Environ. Sci. Technol. **1999**, *33,* 2719−2725

Estimation of Organic Carbon Normalized Sorption Coefficient (K_{OC}) for Soils Using the Fragment Constant Method

SHU TAO,* HAISHAN PIAO, R. DAWSON, XIAOXIA LU, AND HAIYING HU

Department of Urban and Environmental Sciences, Peking University, Beijing 100871, China

A fragment constant model for prediction of K_{OC} was developed and evaluated with a diverse database of 592 chemicals belonging to 17 classes. The range of experimental K_{OC} covered 7.65 log-units. The 592 chemicals were randomly divided into a training set and a testing set for model development and validation. A general model was then established using the entire database having 74 fragment constants and 24 structural factors. Statistically, the regression model accounted for as much as 96.96% of the variation in the measured log K_{OC}. The mean residual between the experimental and predicted K_{OC} values was 0.366 log-units. In more than 74% of the chemicals studied the residual values were less than 0.5 log-units. The robustness of the regression model, with respect to either specific individual chemicals or particular compound classes, was evaluated through use of jackknife tests. The experimental results confirmed the ability of the fragment model to predict K_{OC} for a wide variety of untested chemicals.

Introduction

Movement of a chemical across the water/soil or water/sediment interface is primarily a sorption−desorption process governed by the solubility of the particular chemical in water and its affinity to solid phase. Harmaker demonstrated that the amount of nonionic organic chemicals sorbed varied from soil to soil and that such variation was primarily due to the organic content of the soil (*1*). As such, an organic carbon normalized sorption coefficient (K_{OC}) represents an important parameter for organic chemicals. There have been numerous efforts attempting to estimate K_{OC} through statistical modeling. These approaches can be divided into two general categories: (1) estimation using other parameters and (2) estimation based on molecular structure. Gawlik et al. performed a thorough review of the various alternative methods and concluded that the application of the octanol−water partition coefficient (K_{OW}) as a descriptor for soil sorption was the most commonly used approach (*2*). A fairly large number of regression equations have been derived describing the relationships between the K_{OC} and other related parameters (*3, 4*). Sabljic and Protic introduced the use of topological indices as predictors for K_{OC} (*5*). Their method has been one of the most extensively studied approaches (*3, 6, 7*). In addition to molecular connectivity indices, other methods including fragment constant approach have been applied for estimation of the K_{OW} and may also be used for K_{OC} prediction.

Hammett postulated that free-energy based parameters are additive and it can be assumed that the K_{OW} depends on the structure of a chemical in an additive-constitutive fashion (*8*). Rekker proposed a general approach for predicting the K_{OW} in which each part, or fragment, of a chemical was given a value such that the sum would yield the log K_{OW} (*9*). Using statistical methods Rekker and co-workers were able to determine the fragment constants that were the average contributions of simple fragments. As with the molecular connectivity index method, the only input required for the fragment constant method is chemical structure. Leo established rules for fragmenting chemicals and even developed a computer algorithm with a coded computer program to calculate the K_{OW} of organic compounds using the fragment method (*10, 11*).

In addition to the contributions of the fragments, the sorption property of a chemical is influenced by the structure of the molecule. Structural effects can be adequately described by structural correction factors which take into account molecular flexibility, unsaturation, multiple halogenation, branching, interactions with polar fragments, and so forth (*10*). Test calculations have shown that these assumptions are justified for most chemicals, but in some cases they deviate seriously from the norm. Hansch and Leo calculated the fragment and factor values for the K_{OW} of 76 compounds. The results indicated that the mean residual of the model was 0.14 log-units with 66% of chemicals having errors of less than 0.1 log-units and 83% of chemicals having errors of less than 0.2 log-units (*12*). Chou and Jurs computerized the model and tested it on several large data sets of organic compounds. They found that the log K_{OW} could be estimated for 84.5% of the compounds (*13*). The results seem to indicate that given unlimited parameters and a large enough database, one could always derive useful predictive equations (*14*). The fragment approach, however, has not been tried for prediction of the K_{OC} although its application in the case of K_{OW} is considered successful. One reason for its lack of use in the case of K_{OC} might lie in the fact that the amount of experimental data available for K_{OC} is much less than that for K_{OW}, while the experimental error associated with experimental K_{OC} is higher than that with measured K_{OW}.

During the past decade, quite a number of experimental K_{OC} have been reported. The objective of this study is to develop a method, using fragment constants and structural correction factors, for estimating K_{OC}. The residual of the model was analyzed, and modified jackknife tests were used to test the robustness of the regression model.

Methodology

Collection of Experimental K_{OC}. Measured K_{OC} for 592 chemicals were collected from the literature and then compiled into a database. Many of the chemicals in the database had more than one K_{OC} value as a result of their derivation from different sources. There were 1010 total entries in the database with a range of K_{OC} values exceeding seven log-units. While a large proportion of the measured K_{OC} were collected from several comprehensive studies (*7, 15, 16*), additional data were added to the data set for completeness (*17*−*28*). For data reported in terms of K_{OM}, a conversion factor of 1.724 was adopted. Only a single K_{OC} value for each chemical was required for the regression modeling. When more than one value was available for a single chemical the median was adopted. The uncertainty in

* Corresponding author phone/fax: (86)10-62751938; e-mail: taos@urban.pku.edu.cn.

10.1021/es980833d CCC: $18.00
Published on Web 07/02/1999

TABLE 1. Structural Correction Factors

no.	symbol	factors	explanation
1	F_b	aliphatic chain bond factor	number of aliphatic chain bond − 1
2	F_b^o	aliphatic ring bond factor	number of aliphatic ring bond −1
3	F_{CBr}	chain branch factor	number of branch[a]
4	$F_=$	aliphatic double bond factor	number of aliphatic chain double bond
5	$F_=^o$	aliphatic ring double bond factor	number of aliphatic ring double bond
6	$F_\equiv$	triple bond factor	number of triple bond
7	F_{mhG1}	double halogenation factor	number of carbon bonds to two halogens
8	F_{mhG2}	triple halogenation factor	number of carbon bonds to three halogens
9	F_{mhG3}	four halogenation factor	number of carbon bonds to four halogens
10	F_{mhV}	vicinal multiple halogenation factor	number of vicinal carbons bonds to multiple halogens
11	F_{P0}	directly connected H-polar factor	number of directly bonds to a H-polar fragment
12	F_{P1}	double aliphatic H-polar factor	number of aliphatic carbon bonds to two H-polar fragments[b]
13	F_{P2}	double aliphatic/aromatic H-polar factor	number of aliphatic−aromatic chain carbon bounds to two H-polar groups
14	F_{P3}	vicinal aliphatic chain double H-polar factor	number of vicinal aliphatic chain carbon bonds to two H-polar groups
15	F_{P4}	double aliphatic chain H-polar factor	number of aliphatic ring carbon bonds to two H-polar groups
16	F_{P5}	vicinal aliphatic ring double H-polar factor	number of vicinal aliphatic ring carbon bonds to two H-polar groups
17	F_{P6}	double aromatic H-polar factor	number of aromatic carbon bonds to two H-polar groups
18	F_{HP1}	aliphatic halogen and H-polar factor	number of vicinal aliphatic carbon bonds to a halogen and a H-polar group
19	F_{HP2}	aromatic halogen and H-polar factor	number of vicinal aromatic carbon bonds to a halogen and a H-polar group
20	F_C	quaternary carbon factor	number of quaternary carbon
21	F_{CH}	triple carbon factor	number of triple bonded carbon
22	F_{CH_2}	secondary carbon factor	number of secondary carbon
23	F_{NH}	secondary nitrogen factor	number of secondary nitrogen
24	F_{NH_2}	primary nitrogen factor	number of primary nitrogen

[a] "One-time" chain branching [*33*]. [b] H-polar fragment is one that can be expected to participate in hydrogen bonding [*33*].

the experimental K_{OC}, introduced through use of data from various sources, was evaluated and compared to the predicted results based on the duplicated K_{OC} values for those overlapping chemicals collected.

Selection of Atomic and Group Fragments. The fragments were identified based on Leo's definition which refers to a fragment as an atom, or atoms, whose exterior bonds are to isolating carbon atoms. An isolating carbon is one that either has four single bonds, at least two of which are to nonheteroatoms, or is multiply bonded to other carbon atoms (*10*). Further, a single-atom fragment can only be an isolating carbon atom or a hydrogen or heteroatom (e.g. −H, −O−). A multiple-atom fundamental fragment can be formed by any combination of nonisolating carbon, hydrogen, and/or heteroatoms (e.g. −C(O)O− and −OH). A multiple-atom derived fragment can be any combination of single-atom or multiple-atom fundamental fragments. It is essential to guarantee that the fragments of a chemical must not be an arbitrarily selected set. Additionally, the contribution of a fragment to the sorption behavior of a chemical also depends on the type of isolating carbon atom to which the fragment is attached. Different constant values should be assigned to fragments attached to different types of isolating carbon. Superscripts on the fragment constant which denote the type of attachment are ϕ, attached to aromatic ring (if bivalent the attachment is from the left); $1/\phi$, attached to aromatic ring (if bivalent the attachment is from the right); $\phi\phi$, bivalent fragment with two aromatic attachments; and AR, fused in aromatic ring. For aliphatic structural attachment, there is no superscript (*10*). Eighty-six fragments were identified from the 592 chemicals at the beginning of the study. The number was reduced to 74 after preliminary modeling.

Selection of Structural Correction Factors. The influence of structural features on K_{OC} should be taken into consideration when characterizing chemicals with a relatively complex structure (*10*). For the model developed in this study, 19 structural factors (nos. 1−19 in Table 1) were identified for the preliminary modeling. This number increased to 24 with the addition of five more factors (nos. 20−24 in Table 1) following evaluation of the preliminary modeling results.

Model Development. The 592 chemicals selected were divided into a training set and a validation set. The 430 chemicals in the training set were randomly selected. The remaining 162 chemicals were included in the independent validation set. A preliminary modeling was conducted based on the experimental K_{OC} and the numbers of individual fragments and structural features of each chemical in the training set. The results of an evaluation on the calculated fragment constants and correction factors provided the basis for modifying the scheme of the fragmentation and structural factor selection. The model was then rebuilt using the revised independent variables and validated using data from the testing set. The final regression model was established using data for all 592 chemicals.

The K_{OC} of a chemical with known structure can be calculated using the following equation

$$\log K_{OC} = \sum_{i=1}^{a} n_i f_i + \sum_{j=1}^{b} m_j F_j \qquad (1)$$

where a and b are the total numbers of the fragments and structural correction factors defined by the model; n_i and m_j are the number of the ith fragment and the jth structural factor of the chemical; f_i is the fragment constant for the ith fragment; and F_j is the structural factor for the jth structural feature. The coefficient of determination (R^2) was taken into consideration in testing the quality of the regression. The mean residual was used for the same purpose. SPSS was used for the regression analysis.

Evaluation of Robustness. Modified jackknife tests were applied in the following ways to evaluate the robustness of the model (*29, 30*):

1. deletion of single chemical randomly selected from the database for 100 times;
2. deletion of single chemical with residual greater than 1.0 log-units (34 chemicals);
3. deletion of a set of 50 chemicals randomly selected from the database for 40 times; and
4. deletion of the 17 classes of chemicals one by one.

A regression was performed after each deletion, and the calculated jackknife R^2s were compared with one another and with R^2 derived prior to deletion for robustness evaluation.

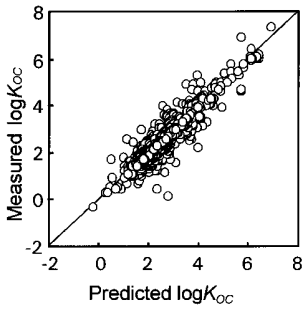

FIGURE 1. Relationship between the measured and the predicted K_{OC} of 430 compounds from the training set (86 fragments and 19 structural factors applied).

Results and Discussion

Preliminary Modeling. A linear multiple regression was conducted on the K_{OC} from the training set. The equation derived using the 86 fragments and the 19 structural correction factors is

$$\log K_{OC} = \sum_{i=1}^{86} n_i f_i + \sum_{j=1}^{19} m_j F_j, \quad n = 430, \quad R^2 = 0.9729 \tag{2}$$

where n_i, m_j, f_i, and F_j are the same as those in eq 1. Statistically, the regression model is able to account for as much as 97.29% of the variation in the experimental log K_{OC} of the 430 chemicals. A scatter plot of the 430 measured K_{OC} against the predicted ones is given in Figure 1. The plot demonstrates that the residuals of the estimation are acceptable, especially when the uncertainty imbedded in the measured data is taken into consideration.

The model based on chemicals in the training set looks good in terms of its predictive certainty. Nevertheless, it was recognized that the signs of several fragment constants derived from the regression were not as expected. For instance, some hydrophobic atomic or group fragments were assigned negative, while positive fragment constants were taken for a few of the hydrophilic fragments. For the fragments which appeared in only a few of the chemicals in the entire data set (e.g. −SP(S)(O−)S−), the unusual values might be the result of the small sample size and the experimental error. The influence of such fragments on the model can be ignored. For several other fragments, such as nonpolar fragments (e.g. −CH< and >C<) and polar fragments (e.g. $-NH_2$ and −NH−), the abnormal values empirically derived from the regression modeling could not be explained by random error, implying there could be additional room for optimizing the means of fragmentation. To accomplish this, >C<, −CH<, $-CH_2-$, and $-CH_3$ were grouped together as one fragment. The different contributions of these groups to overall K_{OC} could be corrected with three new structural factors, F_C, F_{CH}, and F_{CH_2}. Similarly, −N<, −NH−, and $-NH_2$ were assembled together as a single group fragment, and two more structural factors, F_{NH} and F_{NH_2}, were introduced into the model. With the modified fragmentation scheme, a regression was again performed on the 430 chemicals. The performance of the analysis in terms of R^2 (0.9725) was very close to the previous test. The derived values for the fragment constants and the structural factors were more interpretable, with the mean residual falling between the measured and the calculated K_{OC} for 430 chemicals at 0.358 log-units.

Model Validation. Predictions based on the model were proved to be accurate, with the mean residual of 0.358 log-units for the training set. To test the predictive capability of the regression equation and the statistical validity of the modeling, the model was applied to the rest of 162 chemicals in the testing set. The predicted K_{OC} are plotted against the measured ones in Figure 2.

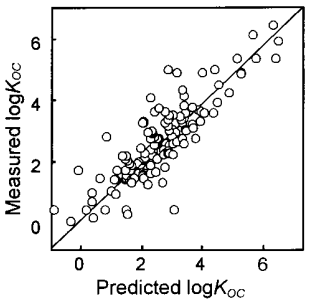

FIGURE 2. Relationship between the measured and the predicted K_{OC} of 162 compounds from the testing set (the model was developed based on the 430 chemicals from the training set).

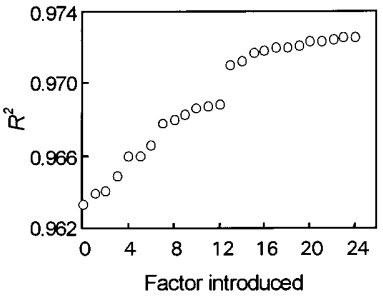

FIGURE 3. Results of the stepwise regression to test the contribution of individual structural factors introduced one by one (The factors were introduced in following order: $F_{=}$, F_{NH_2}, F_C, F_{CH}, F_{P3}, F_{NH}, F_{P0}, F_{P2}, F_{P1}, F_{CH_2}, F_{P4}, F_{HP1}, F^{O}_{b}, F_{P5}, F_{mhG2}, F_{HP2}, F_{mhG3}, F_{mhV}, $F^{O}_{=}$, F_{CBr}, $F_{=}$, F_{P6}, F_{mhG1}, and F_b).

Although the mean residual of the calculated K_{OC} for the chemicals from the testing set were slightly larger than that from the training set (0.468 versus 0.358 log-units), the errors in more than 60% of chemicals from the testing set were less than 0.5 log-units. Because the 162 chemicals were selected randomly from the database, the agreement between the predicted and the observed values of K_{OC} for the members of the testing set is generally satisfactory.

Structural Correction Factors. The structural correction factors were selected in a manner similar to Leo's K_{OW} prediction model (*10*). To justify the use of these factors in the current model, a stepwise multiple regression was conducted using the 430 chemicals from the training set. All 24 structural correction factors were tested for their necessity in the K_{OC} modeling. The factors were introduced into the stepwise model, and the R^2s are plotted against the number of factors introduced in Figure 3. For the first data point having a zero structural factor, the R^2 (0.963) represents the coefficient of determination of the fragment constant model with no structural correction factor introduced.

The results presented in Figure 3 indicate a general increase in R^2 with the rate of increase slowing down gradually. It appears that all of the structural factors selected contributed positively to the regression, although the influence of the first 14 factors on the model was more significant than that of the others. The R^2 for the regression model increased from 0.963 to 0.973 after all 24 factors were introduced.

Establishment of a General Model. Generally, large data sets are preferred for model development. Dividing the entire database into the training and the testing sets is essential for model evaluation. After the validation, the two data sets were merged and analyzed together. The optimal estimation results can be achieved by utilizing all the information available in this study. Fragments and structural factors selected for the

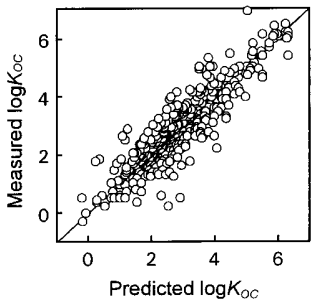

FIGURE 4. Relationship between the measured and the predicted K_{OC} of 592 compounds from both the training and the testing sets.

modeling were the same as those in the preliminary modeling exercise. The equation is exactly the same as eq 1, while the values of the constants and factors derived are slightly different from those obtained in the preliminary modeling. The R^2 is 0.9696 with sample size of 592. The average error of the predicted K_{OC} was 0.366 log-units. For 74% of chemicals used in the modeling, the differences between the measured and the calculated K_{OC} values were less than 0.5 log-units. The predicted K_{OC} values are plotted against the measured ones in Figure 4.

Error Evaluation. Error in the prediction model may be generated from a number of sources. For instance, the measured data are not necessarily accurate themselves; certain structural features of a chemical may not be explained by the model descriptors selected, or a sorption mechanism other than van der Waals interaction and hydrophobic bounding could occur. The linearity of the model can be tested by examining the residuals of the regression against the experimental K_{OC} values to see if there is any systematical error in the model. It is shown that the residuals are more or less randomly scattered about zero, indicating the appropriateness of using a linear model in general. There is no overwhelming trend evidence, except for a slight increase in the error over the range of the experimental log K_{OC} from −0.31 to 7.34.

The model, as finally developed, explains 97% of the variation in the experimental data, leaving only 3% for imperfections. Since the measured K_{OC} were collected from different sources, it is reasonable to assume that the database contained a certain amount of error from the beginning. Presumably, the difference in experimental procedures, data interpretation and presentation, and experimental materials may all contribute to discrepancies in the measured data. For the hydrophilic compounds, error may occur when soil samples with low carbon content are used in an experiment. In these cases, the chemicals sorbed on organic carbon may be overestimated due to the effects of clay minerals. Experimental K_{OC} values can vary up to one log-unit as a result of variation in laboratory measurement procedures (31). For those data reported in K_{OM}, a conversion value of 1.724 was adopted to derive the K_{OC} used in this study. Although this ratio is generally accepted, it has been suggested that the ratio may range from 1.9 to 2.5 (32). These rationales provide logical reasons for the differences among the experimental K_{OC} for a given chemical in the database. To investigate the uncertainty in the experimental K_{OC} data, the variations among duplicated values collected for these overlapping chemicals in the database were calculated against their own mean values. The differences of individual values from the means are illustrated in Figure 5 along with the model predicted residuals as frequency distribution diagrams.

Both the predicted and experimental K_{OC} residuals are symmetrically distributed. The mean residual of the predicted K_{OC} based on the regression model was 0.366 log-units, while the mean residual of the experimental K_{OC} derived from the overlapping data was 0.224 log-units. Since the model was developed based on medians of the overlapping experimental K_{OC}, the uncertainty of the data has already been reduced to a certain extent by this exercise. The two mean residuals cannot be compared directly. Still, it can be seen that a portion of the prediction error could be the result of experimental uncertainty. A more accurate model can be developed through use of a more accurate experimental data set. As recommended by Leo, who developed the fragment constant method for predicting K_{OW}, the average residual in log-units (0.366 for this method) can be used as the error term of uncertainty for model application (33).

Robustness of the Prediction Model. The robustness of the prediction model was evaluated using a set of modified jackknife tests in a number of ways. For the random deletion of 100 single chemicals one at a time, the frequency distribution of the 100 jackknifed R^2 thus calculated is plotted in Figure 6 as a histogram.

Most of the calculated jackknifed R^2s are slightly higher than the original value (0.9696) derived from the general model with a few lower than the value. The robustness of the model with regard to a majority of the individual chemicals was demonstrated. Among the 100 chemicals randomly picked from the database, the jackknifed R^2 of pendimethalin (0.9703) is slightly higher than those of the others (Figure 6). The measured log K_{OC} of the chemical was 2.21, while a value as high as 4.00 was predicted. It appears that for those chemicals with relatively high prediction residuals, the jackknifed R^2s tend to be higher than those of the others. To test this assumption, a jackknife test was performed on the 34 compounds having the highest residuals among all data. The jackknifed R^2s are plotted against their prediction residuals in Figure 7. A linear increase in the jackknifed R^2 with increase in the residuals can be clearly seen with only one exception.

It can be concluded that the robustness of the prediction model was high and the largest deviation from the original R^2 (0.9696 for the general model) was only 0.0013 (0.9709 for deletion of 1,2,3,4-tetrachloro-5,6-dimethyloxybenzene from the data set). The robustness of the general model was influenced most significantly by the chemicals having

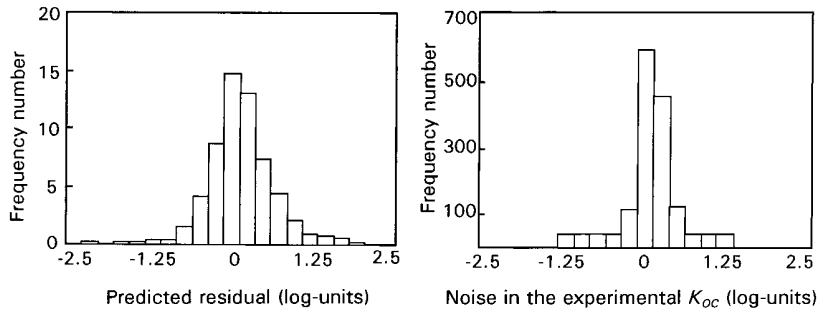

FIGURE 5. Frequency distributions of the predicted and the experimental residuals.

TABLE 2. Fragment Constants Derived from the Final Modeling on 592 Chemicals

fragment constant	f	f^{ϕ}	$f^{I\phi}$	$f^{\phi\phi}$	fragment constant	f	f^{ϕ}	$f^{I\phi}$	$f^{\phi\phi}$
Without C or H					**With H, Without C**				
−F	1.149	0.087			−H	1.487			
−Cl	0.523	0.439			−OH	−0.300	−0.176		
−Br	0.558	0.404			−OP(O)(NH−)O−		−1.715		
−N═	−0.545	−0.626		−0.562	**With C and H**				
−O−	−0.584	−0.723		−0.719	−C(O)H	−1.109			
−S−	−0.074	−0.307			−C(O)OH	−0.678	−0.425		
$-NO_2$		0.168			−C(O)NH−	−1.406	−0.534	−0.875	−2.515
$-SO_2-$	−0.970	−1.153			$-C(O)NH_2$		−0.315		
−S(O)−	−0.709	−1.133			−OC(O)NH−		−0.837	−0.600	−0.522
−SP(S)(O−)O−	0.002				$-C(O)ONH_2$		−0.479		
−OP(O)(O−)O−	−0.964				−CH═N−		−1.738		
−OP(S)(O−)O−		−0.456			−HNC(O)NH−		−1.158		
−OS(O)O−		−1.455			$-HNC(O)NH_2$		−0.219		
−P(S)(O−)O−	−0.934				−CH═NOC(O)NH−	−0.533			
−SP(O)(O−)O−	−1.309	−1.204			−HNC(O)N		−1.476	−1.204	
−P(O)	−1.906				−HNC(O)NO−		−1.026		
−P(S)	−1.160				−S(O)(O)NHC(O)HN−				−1.571
With C, without H					**Fused in Aromatic Ring** (f^{AR})				
−C−	0.519	0.423		0.010	−C(H)═	0.305			
$-CF_3$		0.521			−C═	0.251			
−C≡N		0.075			−N═	−0.308			
−C(O)N	−1.767		−1.833		−S−	0.748			
−C(O)−	−1.355	−0.839			−C(O)−	−0.898			
−C(O)O−	−0.434	−0.427			−N	−0.739			
−C═N−	−0.272				−NH−	0.413			
−OC(O)N	−1.831				−O−	0.533			
−SC(O)N	−0.365								

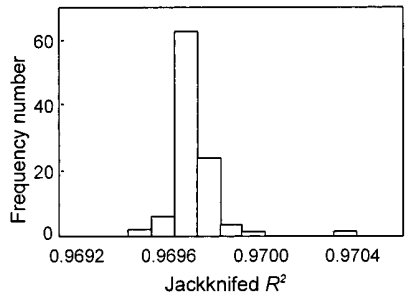

FIGURE 6. Frequency distribution of the jackknifed R^2 calculated based on 100 single-chemical random deletion.

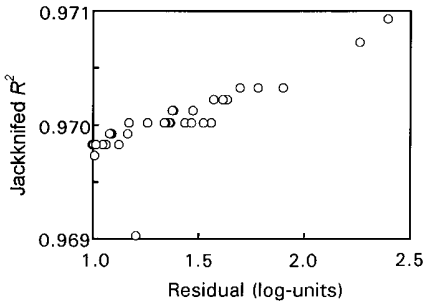

FIGURE 7. Plot of the jackknifed R^2 against the residual for individual chemicals with residuals above 1.0 log-units.

residuals above 1.40 log-units.

A subset jackknife deletion test was conducted to examine the robustness of the derived regression model based on the 592 chemicals. Groups of 50 chemicals in the data set were deleted in turn, and the regression was carried out for the remaining 542 chemicals. The deleted chemicals were randomly selected from the database, and each observation was deleted at least once, while no observation was deleted more than five times. The procedure was repeated 40 times, and the results are shown in Figure 8.

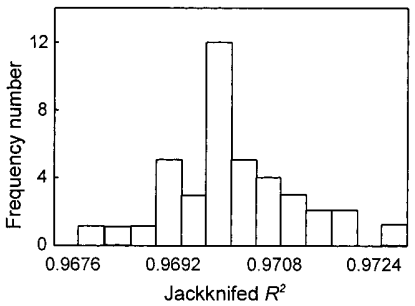

FIGURE 8. Frequency distribution of the R^2 as the result of a jackknife test with random deletion of 50 chemicals.

TABLE 3. Structural Correction Factors Derived from the Final Modeling on 592 Chemicals

feature	factor	feature	factor	feature	factor	feature	factor
F_b	0.088	F_{mhG1}	0.045	F_{P2}	0.075	F_{HP2}	−0.107
F_b^o	0.256	F_{mhG2}	0.131	F_{P3}	0.429	F_C	−1.023
F_{CBr}	−0.102	F_{mhG3}	−0.035	F_{P4}	0.636	F_{CH}	−0.731
$F_=$	0.114	F_{mhV}	0.009	F_{P5}	−0.085	F_{CH_2}	−0.376
$F_=^o$	0.126	F_{P0}	0.560	F_{P6}	0.166	F_{NH}	0.430
$F_≡$	−1.028	F_{P1}	0.373	F_{HP1}	−0.133	F_{NH_2}	0.735

The results shown in Figure 8 indicate that jackknifed R^2s (0.9680−0.9727) are more or less randomly scattered about their mean value of 0.9700 which is very close to the R^2 derived from the full data set (0.9696). It would also be desirable to examine the residuals on a class-by-class basis. The 592 chemicals used in this study were arranged into 17 classes of compounds in Table 4. A class deletion jackknife test was carried out so that each class of chemicals was removed from the database once, while a regression analysis was conducted using the remained 16 classes. The jackknifed R^2s thus obtained are tabulated in the last column of Table 4.

A brief examination of the jackknifed R^2s of various groups shows that only PAH (0.9652) and PCB (0.9691) have

TABLE 4. Result of the Category Deletion Jackknife Test

no.	compound class	no. of chemicals	absolute error	relative error (%)	R^2
1	acid	30	0.522	32.0	0.9722
2	ether	15	0.680	27.5	0.9727
3	alcohol, phenol	55	0.486	23.6	0.9718
4	amine, aniline	50	0.441	16.5	0.9710
5	organophosphate	57	0.346	15.5	0.9705
6	ester	28	0.438	14.7	0.9696
7	carbamate, thiocarbamate	45	0.418	14.7	0.9696
8	diazole, triazole, pyridine, pyrimidine, triazine	40	0.283	14.6	0.9707
9	monocyclic aromatic hydrocarbon	16	0.382	13.5	0.9697
10	azo, nitrobenzene, nitrile	13	0.409	12.9	0.9701
11	aldehyde, ketone	22	0.340	12.8	0.9703
12	amide, acetanilide, benzamide	51	0.248	12.3	0.9699
13	urea	56	0.291	12.2	0.9702
14	halogenated hydrocarbon	45	0.243	9.0	0.9691
15	chlorinated biphenyl	20	0.343	6.7	0.9669
16	polycyclic aromatic hydrocarbon	42	0.246	5.2	0.9652
17	others	7	0.123	3.8	0.9693

TABLE 5. Sample Calculations[a]

chemical	formula	exp K_{OC}	residual
oxydemeton−methyl $[-S(O)-] + [-SP(O)(O-)O-] + 6[-C] + 6[F_b] + [F_{P3}] + 3[F_{CH_2}] = -0.7092 - 1.3090 + 3.1113 + 0.5303 + 0.4291 - 1.1293 = 0.9232$	$C_6H_{15}O_4PS_2$	1.10	0.177
3,5,6-trichloro-2-pyridyloxy acetic acid $[-C(H)=] + 4[-C=] + [-N=^{AR}] + 3[-Cl^{\phi}] + [-O-^{\phi}] + [-C] + [-C(O)OH] + 2[F_b] + [F_{P1}] + [F_{P2}] + [FH_{P2}] + [F_{CH_2}] = 0.3047 + 1.0051 - 0.3081 + 1.3166 - 0.7234 + 0.5185 - 0.6781 + 0.1768 + 0.3731 + 0.0748 - 0.1074 - 0.3764 = 1.5761$	$C_7H_4Cl_3NO_3$	1.43	−0.146
N-(1,1-dimethyl-2-propynyl)benzamide $5[-C(H)=] + [-C=] + 5[-C] + [-C(O)NH-^{\phi}] + 5[F_b] + 2[F_{CBr}] + [F_{\equiv}] + [FC] + [F_{CH_2}] = 1.5237 + 0.2513 + 2.5927 - 0.5337 + 0.4419 - 0.2038 - 1.0277 - 1.0231 - 0.3764 = 1.6448$	$C_{12}H_{13}NO$	1.54	−0.105
1,1,2-trichloroethane $3[-Cl] + 2[-C] + 3[F_b] + 2[F_{mhG1}] + 2[F_{mhv}] + [F_{CH}] + [F_{CH_2}] = 1.5701 + 1.0371 + 0.2651 + 0.0892 + 0.0170 - 0.7309 - 0.3764 = 1.8712$	$C_2H_3Cl_3$	1.75	−0.121
2-chloro-4-(ethylamino)-6-isopropylamino-1,3,5-triazine $3[-C=] + 3[-N=^{AR}] + [-Cl^{\phi}] + 2[-N=^{\phi}] + 5[-C] + 5[F_b] + [F_{CBr}] + 4[F_{P2}] + 3[F_{P6}] + 2[F_{HP2}] + [F_{CH}] + [F_{CH_2}] + 2[F_{NH}] = 0.7538 - 0.9243 + 0.4389 - 1.2527 + 2.5927 + 0.4419 - 0.1019 + 0.2990 + 0.4973 - 0.2149 - 0.7309 - 0.3764 + 0.8604 = 2.2829$	$C_8H_{14}ClN_5$	2.17	−0.113
tetrachloroethylene $4[-Cl] + 2[-C] + 4[F_b] + [F=] + 4[F_{mhG1}] + 2[F_{CH}] = 2.0934 + 1.0371 + 0.3535 + 0.1143 + 0.1784 - 1.4618 = 2.3149$	C_2Cl_4	2.56	0.245
benzidine $8[-C(H)=] + 4[-C=] + 2[-N=^{\phi}] + 2[F_{NH_2}] = 2.4380 + 1.0051 - 1.2527 + 1.4690 = 3.6593$	$C_{12}H_{12}N_2$	3.46	−0.199
N-sec-butyl-4-*tert*-butyl-2,6-dinitroaniline $2[-C(H)=] + 4[-C=] + [-N=^{\phi}] + 7[-C] + [-C^{\phi}] + 2[-NO_2{}^{\phi}] + 7[F_b] + 3[F_{CBr}] + [F_C] + [F_{CH}] + [F_{CH_2}] + [FNH] = 0.6095 + 1.0051 - 0.6264 + 3.6298 + 0.4229 + 0.3367 + 0.6186 - 0.3057 - 1.0231 - 0.7309 - 0.3764 + 0.4302 = 3.9903$	$C_{14}H_{21}N_3O_4$	3.91	−0.080
S-{((*p*-chlorophenyl)thio)methyl}*O*,*O*-diethyl phosphorodithioate $4[-C(H)=] + 2[-C=] + [-SP(S)(O-)O-] + [-Cl^{\phi}] + 5[-C] + [-S-^{\phi}] + 6[F_b] + [F_{CBr}] + [F_{P1}] + 3[F_{CH_2}] = 1.2190 + 0.5025 + 0.0017 + 0.4389 + 2.5927 - 0.3067 + 0.5303 - 0.1019 + 0.3731 - 1.1293 = 4.1204$	$C_{11}H_{16}ClO_2PS_3$	4.66	0.540
1,2-benzofluorene $10[-C(H)=] + 6[-C=] + [-C^{\phi\phi}] + [F_b{}^{o}] + [F_{CH_2}] = 3.0474 + 1.5076 + 0.0101 + 0.2560 - 0.3764 = 4.4447$	$C_{17}H_{12}$	5.46	1.015
2,3,4,2′,3′,4′-hexachlorobiphenyl $4[-C(H)=] + 8[-C=] + 6[-Cl^{\phi}] = 1.2190 + 2.0102 + 2.6332 = 5.8623$	$C_{12}H_4Cl_6$	6.42	0.558
3-(trifluoromethyl)phenylurea $4[-C(H)=] + 2[-C=] + [-CF_3{}^{\phi}] + [-HNC(O)NH_2] = 1.2190 + 0.5025 + 0.5213 - 0.2187 = 2.0241$	$C_8H_7F_3N_2O$	1.96	−0.064
phenylurea $5[-C(H)=] + [-C=] + [-HNC(O)NH_2{}^{\phi}] = 1.5237 + 0.2513 - 0.2187 = 1.5563$	$C_7H_8N_2O$	1.35	−0.206
trichlorofluoromethane $[-F] + 3[-Cl] + [-C] + 3[F_b] + 4[F_{mhG3}] + [FC] = 1.1488 + 1.5701 + 0.5185 + 0.2651 - 0.1417 - 1.0231 = 2.3377$	CCl_3F	2.20	−0.138

[a] −C is same as >C< in Table 2; all superscripts (ϕ, $1/\phi$, $\phi\phi$, AR) are the same as those used in Table 2.

jackknifed R^2s lower than 0.9696 (obtained in the case without deletion), while all others in the set are reasonably constant (between 0.9696 and 0.9727). The test result did not indicate any unduly positive high variation, suggesting that the model was not biased by any of the particular chemical groups studied. According to the results listed in Table 5, it appears that the jackknifed R^2 is positively correlated to the residual also listed in the table. A trend is clearly demonstrated by plotting the jackknifed R^2 against the residual in Figure 9.

Considering the variety among the 592 members of the data set and the fact that the experimental K_{OC} vary in a range exceeding seven log-units, the nonclass-specific model allows estimation of K_{OC} over a large variety of organic compounds within the constraints of experimental uncertainties. Because of the large numbers of both the fragment

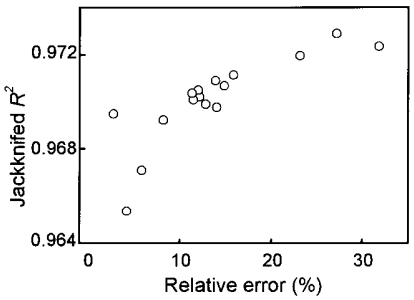

FIGURE 9. Variation in the jackknifed R^2 against relative error for 17 chemical classes.

constants and structural factors, the model developed in this study is a fairly powerful one. There are relatively few synthetic compounds for which a value of K_{OC} could not be calculated based on the model.

Sample Calculations. A number of sample calculations based on 15 representative chemicals from the database are listed in Table 5 to demonstrate how the chemicals were fragmented to derive the predicted results.

Acknowledgments

Funding was provided by The National Natural Science Foundation of China [49632060] and [49525102].

Supporting Information Available

Table of experimental and predicted K_{OC} of 592 chemicals. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Harmaker, J. W. In Environmental Dynamics of Pesticides; Hague, R., Freed, V. H., Eds.; Plenum: New York, 1975; p 115.
(2) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. Chemosphere 1997, 34, 2525.
(3) Gerstl, Z. J. Contam. Hydrol. 1990, 6, 357.
(4) Muller, M.; Kordel, W. Chemosphere 1996, 32, 2493.
(5) Sabljic, A.; Protic, M. Bull. Environ. Contam. Toxicol. 1982, 28, 162.
(6) Meyland, W.; Howard, P. H. Environ. Sci. Technol. 1992, 26, 1560.
(7) Sabljic, A. Environ. Sci. Technol. 1987, 21, 358.
(8) Hammett, L. Physical Organic Chemistry: Reaction Rates, Equilibria and Mechanisms, 2nd ed.; McGraw-Hill: New York, 1970.
(9) Rekker, R. The Hydrophobic Fragmental Constant; Elsevier: Amsterdam, 1977.
(10) Leo, A. J. Symposium on Structure−Activity Correlations in Studies of Toxicity and Bio-concentration with Aquatic Organisms; Great Lakes Research Advisory Board: Burlington, Ontario, 1975; p 151.
(11) Leo, A. J. In QSAR and Strategies in the Design of Bioactive Compounds; Seydel, J. K., Ed.; VCH: Weinheim, 1985; pp 294−298.
(12) Hansch, C.; Leo, A. J. Substituent Constants for Correlation Analysis in Chemistry and Biology; John Wiley: New York, 1979.
(13) Chou, J. T.; Jurs, P. C. J. Chem. Inf. Comput. Sci. 1979, 19, 172.
(14) Hansch, C.; Leo, A. J. Exploring QSAR: Fundamental and Application in Chemistry and Biology; American Chemical Society: Washington, DC, 1995.
(15) Kenaga, E. E.; Goring, C. A. In Aquatic Toxicology, Proceedings of the Third Annual Symposium on Aquatic Toxicology; Eaton, J. G., Parish, P. P., Hendricks, A. C., Eds.; ASTM, STP 707: Philadelphia, 1980; pp 78−115.
(16) Briggs, G. G. Proceeding of the Seventh British Insecticide and Fungicide Conference; 1973; pp 475−478.
(17) Briggs, G. G. J. Agric. Food Chem. 1981, 29, 1050.
(18) Karickhoff, S. W.; Brown, D. S.; Scott, T. A. Water Res. 1979, 13, 241.
(19) Karickhoff, S. W. Chemosphere 1981, 10, 833.
(20) Karickfoff, S. W.; Morris, K. R. Environ. Toxicol. Chem. 1985, 4, 469.
(21) Vowles, P. D.; Mantoura, R. F. C. Chemosphere 1987, 16, 109.
(22) Hassett, J. J. Sorption Properties of Sediments and Energy-Related Pollutants; 1980; EPA-600/3-80-041.
(23) Hassett, J. J. Soil Sci. Soc. Am. J. 1981, 45, 38.
(24) Chiou, C. T.; Peters, L. J.; Freed, V. H. Science 1979, 206, 831.
(25) Chiou, C. T.; Porter, P. E.; Schmeddi, D. W. Environ. Sci. Technol. 1983, 17, 227.
(26) Gustafson, D. I. Pesticides in Drinking Water; Van Nostrand Reinhold: New York, 1993.
(27) Montgomery, J. Agrochemicals Desk Reference-Environmental Data; Lewis Publishers: Boca Raton, FL,1994.
(28) Karel, V. Handbook of Environmental Data on Organic Chemicals; Van Nostrand Reinhold: New York, 1996.
(29) Nirmalakhandan, N. N.; Speece, R. E. Environ. Sci. Technol. 1988, 22, 328.
(30) Dietrich, S. W.; Nicholas, D. D.; Dreyer, D.; Hansch, C. J. Med. Chem. 1980, 23, 1201.
(31) Baker, J. R.; Mihelcic, J. R.; Luehrs, D. C.; Hickey, J. P. Water Environ. Res. 1997, 69, 136.
(32) Nelson, D. W.; Sommers, L. E. In Methods of Soil Analysis, Prt 2; Page, A. L., Ed.; American Society of Agronomy and Soil Science Society of America: Madison, WI, 1982; p 537.
(33) Leo, A. J. In Handbook of Chemical Property Estimation Methods; Lyman, W. J., Rosenblatt, D. H., Eds.; McGraw-Hill: New York, 1982; pp 1−54.

Received for review August 14, 1998. Revised manuscript received April 12, 1999. Accepted May 17, 1999.

ES980833D