Science of the Total Environment 670 (2019) 226–235

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A general-applicable model for estimating the binding coefficient of organic pollutants with dissolved organic matter

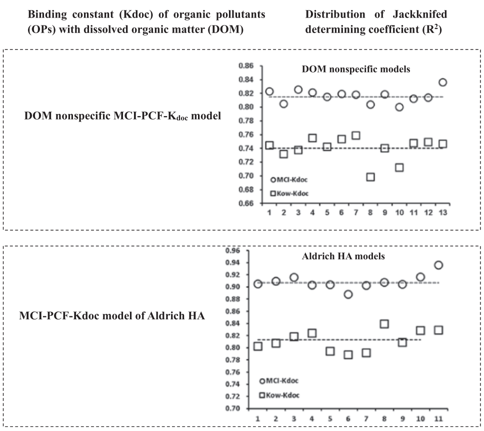
Yi-Long Li [a], Wei He [a,b], Rui-Lin Wu [a], Baoshan Xing [c], Fu-Liu Xu [a,*]

[a] *MOE Laboratory for Earth Surface Processes, College of Urban & Environmental Sciences, Peking University, Beijing 100871, China*
[b] *MOE Key Laboratory of Groundwater Circulation and Environmental Evolution, School of Water Resources and Environment, China University of Geosciences (Beijing), Beijing 100083, China*
[c] *Stockbridge School of Agriculture, University of Massachusetts, Amherst, MA 01003, USA*

HIGHLIGHTS

- Two types of MCI-PCF-K_{doc} models and the Kow-Kdoc models were developed.
- Both MCI-PCF-K_{doc} models exhibited better fit than the Kow-Kdoc models.
- The Aldrich HA model showed higher pertinence to the DOM nonspecific MCI-PCF-K_{doc} model.
- Both MCI-PCF-K_{doc} models were sensitive to the robust parameters.
- Both MCI-PCF-K_{doc} models were not altered with the dipole moment.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 12 December 2018
Received in revised form 9 March 2019
Accepted 9 March 2019
Available online 11 March 2019

Editor: Shiming Ding

Keywords:
Molecular connectivity indices (MCI)
Binding constant (K_{doc})
Multiple linear regression model
Polarity correction factors (PCF)
Dipole moment

ABSTRACT

The binding constant (K_{doc}) of organic pollutants (OPs) with dissolved organic matter (DOM) is an important parameter in determining the partitioning of OPs in the aquatic environment. Most estimation models have focused on calculating the K_{doc} of a specific group of OPs but failed to obtain K_{doc} values of different OPs effectively over the last three decades. In this study, we attempted to build a general-applicable K_{doc} model based on various organic compounds' K_{doc} values from the literature since 1973. Two multiple linear regression models, a DOM nonspecific model and an Aldrich HA model, were developed based on two solid and easy to access parameters—molecular connectivity indices (MCI) and polarity correction factors (PCF). In addition, the models' corresponding K_{ow}-K_{doc} models, which were mostly used in previous model studies, were developed for comparison. The adjusted determining coefficient (adj-R^2) and standard error of the estimate (SEE) of the DOM nonspecific MCI-PCF-K_{doc} model were 0.815 and 0.579, respectively, whereas the adj-R^2 and SEE for the MCI-PCF-K_{doc} model of Aldrich HA reached 0.907 and 0.438, respectively. The Aldrich HA model showed higher pertinence to the nonspecific model. Furthermore, both models exhibited better fit than the K_{ow}-K_{doc} models. The dipole moment modification attempts did not significantly improve either MCI-PCF-K_{doc} models; hence, the two models were not altered with the dipole moment. The robustness tests by a Jackknifed method showed that the two MCI-

* Corresponding author.
E-mail address: xufl@urban.pku.edu.cn (F.-L. Xu).

https://doi.org/10.1016/j.scitotenv.2019.03.146

PCF-K_{doc} models exhibited higher robustness than the K_{ow}-K_{doc}. Of all of the OPs, the phenols contributed the most to their robustness. Furthermore, a sensitivity analysis showed that the two MCI-PCF-K_{doc} models were sensitive to the robust parameters.

1. Introduction

Organic pollutants (OPs) are a series of organic materials that are harmful or potentially harmful to humans or the eco-system. The interaction between OPs and dissolved organic matters (DOMs) is of great concern because it helps to elucidate the environmental fate of OPs, such as their transport, transfer, and adsorption in water bodies (Cho et al., 2002; Laor and Rebhun, 1997). Generally, DOMs are believed to increase the solubility of certain OPs due to the formation of DOM-OP colloids, and hence, affect the partitioning of OPs in groundwater, pore water and sediments (Backhus and Gschwend, 1990; Johnson and Amy, 1995). Many studies focus on the DOM's effect on the bioavailability of OPs. Agnola et al. (1981) reported that the presence of DOM can reduce atrazine's inhibitory action on sulfate uptake of terrestrial plants, and other researchers found the intake and enrichment of OPs by aquatic organisms is generally lowered by non-low level DOMs (Boehm and Quinn, 1973; Haitzer et al., 1998; Hassett and Anderson, 1982; Landrum et al., 1987). For example, Kukkonen and Oikari (1991) reported that the bioconcentration factors (BCFs) of benzo[*a*]pyrene on water flea decreases when the DOM concentration increases in 20 natural waters. A potential mechanism of this bioavailability influence by DOMs is the binding/adsorption effect of OPs to DOMs. The free OPs in the water system are more likely to be concentrated by organism uptake, whereas the binding parts are not, and therefore, the DOM-bond OPs show less toxicity effects (Agnola et al., 1981; Burns et al., 1996; Johnson and John, 1999; Krop et al., 2001). In other words, obtaining the partitioning ratio of a certain OP between the DOMs and the water phase, which was the binding constant—K_{doc}, can be regarded as a basic step for further analysis of its environmental fate.

Typically, some OPs of priority concern, including polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), dichlorodiphenyltrichloroethanes (DDTs), hexachlorocyclohexanes (HCHs), and di-(2-ethylhexyl) phthalate (DEHP), were measured experimentally since the early 1970s (Backhus and Gschwend, 1990; Carter and Suffet, 1983; Carter and Suffet, 1982; Chiou et al., 1987; Chiou et al., 1986; Cho et al., 2002; Khan, 1973; Krop et al., 2001; Lafrance et al., 1991; Laor and Rebhun, 1997; Mackay et al., 2010; Rav-Acha and Rebhun, 1992; Traina et al., 1989). Currently, the K_{doc} values of emerging organic pollutants, such as endocrine disrupters, pyrethroids, pharmaceutical and personal care products (PPCPs), have been investigated (Delgado-Moreno et al., 2010; Lee et al., 2011; Maoz and Chefetz, 2010). Although plenty of experimental methods for the determination of K_{doc} are available, it is yet hardly to experimentally measure the K_{doc} of hundreds of thousands of new compounds, which are being synthesized for commercial application each year. As a result, model prediction is often utilized to predict the K_{doc} values of certain compounds based on the existing K_{doc} values of OPs with similar molecular structures. Presently, the proposed quantitative structure-activity relationship (QSAR) models for K_{doc} can be classified into three categories: (1) models based on physicochemical parameters, such as the linear free energy relationship (LFER) model that primarily uses the octanol-water partition coefficient (K_{ow}) as input variable, the linear solvation energy relationship (LSER) model that uses water solubility (s) as input variable, and other linear relationship models that use other parameters, such as absorptivity; (2) models based on solvation theory, such as Flory-Huggins solvation theory; and (3) models based on structure parameters such as fragment constants or topological indices, such as molecular connectivity indices (MCIs) (Krop et al., 2001; Lu et al., 2000a; Yu et al., 1990). The majority of existing models are built with data of specific OP classes with abundant congeners, such as PAHs and PCBs, thus they cannot successfully predict the K_{doc} values of other OPs, especially emerging OPs. Therefore, building a prediction model, which is suitable for OPs of various classes, is an important orientation for the K_{doc} model study. Among all of the models, the K_{ow}-based LLER-K_{doc} model has the simplest form and thus is most widely used. However, due to the difficulty in obtaining experimental K_{ow} values of new compounds in time, the applicability of the K_{ow}-K_{doc} model is limited. Neale proposed a pp-LFER K_{doc} model mainly based on disinfection by-products and few other chemicals like halogenated alkanes and alkenes (Neale et al., 2012). Although its potential application range is a little wider compared to the single OP class models, the difficulty in obtaining accurate model parameters is still a limitation. The Flory-Huggins models have a clear theory, whereas the majority of the parameters are hard and complicated to calculate or quantify, which also limits their applicability. MCIs are a group of molecule descriptors based on Randic's branching theory. Modified by Kier and Hall (1986), MCIs quantitatively describe the topological relationships of the non-hydrogen structure of organic compounds to comprehensively reflect their molecular shapes, volumes and electrical information (Li et al., 2000; Lu et al., 2000c). Due to the relatively large amount of structural information MCI contains and the ready accessibility of MCI values for all chemicals (MCIs can be calculated as long as the molecular structure is known), MCIs have already been utilized in predicting several physicochemical properties including Henry's law constant, K_{ow}, solubility, BCF, K_{doc} and a similar binding constant—K_{oc} (soil organic matter) (Lu et al., 2000c; Nirmalakhandan and Speece, 1988; Pavan et al., 2006; Sabljić and Protić, 1982). Because MCIs are very easy to calculate and the affiliation property of any organic molecule theoretically depends on its structure, the MCI-K_{doc} model is potentially applicable for almost any OPs. Besides, taking steady structural related parameters like MCI as direct model inputs can also erase the inconvenience and the deviation resulting from secondary calculation when the parameters in other models need to be calculated first from chemical structures, especially for emerging OPs.

For the construction of MCI-based binding constant models, Sekušak and Sabljić reported that there were clear linear relationships between the low-order path-type MCIs and the corresponding K_{oc} values of the acetanilides, amides, dinitroanilines, and triazoles (Sabljic, 1987; Sekušak and Sabljić, 1992). Therefore, it was possible to build the K_{oc} prediction models upon the MCIs for these OPs. Sabljic then established a linear MCI-K_{oc} model for four groups of OPs with experimental K_{doc} values and calculated $^{1}\chi_{p}$ values, and the model was adjusted by semi-empirical variables (Sabljic, 1987). In regard to the MCI-K_{doc} models, Sabljic found that the K_{doc} values of 26 PCBs and their $^{1}\chi_{P}$ showed a good parabolic relationship (Sabljić, 1991), whereas Evers and Velzen showed that the K_{doc} values of polychlorinated dibenzopdioxins (PCDDs) had good linear relationships with their $^{1}\chi_{P}^{V}$ or $^{2}\chi_{P}^{V}$ (Evers et al., 1991). Although those proposed MCI-K_{oc}/K_{doc} models had good regressions, their applications were quite limited and applicable within nearly only one group of OPs. To broaden the applicability of the MCI-based binding constant models, Lu and her coworkers analyzed the relationship between K_{oc} values and the MCIs of 330 OPs and found that a good multiple linear regression model can be established for nonpolar OPs, whereas polarity correction factors are needed for polar OPs (Lu et al., 1999b; Lu et al., 1999c; Tao and Lu, 1999). In addition, Lu also raised two MCI-BCF models, which are applicable for multiple OP groups (Lu et al., 1999a, 2000b). According to the above studies, MCI models showed advantages for the prediction of the physicochemical properties for multiple OPs. Conversely, there are presently no sufficient K_{doc} models based on multi-class OPs. Therefore, the objective of our study is (1) to build MCI-K_{doc} prediction models suitable for multiple OP groups and (2) to test the model robustness and compare them with the most commonly used K_{ow}-K_{doc} models.

2. Materials and methods

2.1. Data collection and calculation

Experimental K_{doc} values from 1973 to 2013 were first selected from the SCI publication database. Only the K_{doc} values where the studied compound was predominately neutral were considered for model construction, thus compounds like Diquat (dichloride) are disregarded. In the end, the dataset contains 202 OPs of more than ten OP groups including PAHs, PCBs, polybrominated diphenyl ethers (PBDEs), phenols, halogenated PAHs (X-PAHs), PCDDs, organic chlorine pesticides (OCPs), amides, pyrethroids, triazines, etc. Among 202 OPs, a total of 127 chemicals had more than one reported K_{doc} values based on different methods and conditions. Based on previous study suggestions (Lu et al., 1999a, 2000a, 2000b, 2000c; Tao and Lu, 1999; Tao et al., 2000, 2001) and relatively small differences between the median and the average logK_{doc} values (0.08 log unit on average for 70 OPs with over two K_{doc}), the typically used median values of each chemical were then calculated and used for model construction to minimize the variations of K_{doc} values associated with experiment conditions (e.g. methods, temperature, and pH). The K_{doc} values are normally shown in the log-form, and the logK_{doc} (median) values of all 202 OPs ranged from 0.95 to 7.14. Furthermore, the relative logK_{ow} values for each of the OPs were also selected (shown in Supporting Information, SI). The structural formulas of all OPs were searched in the Pubchem database (https://www.ncbi.nlm.nih.gov/pccompound) and then 22 common MCIs ($^{0}\chi_{P}$, $^{1}\chi_{P}$, $^{2}\chi_{P}$, $^{3}\chi_{P}$, $^{4}\chi_{P}$, $^{5}\chi_{P}$, $^{6}\chi_{P}$, $^{3}\chi_{C}$, $^{4}\chi_{PC}$, $^{5}\chi_{PC}$, $^{6}\chi_{PC}$ and $^{6}\chi_{CH}$; $^{0}\chi^{V}_{P}$, $^{1}\chi^{V}_{P}$, $^{2}\chi^{V}_{P}$, $^{3}\chi^{V}_{P}$, $^{4}\chi^{V}_{P}$, $^{5}\chi^{V}_{P}$, $^{6}\chi^{V}_{P}$, $^{3}\chi^{V}_{C}$, $^{4}\chi^{V}_{PC}$, $^{5}\chi^{V}_{PC}$, $^{6}\chi^{V}_{PC}$ and $^{6}\chi^{V}_{CH}$) of each OP were calculated as follows.

$$^{m}\chi = \sum_{j=1}^{n}\left(\prod_{i=1}^{m+1}\delta_i\right)^{-0.5}, \quad ^{m}\chi^{V} = \sum_{j=1}^{n}\left(\prod_{i=1}^{m+1}\delta_i^{V}\right)^{-0.5} \qquad \text{(A)}$$

where δ is the atomic delta value, δ^{V} refers to the valence atomic delta value, i indicates the non-hydrogen atom number, and m and n correspond to the connectivity level and the subgraph number, respectively (Lu et al., 1999a; Lu et al., 2000c). There are four MCI subclasses according to their subgraph differences—path, cluster, path/cluster and chain-type indices (Li et al., 2000). Although there is no specific corresponding physicochemical meaning for each subclass, a number of studies suggest that each subclass describes a different aspect of the structure properties. Low-order path-type indices generally describe more about molecular size, surface and volume with $^{1}\chi_{P}$ and $^{0}\chi^{V}_{P}$ are believed to relate well with molecular surface area and volume, respectively. The normal cluster and path/cluster-type indices, such as $^{3}\chi_{C}$ and $^{4}\chi_{PC}$, mainly describe the extent of different branching in a molecule and they are very sensitive when there are branching changes. Furthermore, $^{4}\chi_{PC}$ also contains information of substitution pattern on benzene rings. In addition, the chain-type indices like $^{6}\chi_{CH}$ reflect the rings in the molecule and the substitution patterns on these rings (Kier and Hall, 1986; Sabljić, 1991). Technically, those indices were calculated in Wintox software by Jorgensen and Sorensen for construction of the model (Jorgensen et al., 1997).

2.2. Model development

Based on previous studies (Lu et al., 1999a; Lu et al., 2000b; Lu et al., 2000c; Lu et al., 1999c; Tao and Lu, 1999), the following steps were set for the construct and test MCI-K_{doc} models in this study: first, 22 indices were selected by stepwise multiple linear regressions to build the MCI-K_{doc} model; second, the obtained regression models from step one were evaluated by the fitting results and then modified with other parameters if necessary; and third, the stability and sensitivity of the models were tested to further evaluate the model performance. The stability tests were conducted using a modified Jackknifed method (Lu et al., 2000a; Tao et al., 2000; Tao et al., 1999). The basic idea of the method is to randomly remove a group of the modeling data in the model construction, and then compare the deviations of each parameter among the original model and different removal choices to evaluate the robustness of the model and its influence factors. Finally, the MCI-K_{doc} models were compared with the most commonly used K_{ow}-K_{doc} model for judgment of the advantage. All of the data simulation and analysis were conducted by IBM SPSS 20.0.

3. Results and discussion

3.1. Development of overall MCI-K_{doc} models

In the previous studies (Lu et al., 1999a; Lu et al., 2000b; Lu et al., 2000c; Lu et al., 1999c; Tao and Lu, 1999), there were significantly differences between polar and nonpolar OPs when establishing some physicochemical property models (K_{oc} and BCF) based on MCIs. Therefore, our studies utilized their classification method to divide the 202 OPs into 104 nonpolar OPs, which only have carbon (C), hydrogen (H) and halogen (X) in their molecular structure and 98 polar OPs as the rest of the tested OPs. However, to test whether there is a same discrimination pattern for polar and nonpolar OP groups in MCI-K_{doc} models, an overall MCI-K_{doc} model based on all 202 compounds was first established (Eq. (1)). In addition, the corresponding K_{ow}-K_{doc} model was built for a comparison purpose (Eq. (2)). However, there were no reported K_{ow} values for 9-formylanthracene, anthraquinone, and napropamide, and therefore, the K_{ow}-K_{doc} model is based on the remaining 199 OPs.

$$\begin{aligned}\log_{10}K_{doc} &= -0.716^{1}\chi_{P} + 0.424^{3}\chi_{P} + 9.99^{6}\chi_{CH} \\ &+ 0.487^{0}\chi^{V}_{P} - 0.086^{6}\chi^{V}_{PC} + 1.193; n \\ &= 202, adj{-}R^{2} = 0.725, SEE = 0.705\end{aligned} \qquad (1)$$

$$\begin{aligned}&\log_{10}K_{doc} = 0.598\ \ \log_{10}K_{ow} + 1.441; \\ &n = 199, adj{-}R^{2} = 0.740, SEE = 0.689\end{aligned} \qquad (2)$$

As analyzed by the stepwise multiple linear regression, 5 indices were selected for the MCI-K_{doc} model (selected MCI values for each chemical are shown in SI). These indices cover both general ($^{1}\chi_{P}$, $^{3}\chi_{P}$, and $^{0}\chi^{V}_{P}$) and local ($^{6}\chi^{V}_{PH}$ and $^{6}\chi_{CH}$) molecular information of the studied OPs. As mentioned before, $^{1}\chi_{P}$ (molecular surface area), $^{0}\chi^{V}_{P}$ (molecular volume) and $^{6}\chi_{CH}$ (rings & substitution patterns on the rings) have relatively clear structural meanings, whereas $^{3}\chi_{P}$ and $^{6}\chi^{V}_{PC}$ may be related to molecular density and branches (Kier and Hall, 1986). The selection of these indices indicates these structural information of OPs all influence on their binding ability to DOM. However, MCIs are non-dimensional parameters and do not have specific physicochemical meanings, so the coefficients in Eq. (1) do not necessarily reflect the relatively contribution of the chemical properties on the binding abilities. Although the coefficient of $^{6}\chi_{CH}$ is remarkably high, the 9.99 $^{6}\chi_{CH}$ term is generally of same order of magnitude with other terms due to the fact that the $^{6}\chi_{CH}$ values are quite small (0–0.34 for studied OPs) compared to other MCIs. In Sabljic's study based on single MCI, and molecular surface area (indicated by $^{1}\chi_{P}$) was found to have a parabolic relationship with K_{doc} values (Sabljić, 1991), however, in our study, the negative coefficient indicates that the molecular surface area may have overall a negative contribution to K_{doc} values when other structural information indices are also considered. This is in line with Lu's two MCI-K_{oc} models (Lu et al., 2000a; Lu et al., 2000b).

The modeled residual values as well as the errors on all coefficients of both equations are listed in Table S1 of the Supplementary materials, and the relationships between the original and modeled values are shown in Fig. 1. For Eq. (1), the average value of the absolute residuals is 0.50 log-unit (with 36.6% OPs over 0.50 log-unit), and the average residual of Eq. (2) is 0.53 log-unit (with 45.0% OPs over 0.50 log-unit). When comparing the two models, the residual level of the MCI-K_{doc} model is slightly better while the K_{ow}-K_{doc} model has little advantage in fitting effect (marked by the adjusted coefficient of determination (adj-R^2) value and standard error of the estimate (SEE)) and model complexity. Therefore, the overall MCI-K_{doc} model is not yet a potential substitution option for the corresponding K_{ow}-K_{doc} model. Furthermore, the fittings themselves for both models are not sufficient enough for the prediction of K_{doc} values, as indicated by the scatter points away from the Y = X reference line within the whole logK_{doc} range (Fig. 1a and b). Therefore, certain modifications should be conducted for the overall MCI-K_{doc} model.

To modify the MCI-K_{doc} model, the differences between the polar OPs and the nonpolar OPs raised by Lu (Lu et al., 1999a, 2000b, 2000c, 1999c; Tao and Lu, 1999) were first considered. However, unlike their MCI-K_{oc} model studies, the polar spots and the nonpolar spots in our MCI-K_{doc} model do not show a clear distributional difference around the Y = X line ($p > 0.01$). On one hand, this finding is probably observed because the interactions between the OPs and DOMs are more intense than the soil organic matters since the polarity of DOMs are stronger (Krop et al., 2001). On the other hand, the definition of nonpolar OPs in these studies were not strict; compounds with halogen atoms, such as PCBs and X-PAHs, were classified as nonpolar and it might cause bias because the electron-withdrawing properties of the chlorine and bromine atoms in OPs are believed to infect their abilities to bind DOMs (Nuerla et al., 2013). Therefore, the pattern in the MCI-K_{oc} models might not be suitable for our K_{doc} study, and the modification of MCI-K_{doc} model was not separated for polar and nonpolar groups in this study. However, the polarity correction factor (PCF) of the polar MCI-K_{oc} model raised by Lu (Lu et al., 1999a, 2000b, 2000c, 1999c; Tao and Lu, 1999) was still considered in our modification for all OPs since the polar functional groups in both polar and nonpolar OPs were believed to be responsible for the deviation of the Y = X line.

The modification of PCF followed two assumptions: (1) the degrees of binding influence depend upon polar functional groups, and each polar group needs an independent factor to describe its contribution to the K_{doc} values; and (2) the factors are attributed by the polar functional groups and their corresponding amounts. By analyzing the molecular structure of 202 OPs, 15 factors, including factors of hydroxyl (—OH), amino (—NH2, —NH—, —N—), azo/nitrile (—N═/—CN), nitro (—NO2), carbonyl (—CO—), aminocarbonyloxy (—NCOO—), oxycarbonyl (—COO—), carboxyl (-COOH), oxy(—O—), sulfur(—S—), phosphor(—PO3) and four halogens (—F, —Cl, —Br, —I) were screened and then introduced to the original MCI-K_{doc} model. In this way, the model was revised as follows:

$$\log_{10}K_{doc} = \sum_i a_i \cdot \chi_i + \sum_i n_i \cdot F_i + c \tag{B}$$

where χ_i is the MCI selected in the original model and a_i refers to the new coefficient of each MCI; F_i indicates the PCF of the i-th polar functional group and it is calculated as the coefficient of n_i, where n_i is the number of polar functional groups i in a molecule, and c is the intercept of the model.

After introducing the PCFs in Table 1, the new MCI-K_{doc} model, namely, the MCI-PCF-K_{doc} model, was obtained and is shown in Eq. (3).

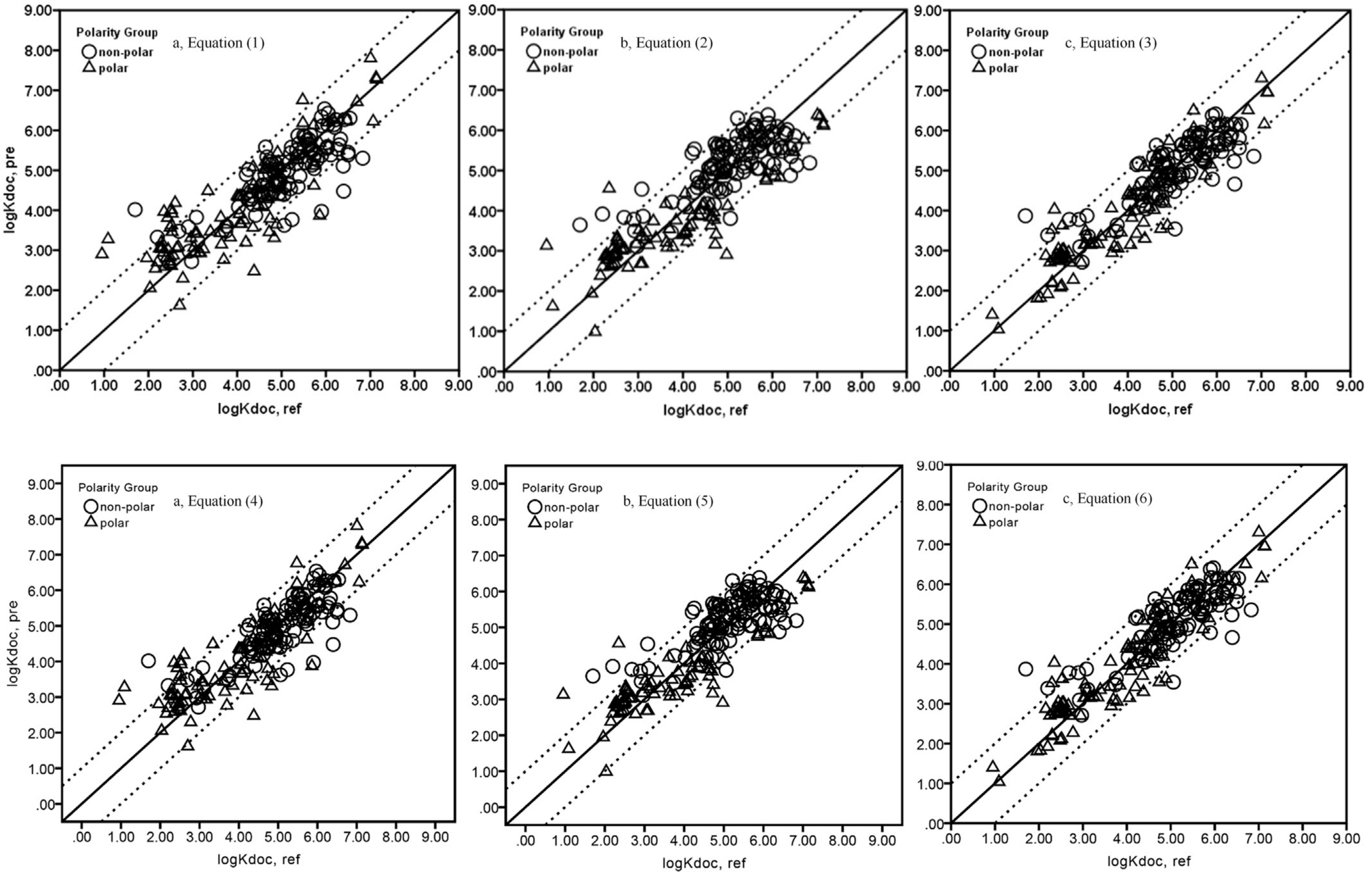


Fig. 1. The relationships between the experimental and modeled log K_{doc} values of Eqs. (1), (2), and (3).

Table 1
The PCF values of the polar functional groups in Eqs. (3), (6), and (7).

Polar functional groups	PCFs		
	Eq.(3)	Eq.(6)	Eq.(7)
—OH	−0.063	−0.221	−0.052
—NH2/—NH—/—N	−0.498	−0.435	−0.464
—N=/—CN	0.014	−0.177	0.025
—NCOO—	−1.227	−1.782	−1.172
—COOH	−2.409		−2.391
—COO—	−0.736	−0.279	−0.725
—CO—	−0.190	0.103	−0.181
—O—	−0.046	−0.112	−0.052
—S—	−0.022	0.236	−0.057
—F	−0.436	−0.283	−0.42
—Cl	0.125	0.154	0.129
—Br	0.345	0.142	0.353
—NO2	−0.076	−0.198	−0.015
—PO3	2.915		2.872
—I	−0.124	0.081	−0.102
μ			−0.02

Compared with Eq. (1), the coefficients of ${}^0\chi_P^V$, ${}^1\chi_P$ and ${}^6\chi_{CH}$ in Eq. (3) changed evidently with a decrease in their absolute values. These decrease is probably due to the fact that the selected polar functional groups also partly reflect information like molecular volume and substitution numbers on the rings, therefore the contributions of these MCI values on the K_{doc} values decreases when substitutional groups with larger volumes or commonly found on the rings like chlorine and bromine are introduced.

Considering the fitting effects, the PCF-modified model shows a higher fitting effect to the original one with the adj-R^2 rising from 0.725 to 0.815 and the SEE declining from 0.705 to 0.577. As shown in Fig. 1(c), the spots of predicted and experimental values are much closer to the Y = X line as only 25.7% and 7.4% of OPs having an absolute residual value over 0.50 and 1.0 log-unit, and some obvious outliers in the original model are clearly modified. Similarly, for the comparison of the K_{ow}-K_{doc} model, the MCI-PCF-K_{doc} model also performs much better. The evidence of these examples above demonstrates that the PCF modification is beneficial for the precision of the MCI-K_{doc} prediction. However, there are still some non-negligible outliers of the reference line in Fig. 1(c), which indicates that the precision and applicability of the MCI-PCF-K_{doc} model may still be limited to some extent.

$$\log_{10}K_{doc} = 0.211^1\chi_P + 0.122^3\chi_P + 2.91^6\chi_{CH} + 0.061^0\chi_P^V - 0.067^6\chi_{PC}^V + 1.202 + \sum_i n_i \cdot F_i; n = 202, adj-R^2 = 0.815, SEE = 0.578 \tag{3}$$

To figure out the application condition of the MCI-PCF-K_{doc} model, the 15 OPs, whose absolute residual values were over 1.0 log-unit, are listed in Table 2 as well as their detailed experimental conditions. For 14 out of 15 contaminates, the absolute residuals of the original MCI-K_{doc} model are also above or very close to 1.0 log-unit, which indicates that the PCFs are not very effective for these compounds. Therefore, the data sources of these K_{doc} values were examined, and then, the unchanged outliers were assumed to result from (1) the molecular structure and (2) the experimental conditions.

3.1.1. The molecular structure

The ineffectiveness of PCFs for OPs like 2,3,7,8-tetrachlorodibenzo-p-dioxin (2,3,7,8-TCDD), γ-gexachlorocyclohexane (γ-HCH, lindane), and chloranil, might be attributed to the molecular structure. Since 2,3,7,8-TCDD and chloranil are indeed nonpolar substances that are highly symmetrical, the extent of the binding influence of the chlorine and oxygen atoms in their molecules may differ from other OPs. Studies have already shown that there are obvious differences in some physicochemical properties (like melting point and vapor pressure) of HCH isomerides (Willett et al., 1998). Clearly, these differences come from different geometry configurations. However, sadly, neither MCIs nor PCFs are capable of describing the geometric configuration of the isomerides. Therefore, the PCF-modified model cannot further predict the K_{doc} values of lindane compared to the original model.

3.1.2. The experimental condition

It was proven that different experimental conditions resulted in various K_{doc} values of even the same OPs. The most concerning conditions mainly consist of experimental methods and DOM species. For instance, Krop and his co-workers compared the logK_{doc}-logK_{ow} relationships of PAHs by various experimental methods and they found that the K_{doc} values measured by FQ and the revised-phase method (RP) were slightly higher than the values by HPLC (Krop et al., 2001). Additionally, Kukkonen and Pellinen (1994) reported that the K_{doc} values by the Dialysis method were generally higher than the solvent extraction (LLE) method. However, these method comparison studies mainly concentrate on limited OP groups such as PAHs and PCBs, which makes simply adopting their results to other OPs with potential different binding performance inappropriate. Furthermore, so far only few methods have been evaluated and compared, and there lacks a comprehensive evaluation of experimental methods. As a result, the current study does not exclude any experimental method.

For the influences of the experimental DOMs, Krop et al.(2001) concluded that the K_{doc} values by commercial DOMs were typically higher

Table 2
Data information of the OPs with absolute residuals over 1.0 log-unit of Eqs. (1) and (3).

Compound	Class	Residuals		DOM	Method	Number of studies
		Eq. (1)	Eq. (3)			
Carbamazepine	Amides	−1.39	−1.10	Natural	Dialysis	Single
Metolachlor	Amides	−1.61	−1.68	Natural	Batch	Single
BDE-99	PBDEs	−1.29	−1.03	Both	Dialysis/SPME	Three
2,3,7,8-TCDD	PCDDs	1.53	1.47	Natural	AK	Single
γ-HCH	Others	−2.32	−2.17	Both	AS	Single
Chloranil	Others	1.44	1.52	Both	UVspectr	Single
Acenaphthylene	PAHs	−0.79	−1.09	Both	RP	Single
PCB-126	PCBs	1.28	1.16	Commercial	SPME	Single
PCB-116	PCBs	1.93	1.12	Commercial	Dialysis	Single
PCB-33	PCBs	1.92	1.74	Natural	CS	Single
1-Naphthol	Phenols	1.90	1.08	Natural	FQ	Two
p-tert-octylphenol	Phenols	1.24	1.26	Commercial	FQ	Single
1,2,3-TCB	X-PAHs	−1.12	−1.19	Both	AS	Single
Carbazole	Heterocycles	0.95	1.19	Commercial	HPLC	Single
Thiabendazole	Heterocycles	−0.91	−1.22	Both	SPE	Single

than natural DOMs for the same OPs. Furthermore, they also found that the binding mechanisms for the same OP to seawater DOM clearly differed from the freshwater DOM. As mentioned before, to minimize these K_{doc} value deviations of various researchers and experimental conditions, the median values were utilized in this study. However, for OPs with only one experimental K_{doc} study or K_{doc} value reported, the potential bias resulting from specific conditions cannot be controlled, which then may significantly influence the model development. As shown in Table 2, the K_{doc} values of almost all of the 15 OPs came from a single independent study, and both the DOMs and the methods utilized in these studies varied a lot and therefore the deviations of experimental conditions here are also not negligible. Because of the relatively limited studies, the reasons for the unchanged outliers are unlikely to be confirmed yet.

Despite the unchanged outliers, the MCI-PCF-K_{doc} model can still be considered a useful prediction tool for two reasons. First, the model provides relatively good regression parameters; and second, there is a large dataset containing various OP groups and experimental conditions, especially compared with the existing MCI-Kdoc models for specific OPs. However, in our current dataset, the DOMs utilized are from various sources including different natural water bodies (ponds, rivers, lakes, sea, etc.) and some commercial sources like Aldrich Humic Acid (HA), to further control the deviation of the experimental conditions and to raise the accuracy and applicability of the MCI-K_{doc} model for specific conditions, the most commonly used DOM in the literature, Aldrich HA, was then chosen for the development of a more specific MCI-K_{doc} model in the following section.

3.2. Development of MCI-K_{doc} models for Aldrich HA

Through searching the original dataset, 132 OPs with Aldrich HA as binding DOM were selected and these compounds belong to 10 groups including PAHs, PCBs, PBDEs, phenols, X-PAHs, PCDDs, OCPs, triazines and heterocyclic compounds. Similar to the overall models, the median values of 29 OPs, which have more than one K_{doc} value, were taken for the stepwise regression, and the logK_{doc} range of the Aldrich HA dataset is from 1.69 to 7.28. According to the stepwise regression, 7 indices (${}^{0}\chi_P$, ${}^{1}\chi_P$, ${}^{3}\chi_P$, ${}^{3}\chi_C$, ${}^{6}\chi_{CH}$, ${}^{2}\chi_P^V$ and ${}^{4}\chi_P^V$) were selected for the Aldrich HA model (adj-R^2 = 0.725 and SEE = 0.705). However, the multicollinearity of those MCI could affect the model efficiency. Therefore, those group indices were partly discarded in further model development. To utilize the structure information of the substances to the greatest extent, 5 indices with most relatively clear structural meanings for both molecular and local levels (${}^{1}\chi_P$, ${}^{0}\chi_P^V$, ${}^{3}\chi_C$, ${}^{4}\chi_{PC}$ and ${}^{6}\chi_{CH}$) were then chosen for a multiple linear regression instead (Eq. (4)). Additionally, the K_{ow}-K_{doc} model of Aldrich HA was also developed with 9-formylanthracene and anthraquinone excluded (Eq. (5)).

$$\log_{10}K_{doc} = -0.346^{1}\,\chi_P - 0.438^{3}\chi_C + 0.154^{4}\chi_{PC} + 11.282^{6}\chi_{CH} + 0.496^{0}\chi_P^V + 0.584; n = 132, adj{-}R^2 = 0.815, SEE = 0.617 \quad (4)$$

$$\log_{10}K_{doc} = 0.065\ \ \log_{10}K_{ow} + 1.935; \quad n = 130, adj{-}R^2 = 0.813, SEE = 0.625 \quad (5)$$

Similar to Eq. (1), Eq. (4) shows both general and local molecular properties of OPs contribute to their K_{doc}, and despite the high variation coefficient of each MCI, each term in Eq. (4) is also generally within the same order of magnitude.

For Eq. (4), the absolute residuals (data shown in SI) of 31.8% modeled OPs are over 0.50 log-units and the average residual is 0.46 log-unit. For Eq. (5), 42.4% of the residuals are over 0.50 log-unit with an average residual of 0.50 log-unit. Transversely evaluating the two equations with adj-R^2, SEE, and residual levels, the regression effect of the MCI-K_{doc} model for Aldrich HA is only slightly better than the corresponding K_{ow}-K_{doc} model. Moreover, when comparing the outliers with residuals over 1.0 log-unit, the respective 12 outliers of the two equations are almost completely different, which demonstrates that Eq. (4) may be less valid than Eq. (5) for some specific compounds. However, since MCI values are solid and easy to access whereas K_{ow} values always need to be determined, the uncertainty of Eq. (4) is much less. The vertical comparison of Eqs. (1) and (4) shows that the Aldrich MCI model is indeed more efficient than the overall model. Meanwhile, the observation of the modeled and experimental K_{doc} relationship (Fig. 1) shows that the distribution of the outliers is also within the entire range of Eq. (4), especially for the polar OP spots. As a result, PCFs were also introduced for the Aldrich HA MCI-K_{doc} model. Compared to Eq. (3), 13 factors were added to the MCI-K_{doc} model of Aldrich HA except for carboxyl (—COOH) and phosphate (—PO3) since no OPs contain these groups. The PCFs are listed in Table 1 and the MCI-PCF-K_{doc} model for Aldrich HA is as follows:

$$\log_{10}K_{doc} = 0.280^{1}\,\chi_P + 0.056^{3}\chi_C - 0.164^{4}\chi_{PC} + 4.887^{6}\chi_{CH} + 0.164^{0}\chi_P^V + 0.435 + \sum_i n_i \cdot F_i; n = 132, adj{-}R^2 = 0.907, SEE = 0.438 \quad (6)$$

When comparing the PCF-modified and non-modified models, the adj-R^2 increases from 0.815 to 0.907 and the SEE declines from 0.617 to 0.438. The residual analysis (shown in SI) shows that the number of OPs with residuals over 0.50 and 1.0 log-unit declines to 24 (18.2%) and 3 (2.3%), whereas the average residual declines to 0.31 log-unit. As also shown in Fig. 1, after the PCFs were added, the distribution of the spots is much more centered to the reference line, and thus, the precision of the model is clearly improved. The two MCI-PCF-K_{doc} models were also compared for the percentage of outliers, where the Aldrich model is evidently lower than the overall model. It indicates that the different DOM sources impact the overall model to some extent, and by restricting the DOM, the precision of the prediction model increases markedly. In our current study, the effect of various DOM is only controlled by choosing Aldrich HA as a representative since it is dominating in K_{doc} studies. Other DOM choices in the literature either have very limited K_{doc} studies or simply natural extracted which have little information on DOM properties, so none DOM parameters were further introduced to modified the overall MCI-PCF model. Future studies which interested in other specific DOMs could also consider the establishing a DOM specific MCI-PCF model when there are enough data, and ultimately establishing a MCI-DOM model.

Even though the majority of the outliers are modified in Fig. 1f, the model prediction effects for some OPs evidently decreases. Therefore, the detailed experimental conditions of the outliers in Eq. (6) were also conducted (Table 3) for further analysis on the model effects.

Regarding the aspect of experimental conditions, all three substances listed in Table 3 were only reported in single studies, and more importantly, the UV- spectrum method of the chloranil and the electrokinetic chromatograph (EK) method of ameline are not used in other studies. In this case, until a detailed comparison of all experimental methods based on various OP classes, or at least for these outlier compounds, the extent of method associated influence on our model cannot be confirmed and the deviations cannot be eliminated. Meanwhile, in the aspect of molecular structures, the three OPs are all planar molecules with certain symmetries and conjugated systems (Fig. S1 in SI). Therefore, the polar groups in their structure may contribute less or unequally to their binding abilities while the

Table 3
Data information of outliers in Eqs. (4) and (6).

Compounds	Groups	Residual of Eq. (4)	Residual of Eq. (6)	Method	Number of studies
Chloranil	Others	2.09	1.58	UVspectr	Single
Carbazole	Heterocycles	0.97	1.20	HPLC	Single
Ameline	Triazines	0.15	1.10	EK	Single

Table 4
The Jackknifed R^2 values of both patterns for the overall K_{doc} models.

Method	Model	R^2	Jackknifed R^2		
			Average	Range	CV
The overall K_{doc} models					
Pattern (a)	MCI-PCF	0.815	0.817	0.796–0.854	0.017
	K_{ow}	0.740	0.740	0.699–0.759	0.024
Pattern (b)	MCI-PCF	0.815	0.816	0.800–0.836	0.011
	K_{ow}	0.740	0.741	0.716–0.769	0.018
The Aldrich HA K_{doc} models					
Pattern (a)	MCI-PCF	0.907	0.909	0.896–0.920	0.008
	K_{ow}	0.813	0.812	0.788–0.829	0.020
Pattern (b)	MCI-PCF	0.907	0.908	0.888–0.936	0.012
	K_{ow}	0.813	0.812	0.800–0.837	0.012

molecular structures themselves are more important. Especially for ameline, the nitrogen dominates in its molecule, and thus the excessive PCFs (up to six PCFs) for ameline is probably responsible for the increase of the residuals. Careful attention should be paid for these compounds when applying Eq. (6).

Conclusively, the PCF-modified MCI-K_{doc} model performed well in the K_{doc} values regression for the Aldrich HA with multiple OP groups. Compared with the commonly used K_{ow}-K_{doc} model, the accuracy, the variable availability, and the applicability of the MCI-PCF-K_{doc} model are all better. Additionally, compared with the overall MCI-PCF-K_{doc} model, the specificity of the Aldrich model is much stronger. In this case, for DOMs that are more similar to Aldrich HA in properties and characteristics, the Aldrich HA model could a better choice in predicting the K_{doc} values for new compounds.

3.3. Modification attempt with the dipole moment

According to the residual discussion for both overall and Aldrich HA models, one potential concern of the PCF-modified MCI-K_{doc} models was that PCFs are not sufficient to describe the influences of the overall molecular polarity on the binding effects. To minimize this limitation as much as possible, dipole moment (μ), the easiest and most common parameter representing molecular polarity, was supplemented into the MCI-K_{doc} models as an attempt. Two scenarios were set for the attempt: (a) using the dipole moment as a supplement for PCFs; and (b) using the dipole moment as a substitution of PCFs. The μ-modified attempt was made for both overall and Aldrich K_{doc} models, and the theoretical μ values of all 202 OPs were calculated by their molecular structure with the Gaussian 09, (2009) using B3LYP/6-31G(d).

For scenario (a), when comparing the pre-μ-modified and post-μ-modified (Eq. (7)) overall MCI-PCF-K_{doc} models, the adj-R^2 and SEE values are nearly unchanged. In addition, the differences between the pre and post modified regression coefficients are all within 0.06 (shown in SI), and the relative variations of majority of the coefficients are below 20% except for the coefficients with relatively low values, such as F_{NH2}, F_S, and F_{NO2}. Analysis of the residuals (shown in SI) for each compound indicates that the largest difference of residuals between the two models is only 0.08 log-unit, which also demonstrates that the dipole moment modification is indeed unnecessary for the overall MCI-PCF-K_{doc} model. In fact, the numbers of residuals over 0.50 and 1.0 log-unit are increased after the dipole moment modification. For the supplement of the Aldrich HA model, the differences of the coefficients and the residuals are even smaller with all the outliers unchanged (coefficients and the residuals shown in SI).

$$\log_{10}K_{doc} = 0.228\,{}^{1}\chi_P + 0.110\,{}^{3}\chi_P + 2.888\,{}^{6}\chi_{CH} + 0.054\,{}^{0}\chi_P^V - 0.064\,{}^{6}\chi_{PC}^V - 0.02\mu + 1.226 + \sum_i n_i \cdot F_i; n = 202, adj-R^2 = 0.814, SEE = 0.579 \tag{7}$$

For scenario (b), there were two substitution solutions. The first one was to replace all of the PCFs with μ in the form of Eqs. (3) and (6), and the second one was to reselect all input variables of the K_{doc} model from all of the MCIs and the dipole moment by a stepwise multiple linear regression. For the Aldrich HA models, the general regression effects (indicated by adj-R^2 and SEE) of both solutions were higher than the original MCI-K_{doc} model but still lower than the MCI-PCF-K_{doc} model (shown in SI). Meanwhile, the residual results show that the outlier numbers of the two substitution models are only slightly less than the original model, and some non/low-polar OPs like chloranil and carbazole are still listed as outliers. Furthermore, the residual value of ameline increases remarkably from 0.15 log-unit in Eq. (4) to 1.19 log-unit for both solutions (data shown in SI). The residual increase is probably because ameline has the highest dipole moment value (10.68 D) of all of the OPs, which indicates that the μ-substitution attempts are also no better than the PCF-modified models for some strongly polar OPs. For these kinds of OPs, the polar properties may not be a main factor of binding abilities compared to its molecular structural properties. Similar conclusions can also be found for the overall models (coefficients and the residuals shown in SI).

As discussed above, neither the combination of the dipole moment with the PCFs nor the dipole moment replacement of the PCFs successfully improved the accuracy or the application of the MCI-PCF-K_{doc} models. It is probably because that PCFs themselves have already contain the similar information of molecular polarity that dipole moment does while the dipole moment is lack of the spatial information of the functional groups that PCFs contain. Also, since dipole moment value needs additional calculation which may be inconvenient for some potential users, hence Eqs. (3) and (6) were not revised. However, further studies on different parameters of the overall molecular polarity are still suggested for the improvement of the models.

3.4. Robustness and the sensitivity analysis

The robustness and the sensitivity of the models are important indices of model effects. The robustness of a model represents the ability of the estimation features of a model to remain unchanged when there are minor changes of the modeling conditions (Willett et al., 1998), and the sensitivity of a model evaluates the contribution of the different variables to the model estimation.

As mentioned above, a modified Jackknifed method was utilized to calculate the robustness of the models (Cornish-Bowden and Wong, 1978; Dietrich et al., 1980). This method examines the influences of the modeling data on model robustness by rebuilding the model serval

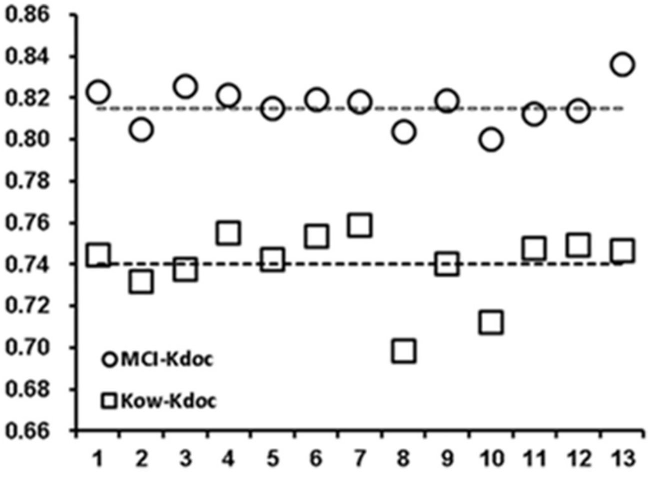


Fig. 2. Jackknifed adj-R^2 values for pattern (b) of the overall models 1-Amides, 2-PBDEs, 3-PCDDs, 4-OCPs, 5-PAHs, 6-PCBs, 7-Hormones, 8-Phenols, 9-Pyrethroids, 10-Triazines, 11-X-PAHs, 12-Heterocycles, 13-Others.

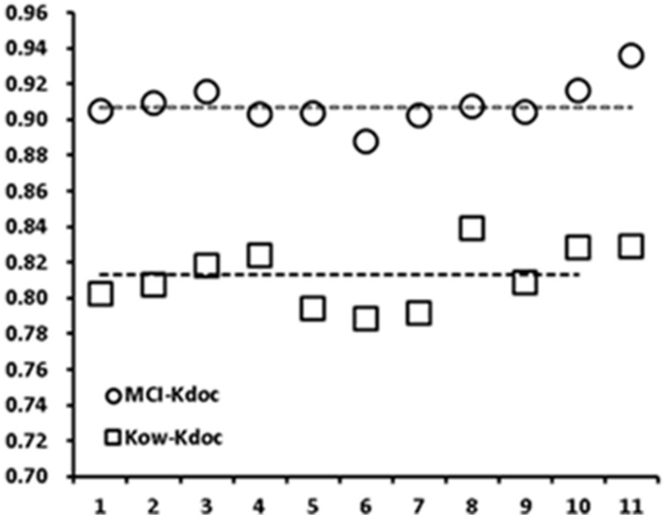


Fig. 3. Jackknifed adj-R^2 values for pattern (b) of the Aldrich HA models 1-PBDEs, 2-PCDDs, 3-OCPs, 4-PAHs, 5-PCBs, 6-Phenols, 7-Triazines, 8-Hormones, 9-X-PAHs, 10-Heterocycles, 11-Others.

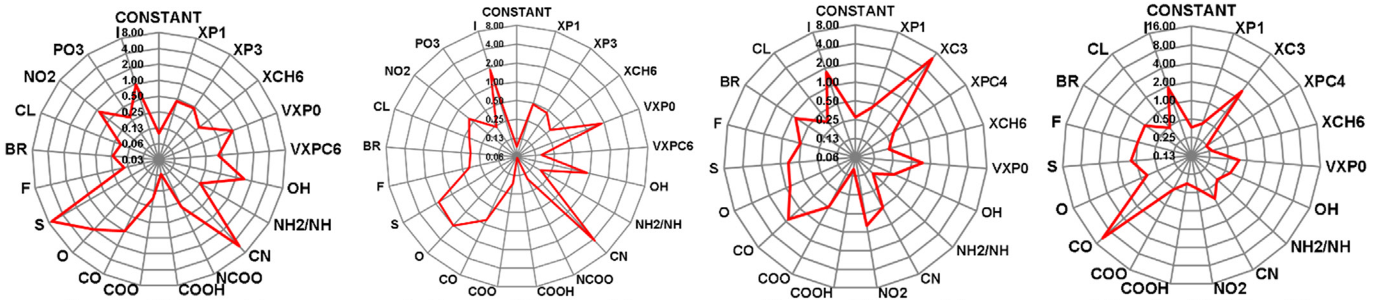
times with different groups of data eliminated from original database, and then comparing the deviations of fitting parameters among the pre- and post-elimination models (Nirmalakhandan and Speece, 1989; Tao et al., 2000). It has been widely used in several prediction models on the properties of OPs (like BCF, Koc, s and EC50) (Nirmalakhandan and Speece, 1988; Tao et al., 1999; Tao et al., 2002). Based on these studies, two different elimination patterns were designed to evaluate the changes of the adj-R^2, MCI coefficients and the PCFs and hence to judge the robustness of the two MCI-PCF-K_{doc} models (Eqs. (3) and (6)) (Lu et al., 1999a; Lu et al., 2000b; Lu et al., 2000c; Tao et al., 2000). The robustness of their corresponding K_{ow}-K_{doc} models were also tested for comparison. The two elimination patterns are as follows: (a) elimination by random groups: 15% of the modeling dataset (30 OPs for the overall models and 20 OPs for the Aldrich models) were eliminated each time and each OP was eliminated at least once and five times at most; (b) elimination by classes: the modeling dataset was classified by polar groups (as stated above) and each time a polar group was eliminated.

For the overall MCI-PCF-K_{doc} and the K_{ow}-K_{doc} models, the Jackknifed adj-R^2 values of both patterns are listed in Table 4. For both patterns, the average Jackknifed adj-R^2 values of both models are very close to their corresponding original adj-R^2 values, which indicates that both overall models are robust in general. However, the MCI-PCF model is more robust than the K_{ow} model since its variable coefficients (CV) of adj-R^2 are lower. Despite the close average of the adj-R^2 values, the deviations of the extreme values for both models are larger, which suggests the relatively strong influences of certain compounds on the model robustness. To identify the influences of different molecular structures on the model robustness, the Jackknifed adj-R^2 values of pattern (b) are plotted in Fig. 2. After the eliminations of the phenol and the triazine classes, the adj-R^2 values evidently drop from the original adj-R^2 (as the reference lines) for both models, and thus, these two OP groups are considered to play more central roles in the models. For the MCI-PCF-K_{doc} model, the "others" group shows a big impact of the model accuracy as the adj-R^2 value increases significantly after it is eliminated. Since only the OPs with unique structures that cannot be classified into an existing group with more than three congeners were classified as "others", the differences of the input variables within this group are indeed much larger. As a result, the general model accuracy is lowered by these OPs.

Similarly, Table 4 and Fig. 3 demonstrate the changes of the Jackknifed adj-R^2 values of the Aldrich HA models. On one hand, both models are generally very robust since the CV values are all below 0.02, and more specifically, the MCI-PCF model is slightly steadier than the K_{ow} model. On the other hand, both models are affected by some OP classes. For the MCI-PCF model, the phenols also play a more central part in construction of the model as they do for the overall model. The reason might come from two areas: (1) the majority of the K_{doc} values of phenols are adopted from single study, therefore the consistency of the data is more guaranteed and less deviation in modeling can be assumed; (2) the structures of the phenols are quite simple and nearly contains only one polar group (carbonyl), and hence, the MCIs together with the PCFs can be more descriptive about their spatial structures and polarities compared to other OPs. Additionally, similar to the overall MCI-PCF model, the "others" group also limits the model accuracy of the Aldrich HA model. Actually, the number one outlier, chloranil, is indeed from this group. Conversely, and slightly different from the overall K_{ow} model, hormones, phenols and triazines all play an important role in the Aldrich HA K_{ow}-K_{doc} model.

The robustness of all of the coefficients in Eqs. (3) and (6) were also checked for the two elimination patterns, and the CV values of the coefficients are shown in Fig. 4. For the overall model, the rose plots for both patterns are generally alike. The majority of the CV values are below 1.0, which shows relatively good robustness of these parameters. However, the CV values of PCFs for some polar groups (S, O, and CN) are quite high, which indicates that their robustness is comparatively weak in the model. For the Aldrich HA model, the rose plots are more different even though the majority of the parameters are still quite stable. For pattern (a), the coefficient of $^3\chi_C$ has the highest CV value, which may come from the relatively small values of the coefficient itself. However, for pattern (b), the CV value of F_{CO} becomes much higher than the value of $^3\chi_C$. This change comes from the elimination of the "others" group since carbonyl group is tightly associated with these OPs, which indicates the strong influence of the "others" group on the model.

For the sensitivity analysis of the models, the degrees of modeled logK_{doc} changes were evaluated when every input variable changes a determined extent each time. Since the MCI-PCF-K_{doc} models are all linear models, the input variables are set to be changed for only one lever—10%. The sensitivity differences (SD) of each variable are defined as the



a, Equation (3)-Pattern (a) b, Equation (3)-Pattern (b) c, Equation (6)-Pattern (a) d, Equation (6)-Pattern (b)

Fig. 4. CV values of the coefficients in Eqs. (3) and (6) for the two patterns.

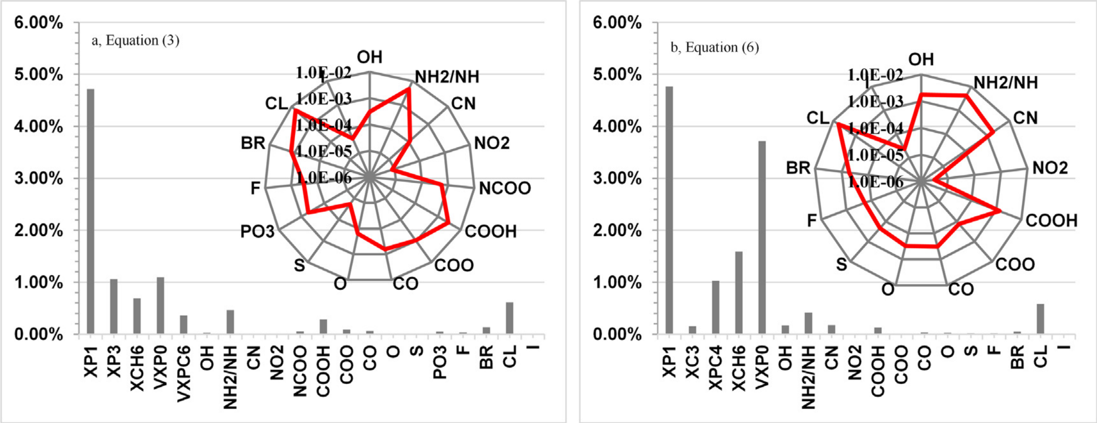


Fig. 5. Parameter sensitivities of Eqs. (3) and (6).

average values of $|\Delta \log K_{doc}|/ \log K_{doc}$(original) for the evaluation. As shown in Fig. 5, generally, the two models are much more sensitive to the MCIs than the PCFs. The ${}^{1}\chi_{P}$ is the most sensitive variable for both models with the ultimate logK_{doc} values changing slightly over 4.7% (4.72% for the overall model and 4.76% for the Aldrich HA) when ${}^{1}\chi_{P}$ is changed by 10%. For the overall model, no other variables are as sensitive as ${}^{1}\chi_{P}$, whereas the Aldrich HA model is also very sensitive to ${}^{0}\chi_{P}^{V}$. In terms of the PCFs, the shared (relatively) sensitive factors (as the SD value over 0.5%) for both models are the F_{Cl} and F_{NH2} as chlorine and amino are among the largest polar groups of the dataset. In addition, for the Aldrich HA model and since it is not very sensitive to F_{CO} and ${}^{3}\chi_{C}$, the potential impact of these two low-robust variables are actually limited.

4. Conclusions

This study constructed K_{doc} prediction models for two different DOM circumstances, nonspecific DOM and Aldrich HA, with MCIs and PCF based on K_{doc} values of various OP groups for the first time. Compared to the commonly constructed K_{ow}-K_{doc} models and the non-PCF modified MCI-K_{doc} models, both models showed higher regression effects for the overall and Aldrich HA DOM scenarios, respectively, which indicates relatively good prediction accuracies. By covering various OP groups in our dataset, the applicability of our models is greatly broadened compared to the existing models that focusing limited OP groups. Furthermore, since our input parameters are calculated solid numbers only based on molecular structures, they are much easier to access compared to models based on measured parameters when it comes to the implication of less studied OPs. As expected, the Aldrich HA model is more selective and pertinent than the overall model for Aldrich HA-like DOMs. However, when dealing with other DOMs that are either less studied or have large characteristic disparities, the overall model will come in handy. Generally, the robustness and sensitivity tests showed that the two models are quite robust and they rely less on the relatively non-robust parameters. In this case, the models can be considered effective tools in the K_{doc} value prediction. However, the effectiveness of the two models on some OPs are quite limited, this is probably due to the discrimination of experimental conditions and molecular structure. The dipole moment was added to the models as an attempt to minimize the deviations, but it was not effective enough either to supplement or to replace the PCFs. However, the addition of other parameters representing overall molecular polarity and DOM characteristics (when sufficient experimental data is available) is suggested for the improvement of the models in future studies.

Acknowledgments

The funding for this study was provided by the National Natural Science Foundation of China (NSFC) (41271462 and 41503083). This work is also supported by a grant from the 111 Project (B14001) and from the Undergraduate Student Research Training Program of the Ministry of Education. A part of calculation in this study is supported by High-performance Computing Platform of Peking University. F.L. Xu would also like to thank the China Scholarship Council for supporting his study at the University of Massachusetts, Amherst.

Appendix A. Supplementary data

Supplementary data to this article (including the logK_{doc} values, MCI values, and model residuals) can be found online at https://doi.org/10.1016/j.scitotenv.2019.03.146.

References

Agnola, G.D., Ferrari, G., Nardi, S., 1981. Antidote action of humic substances on atrazine inhibition of sulfate uptake in barley roots. Pestic. Biochem. Physiol. 15, 101–104.

Backhus, D.A., Gschwend, P.M., 1990. Fluorescent polycyclic aromatic-hydrocarbons as probes for studying the impact of colloids on pollutant transport in groundwater. Environmental Science & Technology 24, 1214–1223.

Boehm, P.D., Quinn, J.G., 1973. Solubilization of hydrocarbons by the dissolved organic matter in sea water. Geochim. Cosmochim. Acta 37, 2459–2477.

Burns, S.E., Hassett, J.P., Rossi, M.V., 1996. Binding effects on humic-mediated photoreaction: intrahumic dechlorination of mirex in water. Environmental science & technology 30, 2934–2941.

Carter, C.W., Suffet, I.H., 1982. Binding of DDT to dissolved humic materials. Environmental science & technology 16, 735–740.

Carter, C.W., Suffet, I., 1983. Interactions between Dissolved Humic and Fulvic Acids and Pollutants in Aquatic Environments. ACS Symposium series American Chemical Society. .

Chiou, C.T., Malcolm, R.L., Brinton, T.I., Kile, D.E., 1986. Water solubility enhancement of some organic pollutants and pesticides by dissolved humic and fulvic acids. Environmental science & technology 20, 502–508.

Chiou, C.T., Kile, D.E., Brinton, T.I., Malcolm, R.L., Leenheer, J.A., MacCarthy, P., 1987. A comparison of water solubility enhancements of organic solutes by aquatic humic materials and commercial humic acids. Environmental science & technology 21, 1231–1234.

Cho, H.H., Park, J.W., Liu, C.C.K., 2002. Effect of molecular structures on the solubility enhancement of hydrophobic organic compounds by environmental amphiphiles. Environ. Toxicol. Chem. 21, 999–1003.

Cornish-Bowden, A., Wong, J., 1978. Evaluation of rate constants for enzyme-catalysed reactions by the jackknife technique. Application to liver alcohol dehydrogenase. Biochem. J. 175, 969–976.

Delgado-Moreno, L., Wu, L., Gan, J., 2010. Effect of dissolved organic carbon on sorption of pyrethroids to sediments. Environmental science & technology 44, 8473–8478.

Dietrich, S.W., Dreyer, N.D., Hansch, C., Bentley, D.L., 1980. Confidence interval estimators for parameters associated with quantitative structure-activity relationships. J. Med. Chem. 23, 1201–1205.

Evers E, Velzen Mv, Oele M, Govers H. Estimating binding coefficients of chlorinated aromatics and aquatic humic substances from molecular properties. Organic Micropollutants in the Aquatic Environment. Springer, 1991, pp. 437–443.

Gaussian 09, Revision **E.01**, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc., Wallingford CT, 2009.

Haitzer, M., Hoss, S., Traunspurger, W., Steinberg, C., 1998. Effects of dissolved organic matter (DOM) on the bioconcentration of organic chemicals in aquatic organisms - a review. Chemosphere 37, 1335–1362.

Hassett, J.P., Anderson, M.A., 1982. Effects of dissolved organic matter on adsorption of hydrophobic organic compounds by river-and sewage-borne particles. Water Res. 16, 681–686.

Johnson, W.P., Amy, G.L., 1995. Facilitated transport and enhanced desorption of polycyclic aromatic-hydrocarbons by natural organic-matter in aquifer sediments. Environmental Science & Technology 29, 807–817.

Johnson, W.P., John, W.W., 1999. PCE solubilization and mobilization by commercial humic acid. J. Contam. Hydrol. 35, 343–362.

Jorgensen SE, Sorensen BH, Mahler H. Handbook of Estimation Methods in Ecotoxicology and Environmental Chemistry. Vol 2: CRC Press, 1997.

Khan, S., 1973. Equilibrium and kinetic studies of the adsorption of 2, 4-D and picloram on humic acid. Can. J. Soil Sci. 53, 429–434.

Kier, L.B., Hall, L.H., 1986. Molecular Connectivity in Structure-Activity Analysis.

Krop HB, van Noort PCM, Govers HAJ. Determination and theoretical aspects of the equilibrium between dissolved organic matter and hydrophobic organic micropollutants in water (K-doc). Reviews of Environmental Contamination and Toxicology, Vol 169 2001; 169: 1–122.

Kukkonen, J., Oikari, A., 1991. Bioavailability of organic pollutants in boreal waters with varying levels of dissolved organic material. Water Res. 25, 455–463.

Kukkonen, J., Pellinen, J., 1994. Binding of organic xenobiotics to dissolved organic macromolecules: comparison of analytical methods. Sci. Total Environ. 152, 19–29.

Lafrance, P., Villeneuve, J., Mazet, M., Ayele, J., Fabre, B., 1991. Organic compounds adsorption onto activated carbon: the effect of association between dissolved humic substances and pesticides. Environ. Pollut. 72, 331–344.

Landrum, P.F., Nihart, S.R., Eadie, B.J., Herche, L.R., 1987. Reduction in bioavailability of organic contaminants to the amphipod Pontoporeia hoyi by dissolved organic matter of sediment interstitial waters. Environ. Toxicol. Chem. 6, 11–20.

Laor, Y., Rebhun, M., 1997. Complexation-flocculation: a new method to determine binding coefficients of organic contaminants to dissolved humic substances. Environmental Science & Technology 31, 3558–3564.

Lee, J., Cho, J., Kim, S.H., Kim, S.D., 2011. Influence of 17β-estradiol binding by dissolved organic matter isolated from wastewater effluent on estrogenic activity. Ecotoxicol. Environ. Saf. 74, 1280–1287.

Li, X.-H., Zhu, Y.-G., Yu, Q.-S., 2000. A novel index of vertex for hetero—atoms in molecules (in chinses). Acta Chim. Sin. 58.

Lu, X., Tao, S., Cao, J., Dawson, R., 1999a. Prediction of fish bioconcentration factors of nonpolar organic pollutants based on molecular connectivity indices. Chemosphere 39, 987–999.

Lu, X., Tao, S., Li, H., 1999b. Estimation of adsorption coefficient values of nonpolar organic compounds based on molecular connectivity indices (in Chinese). Acta Pedol. Sin. 3.

Lu, X., Tao, S., Li, H., 1999c. Prediction of KOC of polar organic compounds based on molecular connec tivity indices (in Chinese). Acta Sci. Circumst. 19, 277–283.

Lu, X., Tao, S., Hu, H., 2000a. Comparison of estimation models for sorption coefficients (Koc) of organic compounds (in Chinese). Chinese Journal of Soil Science 31, 166–170.

Lu, X., Tao, S., Hu, H., 2000b. Prediction of bioconcentration factors of organic compounds in fish by molecular connectivity indices and function correction factors (in Chinese). Chin. J. Appl. Ecol. 11, 277–288.

Lu, X., Tao, S., Hu, H., Dawson, R., 2000c. Estimation of bioconcentration factors of nonionic organic compounds in fish by molecular connectivity indices and polarity correction factors. Chemosphere 41, 1675–1688.

Mackay, D., Shiu, W.-Y., Ma, K.-C., 2010. Lee SC. CRC press, Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals.

Maoz, A., Chefetz, B., 2010. Sorption of the pharmaceuticals carbamazepine and naproxen to dissolved organic matter: role of structural fractions. Water Res. 44, 981–989.

Neale, P.A., Escher, B.I., Goss, K.U., Endo, S., 2012. Evaluating dissolved organic carbon-water partitioning using polyparameter linear free energy relationships: implications for the fate of disinfection by-products. Water Res. 46, 3637–3645.

Nirmalakhandan, N.N., Speece, R.E., 1988. Prediction of aqueous solubility of organic chemicals based on molecular structure. Environmental science & technology 22, 328–338.

Nirmalakhandan, N.N., Speece, R.E., 1989. Prediction of aqueous solubility of organic chemicals based on molecular structure. 2. Application to PNAs, PCBs, PCDDs, etc. Environmental science & technology 23, 708–713.

Nuerla, A., Qiao, X., Li, J., Zhao, D., Yang, X., Xie, Q., et al., 2013. Effects of substituent position on the interactions between PBDEs/PCBs and DOM. Chin. Sci. Bull. 58, 884–889.

Pavan M, Worth AP, Netzeva TI. Review of QSAR models for bioconcentration. European Commission Joint Research centre (EUR 22327 EN) 2006.

Rav-Acha, C., Rebhun, M., 1992. Binding of organic solutes to dissolved humic substances and its effects on adsorption and transport in the aquatic environment. Water Res. 26, 1645–1654.

Sabljic, A., 1987. On the prediction of soil sorption coefficients of organic pollutants from molecular structure: application of molecular topology model. Environmental science & technology 21, 358–366.

Sabljić, A., 1991. Chemical topology and ecotoxicology. Sci. Total Environ. 109, 197–220.

Sabljić, A., Protić, M., 1982. Molecular connectivity: a novel method for prediction of bioconcentration factor of hazardous chemicals. Chem. Biol. Interact. 42, 301–310.

Sekušak, S., Sabljić, A., 1992. Soil sorption and chemical topology. J. Math. Chem. 11, 271–280.

Tao S, Lu X. Estimation of organic carbon normalized sorption coefficient (Koc) for soils by topological indices and polarity factors Chemosphere 1999; 39: 2019-2034.

Tao, S., Piao, H., Dawson, R., Lu, X., Hu, H., 1999. Estimation of organic carbon normalized sorption coefficient (Koc) for soils using the fragment constant method. Environmental science & technology 33, 2719–2725.

Tao S, Lu XX, Cao J, Hu HY. Robustness test of a topological indices and polarity factors model for estimating KOC of organic compounds. Acta Pedologica Sinica 2000.

Tao, S., Lu, X.X., Cao, J., Dawson, R., 2001. A comparison of the fragment constant and molecular connectivity indices models for normalized sorption coefficient estimation. Water Environment Research 73, 307–313.

Tao, S., Xi, X., Xu, F., Li, B., Cao, J., Dawson, R., 2002. A fragment constant QSAR model for evaluating the EC 50 values of organic chemicals to daphnia magna. Environ. Pollut. 116, 57–64.

Traina, S.J., Spontak, D.A., Logan, T.J., 1989. Effects of cations on complexation of naphthalene by water-soluble organic carbon. J. Environ. Qual. 18, 221–227.

Willett, K.L., Ulrich, E.M., Hites, R.A., 1998. Differential toxicity and environmental fates of hexachlorocyclohexane isomers. Environmental Science & Technology 32, 2197–2207.

Yu PC, Weber WJ, Eadie BJ. Estimating the effects of dispersed organic polymers on the sorption of contaminants by natural solids. 2. Sorption in the presence of humic and other natural macromolecules. Environ. sci. technol 1990; 24: 837–842.