Predicting low density polyethylene-air partition coefficients using theoretical linear solvation energy relationships
Tengyi Zhu, Mao Li, Jing Wu, Yajun Wang and Rajendra Prasad Singh

ABSTRACT
Passive air samplers have proven to be a widely used technique in measuring hydrophobic organic compounds (HOCs) in the atmosphere. Chemical partition coefficients between the sorbent material and air ($K_{PE}$) are significant for calculating chemical concentrations from passive sampling devices. We have established a theoretical linear solvation energy relationship (TLSER) model for predicting $K_{PE}$ values by molecular descriptor. With the TLSER model, McGowan volume ($V$), chemical hardness ($\eta$) and dipole moment ($DM$) were screened as the most relevant variables. The model had a high correlation coefficient ($R^2$), leave-one-out cross-validation coefficient ($Q^2_{LOO}$) and bootstrapping coefficient ($Q^2_{BOOT}$). To our knowledge, this is the first attempt to predict the LDPE-air partition coefficient using the TLSER model. Statistical parameters, determination coefficient ($R^2$) and cross-validation coefficients ($Q^2$) ranged from 0.896 to 0.931 and 0.883 to 0.909, respectively, which indicated that the TLSER model appropriately fitted the results, and also showed robustness and predictive capacity. Mechanism interpretation suggested that the factors governing the partition process for LDPE and air were the McGowan volume and molecular orbital energies. The results of this study provide a good tool for predicting log $K_{PE}$ values of HOCs, within the applicability domains to reduce cost and time for innovation.

Key words | applicability domain (AD), hydrophobic organic compounds (HOCs), low density polyethylene-air partition coefficient ($K_{PE}$), theoretical linear solvation energy relationship (TLSER)

INTRODUCTION
Hydrophobic organic compounds (HOCs) are usually carcinogenic, teratogenic and mutagenic, and pose a great danger to the natural and human survival environment (Fujii et al. 2007). The atmosphere is one of the core matrices for global monitoring of persistent organic pollutants (POPs). The urgent need for the effective evaluation of air pollutants has contributed to the innovative development of atmospheric POPs sampling technology. Great progress has been made in the development and application of passive atmospheric sampling (PAS) technology in the last decade (Hazrati & Harrad 2007; Tuduri et al. 2012). PAS is becoming an effective complementary sampling technology for high volume air samplers, with its extraordinary advantage in synchronic monitoring of POPs in the atmosphere across vast territories. Passive sampling technology is a kind of equilibrium sampling technology for enriching organic pollutants in environmental media based on molecular diffusion or the infiltration principle. It integrates sampling, separation, concentration and analysis without the need for external energy. Many different materials have been used as sorbents in various passive sampling devices, including semipermeable membrane devices (Huckins et al. 1990; Petty et al. 1993; Lohmann et al. 2001; Khairy & Lohmann 2014), low density polyethylene (LDPE) film (Bartkow et al. 2006; Denberg et al. 2007; Khairy & Lohmann 2014; Zhu et al. 2018), and polyurethane foam (PUF) devices.

© IWA Publishing 2018
Journal of Water Supply: Research and Technology—AQUA | 67.8 | 2018
doi: 10.2166/aqua.2018.052
The experimental data sets of $K_{\text{PE-a}}$ were collected from literature (Tables S1–S5 in the supplementary material, available with the online version of this paper) for five different classes of HOCs, including 41 PAH congeners, 22 PCB congeners, 4 PBDEs, 10 different OCPs, and 20 polychlorinated dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs).

The total data set covers 97 compounds with a wide range of $K_{\text{PE-a}}$ values (from 4.93 to 11.7). The normal distribution for the total data set with a mean of 8.40 and a standard deviation of 1.42 is shown in Figure 1. All
compounds were randomly divided into a training set and validation set in the ratio of 4:1. The data in the training set was used to establish the model and the data in the validation set was used to validate the model.

Calculation of molecular descriptors

The initial organic molecular structure was generated by ChemBio3D Ultra 12.0 software. The molecular structures were optimized with the MM2 method contained in the ChemBio3D software (at the minimum root mean square (RMS) gradient $= 0.001$) (Zhu et al. 2018). Then, MOPAC2016 software was applied to generate molecular descriptors (MTLSER parameters) with the minimize energy/geometry method, using keywords: PM7, Dielectric Constant $= 78.6$, EF, GNORM $= 0.01$, MULLIK POLAR SHIFT $= 80$ (MOPAC2016 & Stewart 2016). Based on the optimized geometric structures from MOPAC, molecular descriptors were calculated by using the PaDEL-Descriptor software (Version 2.21) (Yap 2011). As a result, 1051 descriptors were retained in the final set.

Model development and evaluation

The TLSER concept is a general approach to describe solute-solvent interactions. The McGowan volume ($V$) describes the cavity term. The dipole moment ($DM$) represents the polarity term. The $q^-$ and $q^+$ are the net charge of the most negative and most positive atoms, respectively, which describe the hydrogen bonding term. The electron donor-acceptor terms are represented by chemical hardness ($\eta$, $(E_{LUMO} - E_{HOMO})/2$), chemical potential ($\mu$, $(E_{LUMO} + E_{HOMO})/2$), ionization potential ($I$, $-E_{HOMO}$), electrophilicity index ($\omega$, $\mu^2/2\eta$) and electron affinity ($A$, $-E_{LUMO}$).

Stepwise multiple linear regression analysis in the SPSS software was performed on selected variables and developed the predictive model based on the following schemes:

$$\text{TLSER: } \log K_{PE-a} = aV + bDM + cq^+ + dq^- + e\eta + f\mu + g\omega + hI + iA + \text{constant}$$

The lowercase letters $a, b, c, d, e, f, g, h$ and $i$ are regression coefficients. The best model was selected, with the lowest number of chemical descriptors and the maximum values of the determination coefficient ($R_{\text{adj}}^2$). The statistical parameters $R_{\text{adj}}^2$, leave-one-out cross-validated ($Q_{\text{LOO}}^2$), bootstrap method ($Q_{\text{BOOT}}^2$) (1/5, 5,000 iterations) and the root mean squared error (RMSE) were calculated to assess the predictive ability of the model. $R_{\text{ext}}^2$, RMSE$_{\text{ext}}$ and external explained variance ($Q_{\text{ext}}^2$) were employed to evaluate the predictive capacity of the models. The severity of multicollinearity of each predictor variance was also characterized by the variance inflation factor (VIF). The applicability domain (AD) was characterized by the Euclidean distance method (OECD 2007). The calculation method of the Williams plot and the statistical parameters of the models are presented in the supplementary material (available with the online version of this paper).

RESULTS AND DISCUSSION

Development and validation of the model

In the scheme of this model, property represents solute-solvent interactions. The McGowan volume ($V$) describes the cavity term. The dipole moment ($DM$) represents the polarity term. The $q^-$ and $q^+$ are the net charge of the most negative and most positive atoms, respectively, which describe the hydrogen bonding term. The electron donor-acceptor terms are represented by chemical hardness ($\eta$, $(E_{LUMO} - E_{HOMO})/2$), chemical potential ($\mu$, $(E_{LUMO} + E_{HOMO})/2$), ionization potential ($I$, $-E_{HOMO}$), electrophilicity index ($\omega$, $\mu^2/2\eta$) and electron affinity ($A$, $-E_{LUMO}$).

Stepwise multiple linear regression analysis in the SPSS software was performed on selected variables and developed the predictive model based on the following schemes:

$$\text{TLSER: } \log K_{PE-a} = aV + bDM + cq^+ + dq^- + e\eta + f\mu + g\omega + hI + iA + \text{constant}$$

The lowercase letters $a, b, c, d, e, f, g, h$ and $i$ are regression coefficients. The best model was selected, with the lowest number of chemical descriptors and the maximum values of the determination coefficient ($R_{\text{adj}}^2$). The statistical parameters $R_{\text{adj}}^2$, leave-one-out cross-validated ($Q_{\text{LOO}}^2$), bootstrap method ($Q_{\text{BOOT}}^2$) (1/5, 5,000 iterations) and the root mean squared error (RMSE) were calculated to assess the predictive ability of the model. $R_{\text{ext}}^2$, RMSE$_{\text{ext}}$ and external explained variance ($Q_{\text{ext}}^2$) were employed to evaluate the predictive capacity of the models. The severity of multicollinearity of each predictor variance was also characterized by the variance inflation factor (VIF). The applicability domain (AD) was characterized by the Euclidean distance method (OECD 2007). The calculation method of the Williams plot and the statistical parameters of the models are presented in the supplementary material (available with the online version of this paper).
follows:

$$\log K_{PE-a} = 9.544 + 3.768V - 1.949\eta + 0.133DM$$  \hspace{0.5cm} (3)$$

$n_{tra} = 77$,  $R^2_{adj} = 0.896$,  $Q^2_{LOO} = 0.883$,  $Q^2_{BOOT} = 0.887$,  $RMSE_{tra} = 0.421$,  $p < 0.001$,  $n_{ext} = 20$,  $R^2_{ext} = 0.931$,  $Q^2_{ext} = 0.909$,  $RMSE_{ext} = 0.465$, where $n_{tra}$ and $n_{ext}$ are the numbers of chemicals in the training set and the validation set, respectively.

The TLSER model showed satisfactory statistical parameters with $R^2_{adj}$, $Q^2_{LOO}$ and $Q^2_{BOOT}$ of 0.896, 0.883 and 0.887 in the training set, and $R^2_{ext}$ and $Q^2_{ext}$ of 0.931 and 0.909 in the validation set, respectively. According to the criteria ($Q^2 > 0.5$, $R^2 > 0.6$) proposed by Golbraikh et al. (2003), these statistical results suggested that the TLSER model had satisfactory stability and predictive ability. The results of the current study show that the importance of molecule descriptors follows the order: cavity term > electron donor-acceptor term > polarity term. The cavity term (McGowan volume, \(V\)) is the most significant variable in the TLSER model. The observations from this study are consistent with the results from Liu et al.’s (2017) investigations.

The model for HOCs had smaller RMSE\(_{tra}\) values (0.421 and 0.465), indicating goodness of fit and predictive capacity. The correlation between observed and predicted $\log K_{PE-a}$ values is presented in Figure 2. The external predictive ability of the model can be validated by a small RMSE\(_{ext}\) (0.465) and a high $Q^2_{ext}$ (0.909). The statistical results indicated that the predicted $\log K_{PE-a}$ values were consistent with the observed values in this study. In addition, the predictive variable involved in the models and corresponding Student’s $t$ test ($t$), statistical significance ($p$) and variance inflation factor (VIF) values are also listed in Table 1. The statistical parameters of the models indicated that there was no multicollinearity among predictive variables (descriptors), with the VIF values < 10 for all the predicted descriptors.

### Applicability domain

The evaluation results of Euclidean distance for the TLSER model are shown in Figure 3. All the compounds in the training and validation sets were located at the accepted domain, which indicated that the training sets for the TLSER model had excellent representativeness. Considering the standardized residuals (\(\delta\)), there was one chemical (2,3,7-chlorinated dibenzo-p-dioxin) in the training set with $|\delta| > 3$, identified as an outlier (Figure 4). This showed that there are no outliers in the TLSER model except for the dioxin. But dioxin was in the TLSER model because other

<table>
<thead>
<tr>
<th>Model</th>
<th>Descriptor</th>
<th>$t$</th>
<th>$p$</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLSER</td>
<td>$V$</td>
<td>18.328</td>
<td>0.000</td>
<td>1.638</td>
</tr>
<tr>
<td></td>
<td>$\eta$</td>
<td>-9.620</td>
<td>0.000</td>
<td>2.359</td>
</tr>
<tr>
<td></td>
<td>$DM$</td>
<td>2.085</td>
<td>0.000</td>
<td>2.361</td>
</tr>
</tbody>
</table>

\[\text{Table 1} \] The descriptors involved in the TLSER model, and corresponding Student’s $t$ test ($t$), statistical significance ($p$) and variance inflation factor (VIF) values

![Figure 2](image-url)  \hspace{0.5cm} \text{Figure 2} | Plot of the predicted versus experimental $\log K_{PE-a}$ values.

![Figure 3](image-url)  \hspace{0.5cm} \text{Figure 3} | AD of developed TLSER model evaluated by the Euclidean distance-based approach.
Mechanism interpretation

In the TLSER model, $V$, $\eta$ and $DM$ are the most important descriptors. Among them, $V$ and $DM$ were positively correlated with $\log K_{PE-a}$, and $\eta$ was negatively correlated with $\log K_{PE-a}$. $V$ is the McGowan volume. This descriptor, which describes the cavity term, has been widely used to analyze physicochemical properties of solutes in chemistry. It can be calculated from the sum of all atomic contributions for the compound, subtracting 6.56 cm$^3$ mol$^{-1}$ for each bond, regardless of whether it is a single, double or triple bond (Zhao et al. 2003a; Liu et al. 2017). The McGowan volume is a measure of the energy needed to overcome the solvent molecular interactions to form a cavity for the solute molecule. Furthermore, the McGowan volume is also related to polarizability and/or hydrophobicity of the solute. So, the McGowan volume can be a measure of the exoergic solute-solvent general dispersion interaction that arises through solute polarizability (Zhao et al. 2003b; Sprunger et al. 2007; Liu et al. 2017). The molecular orbital theory as formulated by Mulliken and Hund has been very successful in explaining and predicting chemical behaviors for an enormous number of molecules (Zhou & Parr 1990). The highest occupied molecular orbitals (HOMO) and the lowest unoccupied molecular orbitals (LUMO) play an important role in governing chemical reactions. The energy difference between the LUMO of an electron acceptor and the HOMO of an electron donor has long been used as a reactivity index (Zhou & Parr 1990). Previous investigations have revealed that the $E_{HOMO}-E_{LUMO}$ gap is an important stability index for the individual species concerned (Zhou & Parr 1990). So, chemical hardness ($\eta$, $(E_{HOMO}-E_{LUMO})/2$) is a good index of orientation of electrophilic aromatic substitution. A large HOMO–LUMO gap implies high stability. High stability of a molecule reflects its lower reactivity toward chemical reactions in some sense (Zhou & Parr 1990). In the sense of lower reactivity in chemical reactions, a large $\eta$ indicates high stability for the molecule of the compound. Dipole moment ($DM$) is an overall descriptor of the electronic interaction among the solvent and solute molecules. It is well known that $DM$ is extremely important for various physicochemical properties, and many descriptions have been made to quantify the polarity effect. For example, molecular polarity indicates chromatographic retention on a polar stationary phase. The most obvious and often used quantity that describes polarity is the molecular dipole moment (Karelson et al. 1996). The dipole moment decreases the adsorption coefficients ($K_{PE-a}$ value). This is reasonable, because greater $DM$ would imply greater dipole-dipole and dipole-induced dipole interactions between solutes and air, resulting in a greater tendency to be partitioned into air and a lower toxicity (Chen et al. 1996). As demonstrated in this study, McGowan volume, chemical hardness and dipole moment were selected as the most relevant variables in the TLSER model. It can be seen that large molecules cannot pass through the cavity.
through the membrane easily, and need more energy to permeate into the polymer membrane, which is not conducive to partitioning behaviors (Yang et al. 2007; Bao et al. 2011; Liu et al. 2017).

Model comparison

Several researchers have developed different models for predicting the $K_{\text{PE-a}}$ value (Table 2). Bartkow et al. (2004) presented $K_{\text{PE-a}}$ for three PAHs based on the comparison between their accumulation in PE film and gaseous concentrations from active high-volume sampling (Bartkow et al. 2004). In a follow-up study by Kennedy et al. (2007), a more rigorous correlation of $K_{\text{PE-a}}$ versus octanol-air partition coefficient ($K_{\text{oa}}$) was suggested for six PAHs. Ma et al. (2010) reported a correlation of $K_{\text{oa}}$ versus $K_{\text{aw}}$ values for predicting $K_{\text{PE-a}}$. $K_{\text{PE-a}}$ values reported by Khairy & Lohmann (2012) were in very good agreement with those obtained according to the equation proposed by Bartkow et al. (2004), with factor difference less than 46% in all the investigated individual PAHs and similar to the prediction by Lohmann (2012) based on the partitioning of PAHs between PE and water (Khairy & Lohmann 2012). In previous work, Khairy & Lohmann (2013) reported that the $K_{\text{PE-a}}$ versus $K_{\text{oa}}$ relation for OCPs explained only 32% of the total variability in the data, and $K_{\text{oa}}$ values were not suitable as predictors for $K_{\text{PE-a}}$. And they obtained a strong linear relationship between log $K_{\text{PE-a}}$ and log $P_L$ (subcooled liquid vapor pressure), which included only 59 compounds from three different chemical classes (PCBs, PBDEs and OCPs) (Khairy & Lohmann 2014). The relationship with log $K_{\text{PE-a}}$ and log $K_{\text{oa}}$ is shown in Figure S1 (in the supplementary material, available with the online version of this paper). A linear relationship between log $K_{\text{PE-a}}$ and log $K_{\text{oa}}$ was obtained. This suggested that uptake of gase-phase HOCs by PE is not as efficient as their uptake by octanol, probably due to the rigid matrix of the polymer (Lohmann 2012). This field calibration model was only performed for PAHs, PCBs, PBDEs and OCPs. And $K_{\text{pe-a}}$ are not available for other classes of HOCs (Table 2).

Few TLSER models existed in previous studies, which could predict the distribution behavior of molecules between LDPE and the gas phase. In the current study, a TLSER model was developed with the aim of optimally correlating with the LSER model and hence being generally applicable to solute/solvent interactions. Comparing with previous results, the models proposed in this study had many advantages, such as higher correlation coefficient ($R^2$), smaller value of RMSE and higher cross-validated coefficient ($Q^2$). Furthermore, the present model had a broad application domain, which included 97 compounds and contained much more structurally diverse compounds.

### Practical applications

PAS has proven to be a widely used technique for measuring HOCs in the atmosphere. Chemical partition coefficients between the sorbent material and air ($K_{\text{pe-c}}$) are significant for calculating chemical concentrations from passive sampling devices. We have established a TLSER model

---

**Table 2** Overview of different predictive models for $K_{\text{PE-a}}$

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$m^a$</th>
<th>$n^b$</th>
<th>$R^2$</th>
<th>Chemicals</th>
<th>RMSE$^c$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{oa}}$</td>
<td>1</td>
<td>3</td>
<td>0.95</td>
<td>PAHs</td>
<td>NR</td>
<td>Bartkow et al. (2004)</td>
</tr>
<tr>
<td>$K_{\text{oa}}$</td>
<td>1</td>
<td>6</td>
<td>0.84</td>
<td>PAHs</td>
<td>NR</td>
<td>Kennedy et al. (2007)</td>
</tr>
<tr>
<td>$K_{\text{oa}}$</td>
<td>1</td>
<td>15</td>
<td>NR</td>
<td>PAHs</td>
<td>NR</td>
<td>Ma et al. (2010)</td>
</tr>
<tr>
<td>$K_{\text{oa}}$</td>
<td>1</td>
<td>12</td>
<td>0.99</td>
<td>PAHs and alkylated PAHs</td>
<td>0.052</td>
<td>Khairy &amp; Lohmann (2012)</td>
</tr>
<tr>
<td>$P_L$</td>
<td>1</td>
<td>17</td>
<td>0.90</td>
<td>OCPs</td>
<td>0.19</td>
<td>Khairy &amp; Lohmann (2013)</td>
</tr>
<tr>
<td>$P_L$</td>
<td>1</td>
<td>59</td>
<td>0.96</td>
<td>PCBs, PBDEs and OCPs</td>
<td>0.22</td>
<td>Khairy &amp; Lohmann (2014)</td>
</tr>
<tr>
<td>$K_{\text{oa}}$</td>
<td>1</td>
<td>59</td>
<td>0.96</td>
<td>PCBs, PBDEs and OCPs</td>
<td>0.23</td>
<td>Khairy &amp; Lohmann (2014)</td>
</tr>
<tr>
<td>$V, \eta, DM$</td>
<td>3</td>
<td>97</td>
<td>0.90</td>
<td>PAHs, PCBs, PBDEs, OCPs, PCDDs and PCDFs</td>
<td>0.34</td>
<td>Current study</td>
</tr>
</tbody>
</table>

$^a$m: the number of parameters; $^b$n: the number of compounds; $^c$RMSE: the root mean squared error; NR: not reported; $K_{\text{oa}}$: the octanol-air partition coefficient; $P_L$: subcooled liquid vapor pressure; $V$: McGowan volume; $\eta$: chemical hardness; $DM$: dipole moment.
for predicting $K_{\text{PEa}}$ values by molecular descriptor. The model had high correlation coefficient ($R^2$), leave-one-out cross-validation coefficient ($Q^2_{\text{LOO}}$) and bootstrapping coefficient ($Q^2_{\text{BOOT}}$). Statistical parameters indicated that the TLSER model appropriately fitted the results, and also showed robustness and predictive capacity. The results of the current study provide a good tool for predicting log $K_{\text{PEa}}$ values of HOCs, within the applicability domains to reduce cost and time for innovation.

Furthermore, the current study proved the advantages of the TLSER model: (i) the chemical descriptors in the TLSER model can be easily obtained by computation instead of experimentation, so a large amount of expense and time can be saved, and (ii) the quantum chemical descriptors have clear physicochemical interpretations, and interpretation of the correlation equations can suggest modes of interaction between solute and solvent molecules. Future work should focus on developing tasks that assess the application of TLSER models for a wider range of chemicals, including emerging contaminants, the relationship between log $K_{\text{PEa}}$ and many different predictors, and ways to account for the McGowan volume, chemical hardness and dipole moment that result in distribution behavior in PE among classes of chemicals or individual chemicals.

**CONCLUSIONS**

In the current study, the TLSER model was successfully constructed by quantum chemical descriptors for predicting low density polyethylene-air partition coefficients. The TLSER model was based on a large data set, including 97 compounds from five different chemical classes, which implied that it had a wide AD. With the TLSER model, McGowan volume, chemical hardness and dipole moment were screened as the most relevant variables. Mechanism explanation for the TLSER model was also performed. Statistical analysis suggested that the model had good adaptability and high prediction performance. This study demonstrated that the TLSER approach is useful to predict log $K_{\text{PEa}}$ values for complex molecules, within the applicability domains to reduce experimental cost and time for innovation. The method is easy to use and it has general applicability. Meanwhile, it can provide a theoretical basis and technical support for quickly obtaining the partition coefficient of HOCs in passive sampling materials.

**ACKNOWLEDGEMENTS**

The study was supported by the National Natural Science Foundation of China (Grant No. 21607123) and the science and technology projects fund of Yangzhou City (Grant No. YZ2016112). This study was also sponsored by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

**REFERENCES**


First received 30 April 2018; accepted in revised form 25 July 2018. Available online 6 August 2018