Prediction of octanol-water partition coefficients for alkyl-PAHs based on the solvation free energy

W R Li¹, G B Song¹, G H Ding¹,³ and H Gao²

¹College of Environmental Science and Engineering, Dalian Maritime University, Dalian, China
²Department of Physics, Dalian Maritime University, Dalian, China
³E-mail: guanghuiding@dlmu.edu.cn

Abstract. Octanol-water partition coefficients ($K_{OW}$) are of great importance in assessing the environmental fate and risk of hydrophobic organic pollutants including alkyl-poly cyclic aromatic hydrocarbons (alkyl-PAHs). Due to the time-consuming and laborious of experimental determination of $K_{OW}$, it is desirable to develop a rapid and accurate prediction model for the $K_{OW}$ of alkyl-PAHs. In this study, a prediction model for log $K_{OW}$ of alkyl-PAHs was developed according to the thermodynamic relationship between $K_{OW}$ and the solvation free energy from $n$-octanol phase to aqueous phase ($\Delta G_{OW}$). $\Delta G_{OW}$ values of alkyl-PAHs were calculated by the SMD model. The effect of dimer formation on the partition behaviour and partition coefficient was explored. After taking the effect of dimer formation into account, the goodness-of-fit, predictive ability, and robustness of the prediction model was significantly improved. The developed model in this study had the superiorities of universal application domain, low dependence on the experimental values and clear model mechanism.

1. Introduction
Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous persistent organic pollutant composed of carbon and hydrogen atoms that contain at least two condensed or fused aromatic ring structures. They are mainly derived from incomplete combustion of fossil fuels and have been ubiquitously detected at notable concentrations in different environmental compartments. A total of 16 PAHs have been listed as toxic pollutants by US EPA [1]. With the increase concern of the potential toxicity of PAHs, alkyl-PAHs are emerging as a class of structural analogues of PAHs that are more abundant and persistent than their parent counterparts. In addition, alkylated PAHs is associated with an increased incidence of various cancers in humans and are potential threat to human health [2]. Despite the increasing concerns on these pollutants, their basic physicochemical data are still scarce.

Among all the physicochemical parameters, the octanol-water partition coefficient ($K_{OW}$) is a key parameter describing the environmental behavior of a chemical between aqueous phase and organic phase. It approximates the hydrophobicity of a chemical, which is indirectly related to its toxicity and biological activity [3]. For the purpose of environmental risk assessment, it is of great importance to obtain the $K_{OW}$ values systematically. Although a large amount of $K_{OW}$ values have been experimentally determined for PAHs, they are just a small fraction of the PAHs identified in the environment [4,5]. Furthermore, few studies have measured $K_{OW}$ values of alkyl-PAHs. Due to the lack of standard samples and suitable analytical methods for the determination of $K_{OW}$, it is impossible to experimentally
determine the \( K_{OW} \) values for all PAHs and alkyl-PAHs. Therefore, it is necessary to establish a fast and accurate predictive model to estimate \( K_{OW} \) values of alkyl-PAHs.

\( K_{OW} \) values of alkyl-PAHs can be predicted by numerous models, such as quantitative structure-property relationship (QSPR) [6,7], and topological features model [8]. According to the thermodynamic relationship, \( K_{OW} \) can be obtained by using the solvation free energy of a compound between aqueous phase and \( n \)-octanol phase (\( \Delta G_{OW} \)). Compared with QSPR model and topological features model, the \( \Delta G_{OW} \) model was almost independent on the experimental data, was not constrained by the application domain, and it can identify isomers effectively. Therefore, the \( \Delta G_{OW} \) model was superior to the two models mentioned above and was selected to predict \( K_{OW} \) in this study. Since there is no experimental \( K_{OW} \) available for alkyl-PAHs, the \( \Delta G_{OW} \) model was developed with \( K_{OW} \) values of parent PAHs.

The reliability of the \( \Delta G_{OW} \) model depends on the accuracy of the calculated solvation free energy. A large amount of quantum chemical methods for calculating \( \Delta G_{OW} \) have been reported, such as: polarized continuum model [9], the conductor-like screening model [10,11], the SMx series models [12,13]. In addition, a solvation model based on electronic density (SMD) protocol [14] has been developed and widely used in the prediction of partition coefficients [15,16]. Due to the consideration of first solvation shell, SMD model was generally accepted as the most reliable solvation model. Therefore, the SMD model was used to predict the \( \Delta G_{OW} \) value of alkyl-PAHs.

Furthermore, driven by the \( \pi-\pi \) interaction, the hydrophobic organic pollutants with aromatic \( \pi \) system can generate dimeric and polymeric configurations [17,18]. Therefore, in the nature environment, the hydrophobic organic pollutants may exist in the configuration of dimeric and polymeric pollutants, not only single molecules [19]. However, during the experimental determination of \( K_{OW} \) values, the concentrations of chemicals are high, the \( \pi-\pi \) interaction is enhanced and the probability of forming dimers or polymers increases. Thus, the experimental \( K_{OW} \) is the value of PAHs monomer, dimer, and polymer mixture, instead of the value of individual molecule of PAHs. As for the higher hydrophobicity of dimers and polymers, the experimental \( K_{OW} \) will higher than the predicted \( K_{OW} \) just based on PAH monomers. Therefore, it is of great significance to consider the influence of dimers and polymers on the prediction of \( K_{OW} \) values. Due to the low concentrations of PAHs in the natural environment, PAHs are found in a small fraction of polymers and mostly in the configuration of single molecules and dimers [20]. Thus, it is desirable to consider monomers and dimers of PAHs, instead of polymers.

In this study, a \( K_{OW} \) prediction model for alkyl-PAHs was developed with \( \Delta G_{OW} \) calculated by the SMD model. Furthermore, the influence of dimers on the prediction of \( K_{OW} \) was investigated, and was used to calibrate the prediction.

### 2. Materials and Methods

#### 2.1. Experimental log \( K_{OW} \) values

The experimental log \( K_{OW} \) values in the present study were taken from the handbook by Mackay [21]. The range of the variation for experimental log \( K_{OW} \) was 3.37 to 6.75. This data set consisted of 21 PAHs and all the experimental log \( K_{OW} \) were listed in table 1.

#### 2.2. Theory of the \( \Delta G_{OW} \) model

According to the thermodynamic relationship, the \( K_{OW} \) of hydrophobic organic pollutants can be obtained from the \( \Delta G_{OW} \) with equation 1.

\[
\log K_{OW} = -\frac{\Delta G_{OW}}{2.303RT}
\]

where \( R \) (\( 8.314 \text{ J} \cdot \text{mol}^{-1} \cdot \text{K}^{-1} \)) is the gas constant, \( T \) is the absolute temperature (K). The \( \Delta G_{OW} \) defined in the SMD model is consisted of three parts as equation 2.

\[
\Delta G_{OW} = \Delta G_{ENP} + G_{CDS} + \Delta G_{conc}
\]

where \( \Delta G_{ENP} \) is the bulk electrostatic contribution, arising from a self-consistent reaction field treatment; \( G_{CDS} \) is the cavity-dispersion-solvent-structure contribution, arising from short-range interactions
between the solute and solvent molecules in the first solvation shell; and the last term accounts for the difference in concentrations in gas phase standard state and solution-phase one [14].

ChemDraw and Chem3D (Version 14.0) were used to generate the initial molecular structure of PAHs. The calculation of solvation free energy by SMD model consisted of three procedures: geometrical optimization and calculation of frequency and energy, all of them were performed by Gaussian 09-E01 [22]. The frequency calculation was carried out after the geometry optimization to verify the optimized structure was the true minimum energy point. The single-point energy in aqueous phase and n-octanol phase was calculated with the SMD model at HF/MIDI6D level. In our previous study [23], HF/MIDI6D was the optimal level for predicting the log \( K_{OA} \) values of PCBs.

2.3. Estimate the ratio of monomer and dimer PAHs
The experimental log \( K_{OW} \) value was supposed to be the sum of log \( K_{OW} \) values of monomer and dimer PAHs. Consequently, QSAR model was employed to estimate the dimeric ratios of PAHs without experimental log \( K_{OW} \) values.

For the QSAR modelling, a total of 11 molecular structural descriptors representing the overall properties of the molecules were selected in this study. All the molecular structural descriptors were computed with the PM7 method [24] in MOPAC (Version 2016). Before modelling, independent variables should be pre-selected with the parameter of Variable Importance in the Project (VIP) [25]. Simca-S (Version 13.0, Umetri AB & Erisoft AB) software was used to perform the partial least square (PLS) regression for the QSAR model. The performance of PLS regression was mainly measured by the cumulative variances of independent and dependent variables \( (R^2_{Cum})\) and \( (R^2_{YCum})\), which are explained by the extracted components, and cumulative cross-validation coefficient \( (Q^2_{Cum})\). The model is robust and stable when the value of \( Q^2_{Cum} \) exceeds 0.5.

2.4. Statistical analysis parameters
The \( \Delta G_{OW} \) model was established with the simple linear regression by SPSS (Version 25.0). The determination coefficient \( (R^2)\), root mean square error \( (RMSE)\), and leave-one-out cross-validation coefficient \( (Q^2_{cv})\) were employed to characterize the performance of models. The significance of regression model and regression coefficients were verified by the \( F \)-test and \( t \)-test. Abnormal points include outliers and high leverage points, which were tested by using studentized residual and leverage values, respectively.

3. Results and discussion

3.1. \( \Delta G_{OW} \) model
\( \Delta G_{OW} \) of these 21 PAHs were estimated with the SMD model at the level of HF/MIDI6D. Based on the thermodynamic relationship, the log \( K_{OW} \) values can be estimated from \( \Delta G_{OW} \) with equation 1. The solvation free energy model was a simple linear regression model, which was developed based on the experimental and calculated \( \Delta G_{OW} \), and was obtained as equation 3. The experimental \( \Delta G_{OW} \) and log \( K_{OW} \) derived from calculated \( \Delta G_{OW} \) values, were listed in table 1.

\[
\Delta G_{OW} (Cal.) = 0.554 \Delta G_{OW} (Exp.) - 1.517 \\
\log K_{OW} = 25.0 \\
\]

\[
n = 21, R^2 = 0.925, RMSE = 0.207, Q^2_{cv} = 0.913
\]

where \( \Delta G_{OW} (Cal.) \) and \( \Delta G_{OW} (Exp.) \) represent the calculated and experimental \( \Delta G_{OW} \) values. As indicated by the statistics of \( R^2 \), \( RMSE \), and \( Q^2_{cv} \), this model had high goodness-of-fit, robustness, and predictive accuracy. As can be seen from figure 1(a), the experimental \( \Delta G_{OW} \) correlated well with the calculated ones with \( R^2 \) of 0.925. The equivarance of the residuals was confirmed by the scattered standardized residuals (figure 1(b)). The normal distribution of residuals was proved by figure 1(c) and figure 1(d). As for the \( F \)-test and \( t \)-test, the results were \( F=234.007 > F_α (1, n-2) \) and \( |t| ≥ t_{α/2} \), suggesting that the regression equation and the regression coefficient were significant. According to the studentized residuals and leverage values, no outliers and high leverage points were found.
Table 1. Experimental, calculated, and calibrated log $K_{OW}$ of PAHs.

| Name               | Dimeric ratios | $ΔG_{OW}$ | log $K_{OW}$ |
|--------------------|---------------|-----------|--------------|
|                    | Calculated    | Experimental |             | Experimental | Monomeric | Dimeric | Calibrated |
| Naphthalene        | 0.25          | 0.18       | -3.93        | 3.37          | 2.88      | 5.60    | 3.57       |
| 1.2-Dihydroacenaphthylene | 0.15      | 0.14       | -4.85        | 3.92          | 3.55      | 6.14    | 3.95       |
| Acenaphthylene     | 0.23          | 0.28       | -4.31        | 4.00          | 3.16      | 6.16    | 3.85       |
| Fluorene           | 0.23          | 0.17       | -4.94        | 4.18          | 3.62      | 6.87    | 4.36       |
| Anthracene         | 0.31          | 0.27       | -4.94        | 4.54          | 3.62      | 6.97    | 4.67       |
| Phenanthrene       | 0.31          | 0.33       | -4.81        | 4.57          | 3.52      | 6.73    | 4.52       |
| Pyrene             | 0.34          | 0.38       | -5.22        | 5.18          | 3.82      | 7.41    | 5.03       |
| Fluoranthene       | 0.31          | 0.35       | -5.32        | 5.22          | 3.90      | 7.63    | 5.07       |
| Benzo[a]fluorene   | 0.28          | 0.28       | -5.86        | 5.40          | 4.29      | 8.27    | 5.42       |
| Triphenylene       | 0.37          | 0.38       | -5.55        | 5.49          | 4.07      | 7.82    | 5.45       |
| Benzo[b]fluorene   | 0.37          | 0.37       | -5.93        | 5.75          | 4.35      | 8.16    | 5.45       |
| Naphthalene        | 0.38          | 0.36       | -5.95        | 5.76          | 4.36      | 8.21    | 5.81       |
| Benzo[b]fluoranthene | 0.29   | 0.31       | -6.22        | 5.80          | 4.56      | 8.59    | 6.04       |
| Chrysene           | 0.37          | 0.42       | -5.69        | 5.86          | 4.17      | 8.16    | 5.65       |
| Benzo[a]anthracene | 0.37          | 0.42       | -5.79        | 5.91          | 4.25      | 8.19    | 5.71       |
| Benzo[k]fluoranthene | 0.38      | 0.31       | -6.33        | 6.00          | 4.64      | 9.09    | 6.34       |
| Benzo[a]pyrene     | 0.40          | 0.37       | -6.10        | 6.04          | 4.47      | 8.73    | 6.15       |
| Perylene           | 0.40          | 0.48       | -5.94        | 6.25          | 4.35      | 8.31    | 5.92       |
| Benzo[ghi]perylen e | 0.42        | 0.42       | -6.36        | 6.50          | 4.66      | 9.06    | 6.50       |
| Coronene           | 0.44          | 0.39       | -6.77        | 6.75          | 4.96      | 9.58    | 7.01       |
| Dibenzo[a,h]anthracene | 0.43    | 0.43       | -6.67        | 6.75          | 4.89      | 9.25    | 6.77       |

(a) $ΔG_{OW}$(Cal.) = 0.554 $ΔG_{OW}$(Exp.) - 1.517
$R^2 = 0.925$

(b) Standardized residual

Standardized predictive value
3.2. **Calibration of log $K_{OW}$ with dimer formation**

As shown in Table 1, the $\Delta G_{OW}$-derived log $K_{OW}$ values of monomeric PAHs were lower than those experimental ones and the differences between them increased gradually with the increase of molecular weight ($M_w$). The reason may be related with the existence of dimer formation of PAHs. As PAHs might existed mostly in the configuration of single molecules and dimers, but rarely in the form of polymer. Therefore, for predicting log $K_{OW}$ of PAHs, the effect of the dimer was considered rather than polymer, and then the calibration method was obtained as equation 4.

$$\log K_{OW} = \text{monomer}\% \times \log K_{OW} (\text{monomer}) + \text{dimer}\% \times \log K_{OW} (\text{dimer})$$

where monomer% and dimer% stand for the ratio of monomer and dimer of PAHs, and the sum of them is supposed to be 1. The geometrical optimization of dimeric PAHs and calculation of $\Delta G_{OW}$ values were also conducted with SMD model at HF/Midi6D level. According to equation 4, the proportion of PAHs present in monomeric and dimeric formation was determined with experimental log $K_{OW}$ and predicted monomeric and dimeric log $K_{OW}$. As shown in Table 1, the experimental dimer ratios of PAHs increased with the increase of $M_w$. The higher the $M_w$, the stronger the lipophilicity of PAHs, and the more likely the PAHs existed in dimer. Then, it led to the greater deviation between the experimental and the predicted log $K_{OW}$.

QSAR models were established with 11 molecular structural descriptors by PLS regression to predict the dimeric ratios of PAHs without experimental data. Before modeling, the variable selection process presented by Ding [25] was conducted and 10 kinds of QSAR models with 2~11 independent variables were established. A QSAR model as equation 5 had the highest values of $R^2_{X(cum)}$, $R^2_{Y(cum)}$ and $Q^2_{cum}$, which indicated the highest goodness-of-fit and prediction ability. Furthermore, as shown in Table 1, the predicted dimer ratios correlated well with the experimental values.

$$Dimer\% = 0.089 - 2.24 \times 10^{-3}TE + 2.31 \times 10^{-4}M_w - 0.12\mu + 2.63 \times 10^{-4}\alpha + 2.88 \times 10^{-4}CM4 + 3.17 \times 10^{-4}C$$

$$n = 21, R^2_{X(cum)} = 0.997, R^2_{Y(cum)} = 0.699, Q^2_{cum} = 0.579$$
With monomeric and dimeric ratios of PAHs predicted by equation 5, and log KOW values of monomeric and dimeric PAHs, the predicted log KOW values were corrected with equation 4. As depicted in figure 2, the corrected log KOW agreed with the corresponding experimental values and R² increased from 0.925 to 0.963 after the calibration. Furthermore, the RMSE was decreased from 0.207 to 0.187. Therefore, the model performance was improved significantly after corrected with the dimer, and the QSAR model was reliable for calibrating log KOW values.

3.3. Comparison to other studies
Ferreira [6] developed a QSAR model for predicting log KOW values of PAHs with 17 experimental data. Ribeiro and Ferreira [7] predicted log KOW values by a QSAR model with 21 experimental data. Gusten [8] predicted the log KOW values of PAHs by molecular topology. With statistics presented in table 2, the values of R², RMSE, and QCV of this study was compared with the results of previously reported models mentioned above.

| Table 2. Comparison of R², RMSE, and QCV of log KOW model with reported models. |
|-----------------------------|-------|-------|---------|----------|------------------------|
| Model                      | n    | R²    | RMSE   | QCV      | Reference             |
|-----------------------------|------|-------|--------|----------|------------------------|
| QSAR                       | 17   | 0.992 | 0.179  | --       | [6]                    |
| QSAR                       | 21   | 0.975 | 0.181  | 0.968    | [7]                    |
| Topological features       | 21   | 0.972 | 0.207  | --       | [8]                    |
| ΔGOW model                 | 21   | 0.963 | 0.187  | 0.957    | This study             |

Although the statistics of the ΔGOW model was not optimal, the quality of the model was not entirely dependent on the statistical parameters, but also on other characteristics. According to the thermodynamic relationship, the log KOW values were obtained from ΔGOW values. Therefore, the ΔGOW model had universal application domain and clear explanation of model mechanism. However, the QSAR models were constrained by the application domain and were more dependent on the experimental log KOW values. Furthermore, multiple molecular structural descriptors were needed to establish the QSAR model and the topological features model, the mechanism of them was difficult to explain. It is therefore to be concluded that the ΔGOW model is a reliable method for prediction of log KOW of alkyl-PAHs.
4. Conclusion
In this study, a $K_{\text{OW}}$ prediction model for alkyl-PAHs was established based on $\Delta G_{\text{OW}}$. The dimer form of polycyclic aromatic compound had a strong influence on the partition behavior and partition coefficient between aqueous phase and $n$-octanol phase. The $R^2$, RMSE, and $Q^2_{CV}$ of the predictive model was significantly improved after the effect of dimer configuration was considered. Compared with results from other studies, the $\Delta G_{\text{OW}}$ model had comparable $R^2$ and RMSE. In addition, the $\Delta G_{\text{OW}}$ model had the advantages of universal application domain, low dependence on the experimental values and simple model mechanism.

Acknowledgements
This work was financially supported by the National Natural Science Foundation of China (51479016) and the Natural Science Foundation of the Liaoning Province of China (20180510004).

References
[1] Perelo L W 2010 J. Hazard. Mater. 177 8189
[2] Casal C S, Arbilla G, and Correa S M 2014 Atoms. Environ. 96 107-16
[3] Abdel-Shafy H and Mansour M S M 2016 Egypt. J. PET. 25 107-23
[4] Wang Y R and Wang Y H 2007 Anal. Lett. 40 307-19
[5] Sahu S K and Pandit G G 2003 J. Liq. Chromatogr. R. T. 26 135-46
[6] Ferreira M M C 2001 Chemosphere 44 125-46
[7] Ribeiro F A L and Ferreira M M C 2003 J. Mol. Struc-Theochem 663 109-26
[8] Gusten H, Horvatic D and Sabljic A 1991 Chemosphere 23 199-213
[9] Miertus S, Scrocco E and Tomasi J 1981 Chem. Phys. 55 117-29
[10] Klamt A 2005 Elsevier Science Ltd (Amsterdam: The Netherlands)
[11] Klamt A 2011 WIREs: Comput. Mol. Sci. 1 699-709
[12] Cramer C J and Truhlar D G 2008 Acc. Chem. Res. 41 760-8
[13] Gupta M, da Silva E F and Svendsen H F 2012 J. Phys. Chem. B 116 1865-75
[14] Marenich A V, Cramer C J and Truhlar D G 2009 J. Phys. Chem. B 113 6378-96
[15] Ouimet J A and Paluch A S 2020 J. Comput. Aided Mol. Des. 34 575-88
[16] Guan D, Liu R and Matthews S 2020 J. Comput. Aided Mol. Des. 34 511-22
[17] Hwang J, Dell B E, Li P, Kozik M E, Smith M D, Shimizu K D 2015 Chem. Sci. 6 4358-64
[18] Sherrill C D 2013 Acc. Chem. Res. 46 1020-8
[19] Guo X J, Jin X, Lv X F, Pu Y Y, Bai F 2015 Environ. Sci. Technol. 49 7926-33
[20] Wild E, Cabrerizo A, Dachs J and Jones K C 2008 J. Phys. Chem. A 112 11699-703
[21] Mackay D, Shiu W Y and Ma K C 1998 Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals (Lewis: London)
[22] Frisch M J et al 2009 Gaussian Inc. (Wallingford CT)
[23] Li W R, Ding G H, Gao H, Zhuang Y T, Gu X Y, Peijnenburg W J G M 2020 Chemosphere 242 125246
[24] Stewart J J P 2013 J. Mol. Model. 19 1-32
[25] Ding G H, Chen J W, Qiao X L, Huang L P, Lin J and Chen X Y 2006 Chemosphere 62 1057-63