternative models for the rapid assessment of the K_{OW} and K_{OA} of PCBs.

Keywords: Quantitative structure–property relationship; Polychlorinated biphenyls; Octanol/water partition coefficient; Octanol/air partition coefficient; Hologram quantitative structure–activity relationship

38 Abbreviations

39 QSPR, quantitative structure property relationship; PCBs, polychlorinated biphenyls;
40 K_{OW} , octanol/water partition coefficient; K_{OA} , octanol/air partition coefficient; MLR,
41 multiple linear regression; HQSAR, hologram quantitative structure–activity
42 relationship; OPS, ordered predictor selection; A, atomic numbers; B, bond types; C,
43 atomic connections; H, hydrogen; Ch, chirality; DA, donor and acceptor; HL,
44 hologram length; LVs, latent variables; HL, hologram length; SEE, standard error of
45 estimate; LOO, leave-one-out; RMSE, root mean square error.

46

47

1. Introduction

Polychlorinated Biphenyls (PCBs) are highly stable industrial chemicals used for various commercial applications, such as in heat transfer fluids, diluents, plasticizers, adhesives, pesticides, and lubricants (Hansen et al., 1999; de Melo, 2012). However, improper disposal practices and accidental releases of PCBs together with polychlorinated dibenzodioxins and dibenzofurans have made them environmental contaminants that are currently included in many blacklists, such as that of persistent organic pollutants and suspected environmental endocrine disruptors. Theoretically, a total of 209 PCBs exist, whereas approximately 150 PCBs are found in the environment (Zhang et al., 2013). PCBs entering the natural ecosystems are distributed into the solid, liquid, or gaseous phases of environmental compartments (Gawlik et al., 1997; Ross et al., 2004). In the study of the environmental fate of organic chemicals, the octanol/water partition coefficient (K_{OW}) and octanol/air partition coefficient (K_{OA}) are considered as key parameters. K_{OW} is correlated with water solubility, soil/sediment sorption coefficient, and bioconcentration. Furthermore, K_{OW} is proven to be a valuable tool in many areas of the natural sciences, including chemistry, biochemistry, biology, pharmacology, and environmental sciences. On the other hand, K_{OA} can be used to explain movement of compounds, and is applied in several cases involving the partitioning of organic substances in the atmosphere to solid or liquid phases (Hippelein and McLachlan, 1998). Because experimental K_{OW} and K_{OA} values are not available for all PCBs, QSPR (Quantitative Structure Property Relationship) models for predicting K_{OW} and K_{OA} values of PCBs were studied in this

present work.

Values for K_{OW} can be obtained by several direct measurements, such as the shake flask, slow-stirring, and generator column method. Direct measurement of K_{OW} with shake flask are considered less accurate because of the formation of colloidal dispersions, volatilization to the atmosphere, and adsorption of solute onto surfaces of transfer vessels (Woodburn et al., 1984). In the slow-stirring method, water and octanol phases are equilibrated under conditions of slow stirring, which may take up to two to four days to achieve equilibration. A rapid, precise, coupled-column system, in particular, the generator column method, can obviate many experimental difficulties, such as colloidal suspensions and solute absorption. Therefore, the log K_{OW} values of a greater number of PCBs that are measurable by the generator column method were superior to those of the other two direct measurements. The K_{OW} values of 73 PCBs used in the current work were determined using the generator column method. Values for K_{OA} can also be obtained through the generator column method; K_{OA} values of 19 PCBs have been reported by Harner and Bidleman (1996).

Numerous QSPR-based methods for predicting log K_{OW} values of PCBs have been reported (Hawker and Connell, 1988; Kamlet et al., 1988; Sabljic et al., 1993; Makino, et al., 1999; Hansen et al., 1999; Yeh and Hong, 2002; Wang et al., 2003; Zhou et al., 2005; Padmanabhan et al., 2006; Lü et al., 2007), and more recently Han et al. (2012) studied the relationship between K_{OW} and chromatographic retention of reversed-phase high performance liquid chromatography (RP-HPLC). Zhang et al. (2013) also reported a QSPR model for log K_{OW} of all PCBs by density functional

theory. Evidently, many studies explored K_{OW} determination through various means and modifications. In contrast, only a few models were generated to predict the log K_{OA} values of PCBs. In particular, one study (Chen et al. 2002) used quantum chemical descriptors to predict the log K_{OA} values obtained by the generator column method.

Dragon descriptors are widely used to develop numerous QSAR models, ranging from predicting physical-chemical properties to pharmacological and toxicological activities (Gharagheizi and Mehrpooya, 2008; García et al., 2011; Ertürk et al., 2012).

Hologram quantitative structure–activity relationship (HQSAR) is a modern 2D QSAR approach based on specialized molecular fragments. In this method, each molecule is broken down into several unique structural fragments, which are then arranged to form a molecular hologram. This method requires no explicit 3D information (e.g., determination of 3D structures, putative binding conformations, and molecular structural alignment) and avoids the calculation and selection of physicochemical descriptors. Only 2D structures of the data set compounds and the corresponding property values are required for the HQSAR application (Salum and Andricopulo, 2009; Leal et al., 2015). Therefore, Dragon-based QSPR and HQSAR were used in this study.

The study aims to generate QSPR models for predicting generator column method derived log K_{OW} and log K_{OA} values of PCBs. Multivariable linear regression (MLR) and HQSAR models were developed to predict log K_{OW} and log K_{OA} values of PCBs, respectively. Furthermore, a comparison among different models was

114 performed. Finally, the models were used to predict $\log K_{OW}$ and $\log K_{OA}$ values of

115 PCBs with unknown experimental measurements.

116

117

118

2. Materials and Methods

2.1. Data set

The log K_{OW} values of 73 PCBs were taken from the reports of Miller et al. (1984), Woodburn et al. (1984), Doucette and Andren (1987), Larsen et al. (1992), Hawker and Connell (1988), Li & Doucette (1993), and Yeh and Hong (2002). For multiple log K_{OW} values for the same PCB measured in different studies, the average log K_{OW} value was calculated (Table S1 in the Supplementary material). The log K_{OA} values of 19 PCBs were collected from Harner and Bidleman (1996) and Chen et al. (2002). Three-dimensional structures of 209 PCBs were constructed using SYBYL-X 1.3 program. Energy minimizations were performed using Tripos force field (SYBYL-X 1.3 Tripos Inc.) followed by the conjugate gradient method with convergence criterion of 0.001 kcal/mol. The Gasteiger Huckel charge was assigned to each PCB. The log K_{OW} and log K_{OA} values determined from aforementioned methods are presented in Table 1. The log K_{OW} values of the 73 PCBs range from 4.42 to 8.22. The Kennard and Stone (KS) method is the best known among uniform designs in chemometrics. It can select a subset of n -dimensional objects, ‘uniformly’ distributed in the experimental space. (Daszykowski et al., 2002; Martin et al., 2012) Thus, the entire data set was split into two groups, namely, a training set (55 PCBs) and a test set (18 PCBs), by the KS method. The log K_{OA} values of the 19 PCBs range from 7.01 to 10.75. The 19 PCBs were also divided into a training set (15 PCBs) and a test set (4 PCBs) by KS. The IUPAC numbers in the training and test sets are listed in Table S2 (Supplementary material).

(Table 1 here)

2.2. DRAGON molecular descriptors

To obtain molecular descriptors, the optimized molecular structures were inputted into the DRAGON program, in which 4885 molecular descriptors of different types were calculated to describe the chemicals' structural diversity. Afterwards, the initial pool of descriptors was reduced by applying the DRAGON built-in variable exclusion procedure. Three types of descriptors were excluded: constant descriptors (relative standard deviation < 0.001), near-constant descriptors (all values are equal except one), and highly correlated descriptors (i.e., for each pair of highly correlated descriptors, $R > 0.9$, the descriptor that has the largest mean correlation coefficient with the rest of the descriptors was removed). After exclusion, 255 molecular descriptors were retained. To further reduce redundancy and multicollinearity in the 255 molecular descriptors, the variable reduction Wootton, Sergent, Phan-Tan-Luu's algorithm (V-WSP) was used (Ballabio et al., 2014). Overall, 130 molecular descriptors were retained in this work.

2.3. Ordered predictor selection procedure

Ordered predictor selection (OPS) (Teófilo et al., 2009) is an iterative algorithm for variable selection, the core of this method is to sort the variables from an informative vector, followed by a systematic investigation of partial least square (PLS) regression models with the aim of finding the most relevant set of variables by comparing the cross-validation parameters of the models obtained. First, an

informative vector, which contains information about the location of the best descriptors for prediction, was constructed. Second, the original descriptors were differentiated according to corresponding absolute values. Third, the absolute values were sorted in descending order. Fourth, multivariate regression models were generated and evaluated using a cross validation strategy. An initial subset of variables (window) was selected to build and evaluate the first model. Then, this matrix was expanded by the addition of a fixed number of variables (increment) and a new model was again constructed and evaluated. New increments were added until all or some percentage of variables were accounted for. Quality parameters of the models were obtained for every evaluation and stored for comparison. Lastly, the descriptors contained in the model with the best quality parameters were selected as the best descriptor subset. The best descriptor subset was used as the independent variable of MLR to establish the QSPR models. The OPS was implemented in MATLAB 7, and is available on the internet at <http://Iqta.igim.unicamp.br>.

2.4. Variation inflation factor

Variation inflation factors (VIF) can be used to evaluate the correlation degree of each independent variable in the equation (Zhou et al., 2005). If $VIF = 1.0$, no intercorrelation exists for each variable; if VIF ranges from 1.0 to 5.0, the related equation is acceptable; and if VIF is larger than 10.0, the regression equation is unstable and a recheck is necessary. VIF was calculated using SPSS 18.0 software.

2.5. Hologram QSAR

Statistical HQSAR modeling was performed in the SYBYL-X 2.1. HQSAR

analysis involves three main steps: the generation of sub-structural fragments for each of the molecules in the training set, the encoding of these fragments in holograms, and the correlation of the latter with the log K_{OW} or log K_{OA} data using PLS. A number of parameters concerning hologram generation, such as hologram length, fragment size, and fragment distinction, prevailingly affect the HQSAR model quality. During hologram generation, certain fragment distinction parameters were considered. These include atomic numbers (A), to distinguish the atom types; bond types (B), to distinguish the bond orders; atomic connections (C), to consider the atomic hybridization states within fragments; hydrogen (H), to identify whether or not hydrogen atoms are included; chirality (Ch), to distinguish fragments based on atom chirality and bond stereochemistry; and donor and acceptor (DA), to initiate the search for defined donor and acceptor atoms in a fragment. Cross-validation was used to establish the optimal number of latent variables (LVs), hologram length (HL), and to assess the robustness and predictive power of the model. HQSAR models can also be graphically displayed as color-coded structure diagrams (contribution maps) in which the colors represent positive (yellow-to-green), neutral (white), and negative (orange-to-red) contribution to potency variation. Additional details on HQSAR are easily found in literature (Moda et al., 2007; Yu et al., 2015; Madhavan et al., 2013).

2.6. Model validation

In this study, statistical parameters, such as squared correlation coefficient r^2 , standard error of estimate (SEE), cross-validated squared correlation coefficient q^2 by leave-one-out (LOO) cross-validation, cross-validated standard error of prediction

(SEP), predictive correlation coefficient r^2_{pred} , and root mean square error (RMSE) were used for model performance characterization. The parameters r^2 and SEE were used to estimate the degrees of overall correlation of the derived model. On the other hand, q^2 and SEP were used to assess the robustness of built models. Lastly, the r^2_{pred} and RMSE obtained from the external test set were used to evaluate the predictive power of the developed models. Other external validation parameters (Q^2_{F1} , Q^2_{F2} , Q^2_{F3} , CCC , k , k' , $(r^2 - r_0^2)/r^2$, $(r^2 - r_0^2)/r^2$) proposed in literatures (Chirico et al., 2012; Golbraikh and Tropsha, 2002) for the external validation of QSPR models were also used in this study. In order to determine the best prediction model for the overall dataset, a modified r^2 parameter (r_m^2) was calculated for the entire dataset. These parameters were defined in the Supplementary material.

3. Results and Discussion

3.1. MLR models for log K_{OW}

One of main assumptions of MLR is normality (Bolboacă and Jäntschi, 2013). The distribution of log K_{OW} in the training set was tested. The null hypothesis of normality was accepted (Kolmogorov-Smirnov p -value = 0.964). Thus, MLR research was performed by following the above-mentioned process, four DRAGON descriptors were selected by OPS; the selected descriptors and relevant statistic parameters are shown in Table S4. An MLR model including the four descriptors was then generated. The results are shown as follows:

$$\log K_{OW} = 0.026 \text{ RDF040m} + 0.724 \text{ Mor17u} + 0.528 \text{ Mor12u} - 0.916 \text{ Mor14u} + 5.575 \quad (1)$$

where $r^2 = 0.923$, $SEE = 0.241$, $q^2 = 0.877$, $SEP = 0.291$, $F = 149.204$, $r^2_{\text{pred}} = 0.860$, and $RMSE = 0.241$.

Moreover, multicollinearity between the descriptors of the Eq. (1) model was tested by their VIFs. The VIF, as well as standard regression coefficients (SR) and t-scores, are listed in Table S5. Evidently, the VIF of descriptor RDF040m is larger than 10.0. Therefore, a forward stepwise procedure was carried out to select the four descriptors. Finally, an MLR model with two descriptors was obtained for predicting the log K_{OW} of PCBs as shown in Eq. (2):

$$\log K_{OW} = 0.040 \text{ RDF040m} - 1.254 \text{ Mor14u} + 4.785 \quad (2)$$

where $r^2 = 0.917$, $SEE = 0.244$, $q^2 = 0.902$, $SEP = 0.261$, $F = 288.124$, $r^2_{\text{pred}} = 0.857$, and $RMSE = 0.253$.

Similarly, the VIF, SR, and t-scores of descriptors in Eq. (2) were also calculated (in Table S5). The VIF values of two descriptors in Eq. (2) were reduced to 2.451. Thus, the multicollinearity of descriptors was eliminated.

The predictivity of the Eq. (2) was checked by the test set. The r^2_{pred} ($r^2_{\text{pred}} = 0.857$) showed satisfactory predictive ability. Several external validation parameters were also used to validate the Eq. (2). Clearly, all performance statistics of Eq. (2) satisfied the criteria in Table S3 (The parameters were shown in Table S6). For the overall dataset, the r^2_m ($r^2_m = 0.846$) showed satisfactory predictive ability, too. The relationship of the observed versus the predicted log K_{OW} values are shown in Fig. 1A. The distribution of residuals for 73 PCBs was shown in Figure S1. The predicted log K_{OW} values are listed in Table 1.

(Figure 1 here)

3.2. HQSAR models for log K_{OW}

The performance of the HQSAR model can be optimized by varying parameters. To determine the best fragment distinction parameters, hologram length and fragment size were set to default values. Particularly, hologram lengths were set to 97, 151, 199, 257, 307, and 353 bins, while the fragment size were set to 4–7. The statistical results of the PLS analyses of the 55 compounds in the training set using different fragment distinction combinations are summarized in Table 2.

(Table 2 here)

In Table 2, the exclusive fragment distinction B term HQSAR model yielded the highest r^2 ($r^2 = 0.903$), which indicates that the fragment distinction B term is

important in predicting the log K_{OW} . As the other fragment distinction terms were combined with B, the fragment distinction A/B term model yielded the highest r^2 ($r^2 = 0.907$) and lowest SEE (SEE = 0.264) with four LVs, and yielded very high r_m^2 ($r_m^2 = 0.846$) for the whole dataset. Thus, the fragment distinction A/B was selected.

After the fragment distribution parameter was identified, fragment size parameters were explored. Results of the models for predicting log K_{OW} are shown in Table S7. Apparently, the r^2 value was initially slightly increased and then decreased with increasing fragment size. At fragment sizes 4–7 or 5–8, the models yielded the best r^2 and lowest SEE. Because the q^2 (which indicate the robustness) of the model with fragment size 4–7 is superior to that of the model with fragment size 5–8, and the model with fragment size 4–7 obtained the highest high r_m^2 ($r_m^2 = 0.846$) for the whole dataset, thus, the fragment size 4–7 was selected. Thus, the HQSAR 17 (shown in bold) was considered as the best HQSAR model for predicting log K_{OW} of PCBs (Table 1). The corresponding graph of observed versus predicted pEC50 values for the training and test set compounds by HQSAR 17 model is depicted in Fig. 1B; the visual correlation between experimental and predicted values for the training and test sets was also satisfactory. Values for r^2 and RMSE were 0.912 and 0.212 for the test set, respectively. To further assess the predictive capability of the best HQSAR model, external validation was also conducted. All parameters of the HQSAR 17 satisfied the strict criterion in Table S3. Based on these findings, HQSAR 17 model can be considered to have good fit, robustness, and good predictive abilities. The individual atomic contributions of three PCBs, according to the HQSAR 17, are displayed in

Figure S2. The HQSAR contribution maps reveal that the Cl substitution position is important, this is because the change of substitution pattern will result in the change of the dispersive force and polarizability of PCBs. Usually molecules with large polarizability are easy to enter octanol phase, therefore, resulting in relatively high log K_{OW} value (Han et al., 2006).

3.3. MLR models for log K_{OA}

The first step of MLR was to investigate the distribution of the experimental log K_{OA} values in the training set. The normality distribution of log K_{OA} in the training set was verified by the Kolmogorov-Smirnov test (p -value = 0.857). Then, different numbers of descriptors were selected by OPS. The chosen descriptors and relevant statistic parameters are shown in Table S8 (in the Supplementary materials). An evident enhancement in q^2 and a reduction in SEP was easily noted after increasing the number of descriptors to four. The model, with the inclusion of the four descriptors, are shown as follows:

$$\log K_{OA} = -0.223 \text{ SM11_AEA(bo)} + 4.593 \text{ TI2_L} - 1.066 \text{ Mor03u} - 0.010 \text{ MW} - 6.156 \quad (3)$$

where $r^2 = 0.989$, $\text{SEE} = 0.135$, $q^2 = 0.970$, $\text{SEP} = 0.191$, $F = 231.348$, $r^2_{\text{pred}} = 0.899$, and $\text{RMSE} = 0.314$.

The VIF, SR, and t-scores of descriptors in Eq. (3) are summarized in Table S9. The VIF value of descriptor SM11_AEA(bo) is higher than 10, which indicates that Eq. (3) is unstable. To overcome this problem, the four descriptors were again

selected through a forward stepwise procedure. Finally, an MLR model containing two descriptors was developed for predicting the log K_{OA} of PCBs.

$$\log K_{OA} = 4.066 \text{ SM11_AEA(bo)} - 1.022 \text{ Mor03u} - 2.008 \quad (4)$$

where $r^2 = 0.944$, $SEE = 0.281$, $q^2 = 0.873$, $SEP = 0.401$, $F = 8.882$, $r^2_{\text{pred}} = 0.958$, and $RMSE = 0.293$.

The r^2 of Eq. (4) was 0.944, which indicates satisfactory fitting power. For the test set, Eq. (4) model shows the following values: $r^2_{\text{pred}} = 0.958$, $Q^2_{F1} = 0.965$, $Q^2_{F2} = 0.950$, $Q^2_{F3} = 0.980$, $CCC = 0.973$, $k = 0.996$, $k' = 1.004$, $r^2_0 = 0.928$, and $r'^2_0 = 0.940$. All external validation parameter values of the Eq. (4) model satisfied the accepted condition. These findings indicate that Eq. (4) has good predictive power. For the overall dataset, the r^2_m ($r^2_m = 0.892$) showed satisfactory predictive ability, too. The prediction values of the training and test sets are shown in Table 1, and the corresponding graph of observed versus predicted log K_{OA} values for the training and test set PCBs obtained using the Eq. (4) model is depicted in Fig. 1C.

3.4. HQSAR models for log K_{OA}

The parameters of the HQSAR model were optimized to improve the accuracy of prediction. To determine the best fragment distinction parameters, hologram lengths and fragment sizes were set to default values. A series of HQSAR models was generated by a single or a combination of two or more of the parameters (A, B, Co, H, Ch, and DA), whereas the fragment size was set to 4–7. Results are shown in Table 3. Table 3 shows that, although the HQSAR 22 and 24 models both show the highest r^2

values, the q^2 and r_m^2 of the HQSAR 22 model is superior to that of the HQSAR 24 model. Thus, fragment distinction A was selected. To obtain the best statistical values, the same fragment distinction (A) and various fragment sizes (1–11) were used to yield different models (Table S10). As indicated in Table S10, the best model, HQSAR 37 (shown in bold), was generated based on the fragment size of 3–6 with the highest r_m^2 of 0.979 for the overall dataset. The model showed good predictability on a test set of 4 molecules ($r_{\text{pred}}^2 = 0.989$). The statistical parameters satisfied the recommended condition by the literatures (Chirico et al., 2012; Golbraikh and Tropsha, 2002) for the test set. The statistical parameters are listed in Table S6. Finally, the HQSAR 37 model was used to predict the $\log K_{\text{OA}}$ of PCBs in the training and test sets, and the 2D scatter plot of observed versus predicted $\log K_{\text{OA}}$ values is shown in Fig. 1D. The $\log K_{\text{OA}}$ of 209 PCBs was predicted by the HQSAR 37 model (in Table 1). The individual atomic contributions maps of three compounds, resulting from the best HQSAR model, are displayed in Figure S3. Obviously, atomic contributions are changed along Cl substitution pattern. One possible reason is the number and position of chlorine atom lead to the change of molar mass and electronic distribution.

Additional, the residuals from the developed four models were checked, too. The plotted points from HQSAR 17 and 37 models are closer to the straight line (zero) than ones from the two models based on DRAGON descriptors (Figure S1). Since all the $\log K_{\text{OW}}$ and $\log K_{\text{OA}}$ values have been predicted, we can easily see that PCB-2, PCB-3, and PCB-3 have a low $\log K_{\text{OW}} < 5$ but a high $\log K_{\text{OA}} \geq 6$ and can exhibit

biomagnify in air-breathing organisms (Kelly, et al., 2007). Some PCBs in the range $6.5 < \log K_{OA} < 10$ may accumulate in the Arctic (Wania, 2003).

(Table 3 here)

3.5 Mechanistic interpretation of the descriptors in the QSPR model

The descriptor RDF040m is a radial distribution function (RDF) descriptor and is calculated at a radius of 4.0 Å from the geometric centers of each molecule to represent the 3D molecular distribution of mass (Hemmer et al., 1999). The descriptor Mor14u (3D MoRSE signal 14 unweighted) is a 3D-MoRSE (Molecule Representation of Structures based on Electron diffraction) descriptor. In Eq. (2), RDF040m showed higher t-score (9.382) and SR (0.586) values than Mor14u, which indicates that RDF040m has a larger influence on the log K_{ow} value. RDF040m is calculated at a radius of 4.0 Å from the geometric centers of each molecule. Thus, the descriptor RDF040m is highly sensitive to chlorine substitutions at the ortho, meta, or para positions, and highly sensitive to mono-, di-, or multi-substitution. The HQSAR 17 with A/B as fragment distinction suggests that atom type and bond order in fragment size 4–7 is important to the prediction of log K_{ow} of PCBs (Cui et al., 2003). Therefore, the position of chlorine substitution influences the value for log K_{ow} . The findings also show that Eq. (2) and HQSAR 17 models are consistent with the reported model by Yeh and Hong (2002).

The descriptor SM11_AEA(bo) in Eq. (4) is the spectral moment of order 11 from the augmented edge adjacency matrix weighted by bond order. The descriptor

Mor03u (3D-MoRSE — signal 03/unweighted) is a 3D-MoRSE descriptor. In Eq. (4), the SR and t-score of the descriptor SM11_AEA(bo) are higher than those of the descriptor Mor03u. This observation indicates that descriptor SM11_AEA(bo) possesses a greater influence on $\log K_{OA}$. The descriptor SM11_AEA(bo) is related to bond order. Bond order is a property considered as intimately related with molecular geometry; it is determined by charge distribution in the neighboring atoms, and by the character of their common molecular orbitals. This quality explains the relationship of the chlorine substitution position in PCB to $\log K_{OA}$. Moreover, the HQSAR 37 model with A as a fragment distinction parameter reveals that the atom type in fragment size 3–6 is also related to $\log K_{OA}$.

Some background information on the selected molecular descriptors has been provided in the Supplementary material. The values of the selected DRAGON molecular descriptors have been provided in Table S11.

3.6. Comparison with the results from other models

To further evaluate the quality of models, a comparison between these presented models and the reported models by Hawker and Connell (1988), Yeh and Hong (2002), and Chen et al. (2002) was undertaken. These models were previously developed to predict generator column measured $\log K_{OW}$ or $\log K_{OA}$ values. Hawker and Connell (1988) predicted generator-column-derived $\log K_{OW}$ values of 46 PCBs using total surface areas (TSA) descriptor, and obtained a simple model: $\log K_{OW} = (3.41 \times 10^{-2})TSA - 2.20$. Yeh and Hong (2002) related the chlorine numbers (N_{cl}) and the number of ortho chlorine atoms (N_{or}) to the $\log K_{OW}$ values of 46 PCBs measured

by the generator column method, and developed a model with two variables: $\log K_{OW}$
 $= 4.35 + 0.453 N_{cl} - 0.164 N_{or}$. Chen et al. (2002) used a PLS model with nine
quantum chemical descriptors to predict the $\log K_{OA}$ values obtained by the generator
column method. Two important statistical parameters (r^2 and r^2_{pred}) were compared
and shown in Table 4. The r^2 of DRAGON and HQSAR models are both inferior to
those reported by Yeh and Hong (2002), and Hawker and Connell, (1988). Although
the r^2 of the best HQSAR model was inferior to corresponding value from Eq. (2), the
 r^2_{pred} of the best HQSAR model was superior to that of Eq. (2). To check the statistical
significance of these differences of models, a paired t -test was applied in 46 out of 73
cases, where the reported model by Yeh and Hong, (2002) gives the best fit, this
difference is statistically significant ($p < 0.05$). For prediction of $\log K_{OA}$ values, the
 r^2 of the best HQSAR model is superior to that obtained by Chen et al. (2002),
whereas the r^2 of DRAGON model is inferior to that reported by the same study. For
the predicted $\log K_{OA}$ values of all 19 compounds, a paired t -test for the Eq. (4) and
the reported model by Chen et al., (2002) with a confidence interval of 95%, resulted
in a two-tailed p -value of 0.905, confirming that the two models are no significant
difference. Evaluation of the 19 predicted $\log K_{OA}$ values by HQSAR 37 and the
model from Chen et al., (2002) with a paired t -test, demonstrated that two models
showed no significant differences ($p = 0.771$). Therefore, the fit and predictive power
of the most effective HQSAR and DRAGON models generated from this work are
considered acceptable.

(Table 4 here)

4. Conclusion

DRAGON-based MLR and HQSAR methods were successfully used to investigate the relationship between some PCBs and their generator-column-derived $\log K_{OW}$ and $\log K_{OA}$. As a result, two models based on DRAGON descriptors and two HQSAR models were generated and have demonstrated relatively satisfactory fitting and external prediction ability. Furthermore, the constructed models indicated that certain molecular details are important for prediction of $\log K_{OW}$ and $\log K_{OA}$. The comparison of models revealed that DRAGON descriptors and HQSAR are suited to predict the $\log K_{OW}$ and $\log K_{OA}$ of PCBs. All results showed that aside from a consistent and reproducible prediction, HQSAR can also be applied to quickly and conveniently derive QSPR models. HQSAR can thus serve as a promising tool for predictive use in environmental health and ecological risk assessment for organic chemicals, such as PCBs.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2014M550387), Postdoctoral Research Sponsorship in Henan Province (Grant No. 2013014), the Joint Fund of National Natural Science Foundation of China and Henan Province (U1404508), National Natural Science Foundation of China (NSFC, Nos. 81273039 and 81273122), Scientific and Technological Key Project Education Department of Henan Province (Grant No. 14B350005), National Science Foundation for Young Scientists of china (Grant No. 81302461), Scientific Research Foundation

440 of Henan University (Grant No. 2013YBZR039) and the open project of Key
441 Laboratory of Environmental Medicine Engineering of Ministry of Education in
442 Southeast University.

443

444

445

References

- Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., Todeschini, R., 2014. A novel variable reduction method adapted from space-filling designs. *Chemom. Intell. Lab. Syst.* 136, 147–154.
- Bolboacă, S.D., Jäntschi L., 2013. Quantitative structure-activity relationships: linear regression modelling and validation strategies by example. *Biomath.* 2, 1–11.
- Chen, J., Xue, X., Schramm, K.W., Quan, X., Yang, F., Kettrup, A., 2002. Quantitative structure–property relationships for octanol–air partition coefficients of polychlorinated biphenyls. *Chemosphere* 2002, 48, 535–544.
- Chirico, N., Gramatica, P., 2012. Real external predictivity of QSAR models. part 2. new intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model* 52, 2044–2058.
- Cui S., Wang X., Liu S., Wang L., 2003. Predicting toxicity of benzene derivatives by molecular hologram derived quantitative structure–activity relationships (QSARS), *SAR QSAR Environ. Res.* 14, 223–231.
- Daszykowski M., Walczak B., Massart D.L., 2002. Representative subset selection. *Anal. Chim. Acta* 468, 91–103.

468

469 de Melo, E.B., 2012. A new quantitative structure–property relationship model to
470 predict bioconcentration factors of polychlorinated biphenyls (PCBs) in fishes
471 using E-state index and topological descriptors. *Ecotoxicol. Environ. Saf.* 75,
472 213–222.

473

474 Doucette, W.J., Andren, A.W., 1987. Correlation of octanol/water partition
475 coefficients and total molecular surface area for highly hydrophobic aromatic
476 compounds. *Environ. Sci. Technol.* 21, 521–524.

477

478 Ertürk M.D., Saçan M.T., Novicb M., Minovski N., 2012. Quantitative
479 structure–activity relationships (QSARs) using the novel marine algal toxicity
480 data of phenols, *J. Mol. Graph. Model.* 38, 90–100.

481

482 García I., Fall Y., García-Mera X., Prado-Prado F., 2011. Theoretical study of GSK-3 α :
483 neural networks QSAR studies for the design of new inhibitors using 2D
484 descriptors, *Mol. Divers.* 15, 947–955.

485

486 Gawlik, B.M., Sotiriou, N., Feicht, E.A., Schulte-Hostede, S., Kettrup, A., 1997.
487 Alternatives for the determination of the soil adsorption coefficient, K_{OC} , of
488 non-ionicorganic compounds - a review. *Chemosphere* 34, 2525–2551.

489

490 Gharagheizi F., Mehrpooya M., 2008. Prediction of some important physical
491 properties of sulfur compounds using quantitative structure–properties
492 relationships, Mol. Divers. 12, 143–155.

493

494 Golbraikh A., Tropsha A., 2002. Beware of q^2 ! J. Mol. Graph. Model. 20, 269–276.

495

496 Hansen, B.G., Paya-Perez, A.B., Rahman, M., Larsen, B.R., QSARs for K_{OW} and K_{OC}
497 of PCB congeners: a critical examination of data, assumptions and statistical
498 approaches. Chemosphere 39, 2209–2228.

499

500 Han, S., Liang, C., Qiao, J., Lian, H., Ge, X., Chen, H., 2012. A novel evaluation
501 method for extrapolated retention factor in determination of n-octanol/water
502 partition coefficient of halogenated organic pollutants by reversed-phase high
503 performance liquid chromatography. Anal, Chim, Acta 713, 130–135.

504

505 Han, X.Y., Wang, Z.Y., Zhai, Z.C., Wang, L.,S., 2006. Estimation of n-octanol/water
506 partition coefficients (K_{ow}) of all PCB congeners by ab initio and a Cl
507 substitution position method. QSAR Comb. Sci. 25, 333 – 341.

508

509 Harner, T., Bidleman, T.F., 1996. Measurement of octanol–air partition coefficients
510 for polychlorinated biphenyls. J. Chem. Eng. Data 41, 895–899.

511

512 Hawker, D.W., Connell, D.W., 1988. Octanol-Water Partition Coefficients of
 513 Polychlorinated Biphenyl Congeners. *Environ. Sci. Technol.* 22, 382–387.
 514
 515 Hemmer, M.C., Steinhauer, V., Gasteiger, J., 1999. Deriving the 3D structure of
 516 organic molecules from their infrared spectra. *Vib. Spectrosc.* 19, 151–164.
 517
 518 Hippelein, M., McLachlan, M.S., 1998. Soil air partitioning of semivolatile organic
 519 chemicals. 1. Method development and influence of physical chemical properties.
 520 *Environ. Sci. Technol.* 32, 310–316.
 521
 522 Kamlet, M.J., Doherty, R.M., Carr, P.W., Mackay, D., Abraham, M.H., Taft, R.W.,
 523 1988. Linear solvation energy relationships. 44. parameter estimation rules that
 524 allow accurate prediction of octano/water partition coefficients and other
 525 solubility and toxicity properties of polychlorinated biphenyls and polycyclic
 526 aromatic hydrocarbons. *Environ. Sci. Technol.* 22, 503–509.
 527
 528 Kelly, B.C., Ikonomou, M.G., Blair, J.D., Morin, A.E., Gobas, F.A.P.C., 2007. Food
 529 web-specific biomagnification of persistent organic pollutants. *Science.* 317,
 530 236–239.
 531
 532 Larsen, B., Skejo-Andreasen, H., Paya-Perez, A., 1992. Octanol-water partition
 533 coefficients of 39 polychlorinated biphenyls in Askarel mixture. *Fresenius*

534 Environ. Bull. 1(Suppl.), S13–S18.

535

536 Leal F.D., da Silva Lima C.H., de Alencastro R.B., Castro H.C., Rodrigues C.R.,
 537 Albuquerque M.G., 2015 Hologram QSAR models of a series of
 538 6-arylquinazolin-4-amine inhibitors of a new Alzheimer's disease target: dual
 539 specificity tyrosine-phosphorylation-regulated kinase-1A enzyme. Int. J. Mol.
 540 Sci. 16, 5235–5253.

541

542 Li, A., Doucette, W.J., 1993. The effect of cosolutes on the aqueous solubilities and
 543 octanol/water partition coefficients of selected polychlorinated biphenyl
 544 congeners. Environ. Toxicol. Chem. 12, 2031–2035.

545

546 Lü, W., Chen, Y., Liu, M., Chen, X., Hu, Z., 2007. QSPR prediction of
 547 n-octanol/water partition coefficient for polychlorinated biphenyls. Chemosphere
 548 69, 469–478.

549

550 Madhavan, T., Gadhe, C.G., Kothandan, G., Cho, S.J., 2013. Enhancement of
 551 P-glycoprotein modulators of arylmethylanilinephenyl derivatives: an integrative
 552 modeling approach. Med. Chem. Res. 22, 2511–2523.

553

554 Makino, M., 1999. Novel classification to predict relative gas chromatographic
 555 retention times and n-octanol/water partition coefficients of polychlorinated

556 biphenyls. Chemosphere 39, 893–903.

557

558 Moda, T.L., Montanari, C.A., Andricopulo, A.D., 2007. Hologram QSAR model for
559 the prediction of human oral bioavailability. Bioorgan. Med. Chem. 15,
560 7738–7745.

561

562 Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H.,
563 Tropsha, A., 2012. Does rational selection of training and test sets improve the
564 outcome of QSAR modeling? J. Chem. Inf. Model 52, 2570–2578.

565

566 Miller, M.M., Ghodbane, S., Wasik, S.P., Tewari, Y.B., Martire, D.E., 1984. Aqueous
567 solubilities, octanol/water partition coefficients and entropies of melting of
568 chlorinated benzenes and biphenyls. J. Chem. Eng. Data 29, 184–190.

569

570 Padmanabhan, J., Parthasarathi, R., Subramaniana, V., Chattaraj, P.K., 2006. QSPR
571 models for polychlorinated biphenyls: n-octanol/water partition coefficient.
572 Bioorgan. Med. Chem. 14, 1021–1028.

573

574 Ross, G., 2004. The public health implications of polychlorinated biphenyls (PCBs) in
575 the environment. Ecotoxicol. Environ. Saf. 59, 275–291.

576

577 Sabljic, A., Güsten, H., Hermens, J., Opperhulzen, A., 1993. Modeling octanol/water

partition coefficients by molecular topology: chlorinated benzenes and biphenyls.

Environ. Sci. Technol. 27, 1394–1402.

Salum L.B., Andricopulo A.D., 2009. Fragment-based QSAR: perspectives in drug design. Mol. Divers.13, 277–285.

Teófilo, R.F., Martins, J.P.A., Ferreira, M.M.C., 2009. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. J. Chemometrics 23, 32–48.

Wang, X., Tang, S., Liu, S., Cui, S., Wang, L., 2003. Molecular hologram derived quantitative structure–property relationships to predict physico-chemical properties of polychlorinated biphenyls. Chemosphere 51, 617–632.

Wania, F., 2003. Assessing the potential of persistent organic chemicals for long-range transport and accumulation in polar regions. Environ. Sci. Technol. 37, 1344–1351.

Woodburn, K.B., Doucette, W.J., Andren, A.W., 1984. Generator column determination of octanol/water partition coefficients for selected polychlorinated biphenyl congeners. Environ. Sci. Technol. 18, 457–459.

600 Yeh, M.F., Hong, C.S., 2002. Octanol-water partition coefficients of non-ortho- and
601 mono-ortho-substituted polychlorinated biphenyls. J. Chem. Eng. Data 47,
602 209–215.

603

604 Yu, S., Yuan, J., Shi, J., Ruan, X., Zhang, T., Wang, Y., Du, Y., 2015. HQSAR and
605 topomer CoMFA for predicting melanocortin-4 receptor binding affinities of
606 trans-4-(4-chlorophenyl)pyrrolidine-3-carboxamides. Chemom. Intell. Lab. Syst.
607 146, 34–41.

608

609 Zhang, F.L., Yang, X.J., Xue, X.L., Tao, X.Q., Lu, G.N., Dang, Z., 2013. Estimation of
610 n-octanol/water partition coefficients ($\log K_{OW}$) of polychlorinated biphenyls by
611 using quantum chemical descriptors and partial least squares. J. Chem. 2013,
612 1–8.

613

614 Zhang, X., Schramm, K.W., Henkelman, B., Klimm, C., Kaune, A., Kettrup, A., Lu, P.,
615 1999. A method to estimate the octanol–air partition coefficient of semivolatile
616 organic compounds. Anal. Chem. 71, 3834–3838.

617

618 Zhou, W., Zhai, Z., Wang, Z., Wang, L., 2005. Estimation of n-octanol/water partition
619 coefficients (K_{OW}) of all PCB congeners by density functional theory. J. Mol.
620 Struc-THEOCHEM 755, 137–145.

Table 1. IUPAC numbers, Names, experimental, predicted (Pred.) and residual (Res.) values of log K_{OW} and log K_{OA} for 209 PCBs for the optimal models.

IUPAC	Nomenclature	Average	Eq. (2)		HQSAR 17		log K_{OA}	Eq. (4)		HQSAR 37	
		log K_{OW}	Pred.	Res.	Pred.	Res.		Pred.	Res.	Pred.	Res.
PCB-1	2-chlorobiphenyl	4.42	4.67	0.25	4.62	0.2		5.7		6.60	
PCB-2	3-chlorobiphenyl	4.58	4.72	0.14	4.77	0.19		6.77		7.08	
PCB-3	4-chlorobiphenyl	4.45	4.7	0.25	4.79	0.34	7.01	6.61	-0.4	6.97	-0.04
PCB-4	2,2'-dichlorobiphenyl	4.90	4.96	0.06	4.93	0.03		6.01		7.00	
PCB-5	2,3-dichlorobiphenyl		5.09		5.06			6.97		7.62	
PCB-6	2,3'-dichlorobiphenyl		5.14		5.07			7.47		7.60	
PCB-7	2,4-dichlorobiphenyl		5.1		5.11			7.3		7.34	
PCB-8	2,4'-dichlorobiphenyl	5.18	5.06	0.12	5.09	-0.09		7.42		7.49	
PCB-9	2,5-dichlorobiphenyl	5.16	5.09	-0.07	5.05	-0.11		7.91		7.36	
PCB-10	2,6-dichlorobiphenyl	4.93	5.17	-0.24	4.95	0.02		7.29		7.26	

PCB-11	3,3'-dichlorobiphenyl		5.15		5.22			8.55		8.08	
PCB-12	3,4-dichlorobiphenyl	5.29	5.09	-0.2	5.2	-0.09		7.08		8.00	
PCB-13	3,4'-dichlorobiphenyl		5.13		5.24			8.37		7.97	
PCB-14	3,5-dichlorobiphenyl		4.69		5.25			7.17		7.93	
PCB-15	4,4'-dichlorobiphenyl	5.33	5.11	-0.22	5.25	-0.08	7.88	8.36	0.48	7.86	-0.02
PCB-16	2,2',3-trichlorobiphenyl		5.42		5.36			7.62		8.02	
PCB-17	2,2',4-trichlorobiphenyl	5.50	5.36	0.14	5.42	-0.08		7.88		7.74	
PCB-18	2,2',5-trichlorobiphenyl	5.56	5.4	-0.16	5.35	-0.21		8.26		7.76	
PCB-19	2,2',6-trichlorobiphenyl		5.32		5.26			7.51		7.54	
PCB-20	2,3,3'-trichlorobiphenyl	5.68	5.55	-0.13	5.5	-0.18		8.8		8.62	
PCB-21	2,3,4-trichlorobiphenyl		5.46		5.48			7.73		8.26	
PCB-22	2,3,4'-trichlorobiphenyl	5.63	5.47	0.16	5.52	-0.11		8.66		8.51	
PCB-23	2,3,5-trichlorobiphenyl		5.53		5.51			8.56		8.23	

PCB-24	2,3,6-trichlorobiphenyl				5.56		5.36			8.31		8.05
PCB-25	2,3',4-trichlorobiphenyl				5.55		5.55			8.53		8.34
PCB-26	2,3',5-trichlorobiphenyl				5.56		5.5			8.73		8.36
PCB-27	2,3',6-trichlorobiphenyl				5.65		5.39			7.7		8.26
PCB-28	2,4,4'-trichlorobiphenyl	5.68			5.48	-0.2	5.57	-0.11		8.49		8.23
PCB-29	2,4,5-trichlorobiphenyl	5.71			5.48	0.23	5.5	-0.21	8.03	8.1	0.07	8.13 0.10
PCB-30	2,4,6-trichlorobiphenyl	5.60			5.62	0.02	5.47	-0.13		7.42		7.86
PCB-31	2,4',5-trichlorobiphenyl	5.79			5.48	-0.31	5.51	-0.28		8.5		8.25
PCB-32	2,4',6-trichlorobiphenyl	5.47			5.53	-0.06	5.41	-0.06		7.62		8.15
PCB-33	2,3',4'-trichlorobiphenyl				5.47		5.49			7.97		8.52
PCB-34	2,3',5'-trichlorobiphenyl				5.62		5.53			7.88		8.45
PCB-35	3,3',4-trichlorobiphenyl				5.52		5.65			9.02		9.00
PCB-36	3,3',5-trichlorobiphenyl				5.61		5.69			8.97		8.93

PCB-37	3,4,4'-trichlorobiphenyl		5.5		5.66			8.98		8.89	
PCB-38	3,4,5-trichlorobiphenyl		5.49		5.61			7.77		8.75	
PCB-39	3,4',5-trichlorobiphenyl		5.58		5.71			8.83		8.82	
PCB-40	2,2',3,3'-tetrachlorobiphenyl	5.55	5.88	-0.33	5.79	0.24		9.03		9.04	
PCB-41	2,2',3,4-tetrachlorobiphenyl		5.78		5.78			8.47		8.66	
PCB-42	2,2',3,4'-tetrachlorobiphenyl		5.82		5.85			8.78		8.76	
PCB-43	2,2',3,5-tetrachlorobiphenyl		5.88		5.8			8.73		8.63	
PCB-44	2,2',3,5'-tetrachlorobiphenyl		5.86		5.78			8.92		8.78	
PCB-45	2,2',3,6-tetrachlorobiphenyl		5.74		5.67			8.84		8.32	
PCB-46	2,2',3,6'-tetrachlorobiphenyl		5.77		5.69			7.93		8.56	
PCB-47	2,2',4,4'-tetrachlorobiphenyl		5.76		5.9			8.59		8.48	
PCB-48	2,2',4,5-tetrachlorobiphenyl		5.77		5.8			8.83		8.53	
PCB-49	2,2',4,5'-tetrachlorobiphenyl	5.73	5.8	0.07	5.84	0.11	8.57	8.6	0.03	8.50	-0.07

PCB-50	2,2',4,6-tetrachlorobiphenyl		5.75		5.77			8.09		8.13	
PCB-51	2,2',4,6'-tetrachlorobiphenyl		5.7		5.75			8.16		8.28	
PCB-52	2,2',5,5'-tetrachlorobiphenyl	5.81	5.83	0.02	5.77	-0.04		8.85		8.52	
PCB-53	2,2',5,6'-tetrachlorobiphenyl	5.46	5.79	0.33	5.68	0.22	8.24	8.54	0.3	8.30	0.06
PCB-54	2,2',6,6'-tetrachlorobiphenyl	5.48	5.53	0.05	5.6	0.12		7.74		7.96	
PCB-55	2,3,3',4-tetrachlorobiphenyl		5.93		5.92			9.26		9.26	
PCB-56	2,3,3',4'-tetrachlorobiphenyl		5.9		5.93			9.45		9.54	
PCB-57	2,3,3',5-tetrachlorobiphenyl		6		5.95			9.11		9.23	
PCB-58	2,3,3',5'-tetrachlorobiphenyl		6.03		5.97			9.26		9.47	
PCB-59	2,3,3',6-tetrachlorobiphenyl		6.03		5.8			9.26		9.04	
PCB-60	2,3,4,4'-tetrachlorobiphenyl		5.84		5.94			9.16		9.15	
PCB-61	2,3,4,5-tetrachlorobiphenyl	6.08	5.86	-0.22	5.85	-0.23	8.9	8.85	-0.05	8.77	-0.13
PCB-62	2,3,4,6-tetrachlorobiphenyl		5.95		5.8			8.25		8.53	

PCB-63	2,3,4',5-tetrachlorobiphenyl		5.91		5.97			8.91	9.12		
PCB-64	2,3,4',6-tetrachlorobiphenyl		5.91		5.82			8.97	8.93		
PCB-65	2,3,5,6-tetrachlorobiphenyl	5.94	5.97	0.03	5.78	-0.16		9.52	8.68		
PCB-66	2,3',4,4'-tetrachlorobiphenyl	6.20	5.91	0.29	5.98	-0.22	9.22	9.12	-0.1	9.26	0.04
PCB-67	2,3',4,5-tetrachlorobiphenyl		5.95		5.94			9.33	9.13		
PCB-68	2,3',4,5'-tetrachlorobiphenyl		6.04		6.02			8.98	9.19		
PCB-69	2,3',4,6-tetrachlorobiphenyl		6.09		5.9			8.65	8.85		
PCB-70	2,3',4',5-tetrachlorobiphenyl	5.99	5.91	-0.08	5.92	-0.07		9.05	9.28		
PCB-71	2,3',4',6-tetrachlorobiphenyl		5.98		5.82			8.21	9.18		
PCB-72	2,3',5,5'-tetrachlorobiphenyl		6.03		5.96			9.16	9.21		
PCB-73	2,3',5',6-tetrachlorobiphenyl		6.15		5.85			8.05	9.11		
PCB-74	2,4,4',5-tetrachlorobiphenyl		5.87		5.96			9.22	9.02		
PCB-75	2,4,4',6-tetrachlorobiphenyl		5.98		5.93			8.6	8.74		

PCB-76	2,3',4',5'-tetrachlorobiphenyl		5.92		5.89			8.35		9.27	
PCB-77	3,3',4,4'-tetrachlorobiphenyl	6.17	5.89	-0.28	6.08	-0.09	9.96	9.98	0.02	9.92	-0.04
PCB-78	3,3',4,5-tetrachlorobiphenyl		5.93		6.05			9.34		9.75	
PCB-79	3,3',4,5'-tetrachlorobiphenyl		5.98		6.12			9.59		9.85	
PCB-80	3,3',5,5'-tetrachlorobiphenyl		6.07		6.17			9.48		9.78	
PCB-81	3,4,4',5-tetrachlorobiphenyl	6.53	5.89	0.64	6.07	-0.46		9.31		9.64	
PCB-82	2,2',3,3',4-pentachlorobiphenyl		6.23		6.21			9.65		9.68	
PCB-83	2,2',3,3',5-pentachlorobiphenyl		6.34		6.23			9.35		9.65	
PCB-84	2,2',3,3',6-pentachlorobiphenyl	6.24	6.18	0.06	6.09	-0.15		9.58		9.34	
PCB-85	2,2',3,4,4'-pentachlorobiphenyl		6.17		6.27			9.28		9.40	
PCB-86	2,2',3,4,5-pentachlorobiphenyl		6.19		6.14			9.29		9.16	
PCB-87	2,2',3,4,5'-pentachlorobiphenyl		6.21		6.2			9.23		9.42	
PCB-88	2,2',3,4,6-pentachlorobiphenyl		6.11		6.1			9.02		8.81	

PCB-89	2,2',3,4,6'-pentachlorobiphenyl		6.1		6.11			8.79		9.20	
PCB-90	2,2',3,4',5-pentachlorobiphenyl		6.28		6.29			9.02		9.37	
PCB-91	2,2',3,4',6-pentachlorobiphenyl		6.11		6.15			9.14		9.06	
PCB-92	2,2',3,5,5'-pentachlorobiphenyl	6.11	6.32	0.21	6.22	0.11		9.2		9.39	
PCB-93	2,2',3,5,6-pentachlorobiphenyl		6.15		6.08			9.75		8.96	
PCB-94	2,2',3,5,6'-pentachlorobiphenyl		6.26		6.12			8.98		9.17	
PCB-95	2,2',3,5',6-pentachlorobiphenyl	6.08	6.18	0.1	6.08	0	9.06	9.35	0.29	9.08	0.02
PCB-96	2,2',3,6,6'-pentachlorobiphenyl		5.94		6		8.77	9.1	0.33	8.74	-0.03
PCB-97	2,2',3,4',5'-pentachlorobiphenyl		6.22		6.23			9.75		9.55	
PCB-98	2,2',3,4',6'-pentachlorobiphenyl		6.19		6.2			8.95		9.15	
PCB-99	2,2',4,4',5-pentachlorobiphenyl		6.17		6.28			9.37		9.27	
PCB-100	2,2',4,4',6-pentachlorobiphenyl		6.13		6.26			8.76		8.87	
PCB-101	2,2',4,5,5'-pentachlorobiphenyl	6.31	6.2	0.11	6.22	-0.09	9.31	9.41	0.1	9.29	-0.02

PCB-102	2,2',4,5,6'-pentachlorobiphenyl		6.13		6.12			9.19		9.07	
PCB-103	2,2',4,5',6-pentachlorobiphenyl		6.21		6.19			8.79		8.89	
PCB-104	2,2',4,6,6'-pentachlorobiphenyl	5.37	5.93	0.56	6.11	0.74		8.37		8.55	
PCB-105	2,3,3',4,4'-pentachlorobiphenyl	6.22	6.27	0.05	6.35	0.13	10.27	10.13	-0.14	10.18	-0.09
PCB-106	2,3,3',4,5-pentachlorobiphenyl		6.33		6.29			9.7		9.77	
PCB-107	2,3,3',4',5-pentachlorobiphenyl		6.34		6.38			9.69		10.15	
PCB-108	2,3,3',4,5'-pentachlorobiphenyl		6.4		6.39			9.8		10.11	
PCB-109	2,3,3',4,6-pentachlorobiphenyl		6.43		6.24			9.48		9.53	
PCB-110	2,3,3',4',6-pentachlorobiphenyl	6.22	6.36	0.14	6.23	0.01		9.67		9.96	
PCB-111	2,3,3',5,5'-pentachlorobiphenyl		6.46		6.41			9.58		10.08	
PCB-112	2,3,3',5,6-pentachlorobiphenyl		6.44		6.22			9.91		9.68	
PCB-113	2,3,3',5',6-pentachlorobiphenyl		6.53		6.26			9.7		9.89	
PCB-114	2,3,4,4',5-pentachlorobiphenyl	6.47	6.24	0.23	6.31	-0.16		9.52		9.66	

PCB-115	2,3,4,4',6-pentachlorobiphenyl		6.31		6.26		9.32	9.42		
PCB-116	2,3,4,5,6-pentachlorobiphenyl	6.30	6.31	0.01	6.11	-0.19	9.48	8.93		
PCB-117	2,3,4',5,6-pentachlorobiphenyl		6.32		6.24		9.48	9.57		
PCB-118	2,3',4,4',5-pentachlorobiphenyl	6.37	6.3	-0.07	6.37	0	10.08	10.09	0.01	10.05 -0.03
PCB-119	2,3',4,4',6-pentachlorobiphenyl		6.42		6.33		9.23	9.77		
PCB-120	2,3',4,5,5'-pentachlorobiphenyl		6.42		6.41		9.93	9.98		
PCB-121	2,3',4,5',6-pentachlorobiphenyl		6.59		6.36		9.09	9.70		
PCB-122	2,3,3',4',5'-pentachlorobiphenyl		6.33		6.33		9.8	10.29		
PCB-123	2,3',4,4',5'-pentachlorobiphenyl	6.50	6.34	-0.16	6.38	-0.12	9.45	10.01		
PCB-124	2,3',4',5,5'-pentachlorobiphenyl		6.34		6.32		9.43	10.03		
PCB-125	2,3',4',5',6-pentachlorobiphenyl		6.44		6.21		8.58	9.93		
PCB-126	3,3',4,4',5-pentachlorobiphenyl	6.56	6.29	-0.27	6.48	-0.08	10.61	10.33	-0.28	10.67 0.06
PCB-127	3,3',4,5,5'-pentachlorobiphenyl		6.39		6.53		9.96	10.60		

PCB-128	2,2',3,3',4,4'-hexachlorobiphenyl	6.86	6.58	-0.28	6.63	-0.23		10.24		10.32
PCB-129	2,2',3,3',4,5-hexachlorobiphenyl		6.65		6.57			10.14		10.19
PCB-130	2,2',3,3',4,5'-hexachlorobiphenyl		6.69		6.65			9.88		10.29
PCB-131	2,2',3,3',4,6-hexachlorobiphenyl		6.56		6.53			9.98		9.83
PCB-132	2,2',3,3',4,6'-hexachlorobiphenyl		6.52		6.51			9.88		9.98
PCB-133	2,2',3,3',5,5'-hexachlorobiphenyl		6.8		6.67			9.64		10.26
PCB-134	2,2',3,3',5,6-hexachlorobiphenyl	6.81	6.6	-0.21	6.5	-0.31		10.15		9.98
PCB-135	2,2',3,3',5,6'-hexachlorobiphenyl		6.67		6.53			9.82		9.96
PCB-136	2,2',3,3',6,6'-hexachlorobiphenyl	6.50	6.35	-0.15	6.4	-0.1		9.8		9.53
PCB-137	2,2',3,4,4',5-hexachlorobiphenyl		6.59		6.63			9.68		9.91
PCB-138	2,2',3,4,4',5'-hexachlorobiphenyl	6.42	6.57	0.15	6.65	0.23	10.09	10.18	0.09	10.19 0.10
PCB-139	2,2',3,4,4',6-hexachlorobiphenyl		6.49		6.59			9.57		9.55
PCB-140	2,2',3,4,4',6'-hexachlorobiphenyl		6.53		6.62			9.45		9.79

PCB-141	2,2',3,4,5,5'-hexachlorobiphenyl		6.63		6.57			9.75		9.92
PCB-142	2,2',3,4,5,6-hexachlorobiphenyl		6.71		6.41			9.49		9.21
PCB-143	2,2',3,4,5,6'-hexachlorobiphenyl		6.79		6.47			9.38		9.71
PCB-144	2,2',3,4,5',6-hexachlorobiphenyl		6.58		6.52			9.57		9.57
PCB-145	2,2',3,4,6,6'-hexachlorobiphenyl		6.29		6.44			9.36		9.23
PCB-146	2,2',3,4',5,5'-hexachlorobiphenyl	6.57	6.68	0.11	6.67	0.1		10		10.16
PCB-147	2,2',3,4',5,6-hexachlorobiphenyl		6.53		6.56			9.65		9.70
PCB-148	2,2',3,4',5,6'-hexachlorobiphenyl		6.68		6.64			9.19		9.77
PCB-149	2,2',3,4',5',6-hexachlorobiphenyl	6.54	6.52	-0.02	6.53	-0.01		9.94		9.85
PCB-150	2,2',3,4',6,6'-hexachlorobiphenyl		7.37		6.51			9.12		9.34
PCB-151	2,2',3,5,5',6-hexachlorobiphenyl	6.45	6.81	0.36	6.49	0.04		9.79		9.72
PCB-152	2,2',3,5,6,6'-hexachlorobiphenyl		7.37		6.41			9.81		9.38
PCB-153	2,2',4,4',5,5'-hexachlorobiphenyl	6.77	6.56	-0.21	6.66	-0.11	10.04	10.17	0.13	10.06 0.02

PCB-154	2,2',4,4',5,6'-hexachlorobiphenyl		6.81		6.64			9.34		9.66	
PCB-155	2,2',4,4',6,6'-hexachlorobiphenyl	7.55	7.39	-0.16	6.63	-0.92	9.16	8.75	-0.41	9.15	-0.01
PCB-156	2,3,3',4,4',5-hexachlorobiphenyl	6.70	6.66	-0.04	6.72	0.02		10.5		10.69	
PCB-157	2,3,3',4,4',5'-hexachlorobiphenyl	6.73	6.7	0.03	6.75	0.02		10.48		10.93	
PCB-158	2,3,3',4,4',6-hexachlorobiphenyl		6.75		6.66			10.26		10.45	
PCB-159	2,3,3',4,5,5'-hexachlorobiphenyl		6.79		6.76			10.27		10.62	
PCB-160	2,3,3',4,5,6-hexachlorobiphenyl		6.78		6.54			9.81		9.93	
PCB-161	2,3,3',4,5',6-hexachlorobiphenyl		6.93		6.7			10.04		10.38	
PCB-162	2,3,3',4',5,5'-hexachlorobiphenyl		6.76		6.77			10.07		10.90	
PCB-163	2,3,3',4',5,6-hexachlorobiphenyl	6.39	6.77	0.38	6.64	0.25		9.88		10.60	
PCB-164	2,3,3',4',5',6-hexachlorobiphenyl		6.82		6.62			10.02		10.71	
PCB-165	2,3,3',5,5',6-hexachlorobiphenyl		6.94		6.68			10.23		10.53	
PCB-166	2,3,4,4',5,6-hexachlorobiphenyl		6.66		6.57			9.5		9.82	

PCB-167	2,3',4,4',5,5'-hexachlorobiphenyl	6.82	6.71	0.11	6.77	-0.05		10.49		10.80
PCB-168	2,3',4,4',5,6-hexachlorobiphenyl		6.88		6.72			9.57		10.52
PCB-169	3,3',4,4',5,5'-hexachlorobiphenyl	7.01	6.7	-0.31	6.89	-0.12		10.75		11.41
PCB-170	2,2',3,3',4,4',5-heptachlorobiphenyl	6.83	7	-0.17	6.99	0.16		10.6		10.82
PCB-171	2,2',3,3',4,4',6-heptachlorobiphenyl	6.68	6.89	0.21	6.95	0.27	10.51	10.44	-0.07	10.47 -0.04
PCB-172	2,2',3,3',4,5,5'-heptachlorobiphenyl	6.85	7.11	0.26	7.01	0.16		10.36		10.80
PCB-173	2,2',3,3',4,5,6-heptachlorobiphenyl		6.91		6.83			10.21		10.23
PCB-174	2,2',3,3',4,5,6'-heptachlorobiphenyl		6.96		6.87			10.3		10.49
PCB-175	2,2',3,3',4,5',6-heptachlorobiphenyl		7.04		6.96			10.2		10.44
PCB-176	2,2',3,3',4,6,6'-heptachlorobiphenyl	6.76	6.7	-0.06	6.84	0.08		10.12		10.02
PCB-177	2,2',3,3',4,5',6'-heptachlorobiphenyl		6.93		6.92			10.05		10.62
PCB-178	2,2',3,3',5,5',6-heptachlorobiphenyl		7.09		6.94			10.35		10.59
PCB-179	2,2',3,3',5,6,6'-heptachlorobiphenyl		6.78		6.8			10.25		10.16

PCB-180	2,2',3,4,4',5,5'-heptachlorobiphenyl	6.56	6.98	-0.42	7.01	0.45	10.75	10.53	-0.22	10.70	-0.05
PCB-181	2,2',3,4,4',5,6-heptachlorobiphenyl		6.84		6.89			9.75		9.95	
PCB-182	2,2',3,4,4',5,6'-heptachlorobiphenyl		6.96		6.98			9.88		10.30	
PCB-183	2,2',3,4,4',5',6-heptachlorobiphenyl	6.78	6.91	-0.13	6.97	0.19		10.35		10.34	
PCB-184	2,2',3,4,4',6,6'-heptachlorobiphenyl	6.76	6.69	-0.07	6.95	0.19		9.74		9.83	
PCB-185	2,2',3,4,5,5',6-heptachlorobiphenyl		6.93		6.82			9.93		9.97	
PCB-186	2,2',3,4,5,6,6'-heptachlorobiphenyl		6.66		6.73			9.95		9.63	
PCB-187	2,2',3,4',5,5',6-heptachlorobiphenyl	6.76	6.95	0.19	6.94	0.18		10.14		10.49	
PCB-188	2,2',3,4',5,6,6'-heptachlorobiphenyl		6.77		6.92			9.81		9.97	
PCB-189	2,3,3',4,4',5,5'-heptachlorobiphenyl	6.95	7.08	0.13	7.12	0.17		10.9		11.43	
PCB-190	2,3,3',4,4',5,6-heptachlorobiphenyl		7.1		6.97			10.51		10.85	
PCB-191	2,3,3',4,4',5',6-heptachlorobiphenyl		7.2		7.06			10.62		11.20	
PCB-192	2,3,3',4,5,5',6-heptachlorobiphenyl		7.28		7.01			10.28		10.78	

PCB-193	2,3,3',4',5,5',6-heptachlorobiphenyl		7.22		7.04		10.21	11.34
PCB-194	2,2',3,3',4,4',5,5'-octachlorobiphenyl	7.31	7.41	0.1	7.36	0.05	10.96	11.33
PCB-195	2,2',3,3',4,4',5,6-octachlorobiphenyl	6.95	7.23	-0.28	7.25	0.3	10.69	10.87
PCB-196	2,2',3,3',4,4',5,6'-octachlorobiphenyl		7.32		7.31		10.76	10.98
PCB-197	2,2',3,3',4,4',6,6'-octachlorobiphenyl		7.05		7.27		10.57	10.51
PCB-198	2,2',3,3',4,5,5',6-octachlorobiphenyl		7.4		7.27		10.46	10.84
PCB-199	2,2',3,3',4,5,5',6'-octachlorobiphenyl		7.37		7.28		10.5	11.12
PCB-200	2,2',3,3',4,5,6,6'-octachlorobiphenyl		7.07		7.13		10.36	10.42
PCB-201	2,2',3,3',4,5',6,6'-octachlorobiphenyl	7.21	7.13	-0.08	7.24	0.03	10.32	10.65
PCB-202	2,2',3,3',5,5',6,6'-octachlorobiphenyl	7.15	7.22	0.07	7.21	0.06	10.47	10.80
PCB-203	2,2',3,4,4',5,5',6-octachlorobiphenyl	6.93	7.26	0.33	7.27	0.34	10.63	10.74
PCB-204	2,2',3,4,4',5,6,6'-octachlorobiphenyl		7.06		7.25		9.95	10.23
PCB-205	2,3,3',4,4',5,5',6-octachlorobiphenyl		7.54		7.36		10.93	11.60

PCB-206	2,2',3,3',4,4',5,5',6-nonachlorobiphenyl	7.07	7.67	0.6	7.61	0.54	11.05	11.38
PCB-207	2,2',3,3',4,4',5,6,6'-nonachlorobiphenyl	7.52	7.41	-0.11	7.57	0.05	10.83	10.90
PCB-208	2,2',3,3',4,5,5',6,6'-nonachlorobiphenyl	8.16	7.5	-0.66	7.54	-0.62	10.59	11.05
PCB-209	2,2',3,3',4,4',5,5',6,6'-decachlorobiphenyl	8.22	7.78	-0.44	7.87	-0.35	11.14	11.30

Table 2. HQSAR statistical results for various fragment distinction parameters using the fragment size (4-7) for the log K_{OW} prediction.

HQSAR	Fragment	HL	LVs	q^2	SEP	r^2	SEE	r_m^2
No.	distinction							
1	A	151	2	0.857	0.320	0.882	0.291	0.792
2	B	257	3	0.859	0.321	0.903	0.267	0.845
3	C	257	2	0.852	0.326	0.882	0.292	0.788
4	H	307	2	0.875	0.300	0.892	0.279	0.817
5	Ch	307	2	0.875	0.300	0.892	0.279	0.817
6	DA	307	2	0.875	0.300	0.892	0.279	0.817
7	A/B	307	4	0.854	0.330	0.907	0.264	0.846
8	A/C	257	2	0.853	0.325	0.884	0.290	0.793
9	B/C	353	4	0.854	0.330	0.907	0.265	0.846
10	A/B/C	353	4	0.854	0.330	0.905	0.267	0.840

11	A/B/C/H	199	2	0.865	0.312	0.899	0.270	0.827
12	A/B/C/H/Ch	199	2	0.865	0.312	0.899	0.270	0.827
13	A/B/C/H/Ch/DA	199	2	0.865	0.312	0.899	0.270	0.827

LVs = Number of latent variables, HL = best hologram length. $r_m^2 = r_m^2$ for the overall dataset

Table 3. HQSAR analysis for the log K_{OA} prediction using various fragment distinction combinations with the fragment size (4-7).

HQSAR No.	Fragment distinction	HL	LVs	q^2	SEP	r^2	SEE	r_m^2
22	A	97	6	0.983	0.188	0.998	0.071	0.973
23	B	199	5	0.985	0.167	0.996	0.089	0.943
24	C	257	6	0.976	0.226	0.998	0.072	0.958
25	H	307	3	0.879	0.433	0.944	0.293	0.854
26	Ch	307	3	0.879	0.433	0.944	0.293	0.854
27	DA	307	3	0.879	0.433	0.944	0.293	0.854
28	A/B	97	5	0.990	0.136	0.997	0.076	0.949
29	A/C	151	6	0.983	0.190	0.997	0.075	0.966
30	B/C	257	5	0.989	0.142	0.997	0.079	0.950
31	A/B/C	199	5	0.990	0.140	0.997	0.076	0.953
32	A/B/C/H	151	4	0.992	0.116	0.997	0.073	0.952

33	A/B/C/H/Ch	151	4	0.992	0.116	0.997	0.073	0.952
34	A/B/C/H/Ch/DA	151	4	0.992	0.116	0.997	0.073	0.952

Table 4. Comparison of squared correlation coefficients of the training and test sets from different models for the prediction of log K_{OW} and log K_{OA} .

References	N/Nt	Predict log K_{OW} or log K_{OA}	Descriptor type	Statistical technique	r^2	r^2_{pred}
Yeh & Hong, (2002)	46/-	log K_{OW}	the number of substitutions	MLR	0.927	-
Hawker & Connell, (1988)	46/-	log K_{OW}	total surface areas	LR	0.920	-
Chen et al. (2002)	19/104 ^a	log K_{OA}	quantum chemical descriptors	PLS	0.976	-
This work	55/18	log K_{OW}	DRAGON	MLR	0.917	0.857
	55/18	log K_{OW}	HQSAR	PLS	0.907	0.912
	15/4	log K_{OA}	DRAGON	MLR	0.944	0.958
	15/4	log K_{OA}	HQSAR	PLS	0.997	0.989

^a In the test set, the log K_{OA} values were determined semi-empirically by Zhang et al. (1999) using a multicolumn method.

N/Nt, the number of compounds in the training and test set, respectively

Figure captions

Figure 1. Predicted versus experimental $\log K_{OW}$ values of the Eq. (2) (A) and HQSAR 17 (B) models for training and test sets. Predicted versus experimental $\log K_{OA}$ values of the Eq. (4) (C) and HQSAR 37 (D) models for training and test sets.

Figure

Figure 1