Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jes

JES JOURNAL OF ENVIRONMENTAL SCIENCES

www.jesc.ac.cn

Novel quantitative structure activity relationship models for predicting hexadecane/air partition coefficients of organic compounds

Ya Wang[1], *Weihao Tang*[2], *Zijun Xiao*[2], *Wenhao Yang*[1], *Yue Peng*[1], *Jingwen Chen*[2], *Junhua Li*[1,*]

[1] *State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China*

[2] *Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Linggong Road 2, Dalian 116024, China*

ARTICLE INFO

Article history:
Received 26 August 2021
Revised 20 October 2021
Accepted 28 October 2021
Available online 1 February 2022

Keywords:
L value
Quantitative structure-activity relationship
Multiple linear regression
Support vector machine
Oraganosilicon compounds

ABSTRACT

Predicting the logarithm of hexadecane/air partition coefficient (*L*) for organic compounds is crucial for understanding the environmental behavior and fate of organic compounds and developing prediction models with polyparameter linear free energy relationships. Herein, two quantitative structure activity relationship (QSAR) models were developed with 1272 *L* values for the organic compounds by using multiple linear regression (MLR) and support vector machine (SVM) algorithms. On the basis of the OECD principles, the goodness of fit, robustness and predictive ability for the developed models were evaluated. The SVM model was first developed, and the predictive capability for the SVM model is slightly better than that for the MLR model. The applicability domain (AD) of these two models has been extended to include more kinds of emerging pollutants, i.e., oraganosilicon compounds. The developed QSAR models can be used for predicting *L* values of various organic compounds. The van der Waals interactions between the organic compound and the hexadecane have a significant effect on the *L* value of the compound. These *in silico* models developed in current study can provide an alternative to experimental method for high-throughput obtaining *L* values of organic compounds.

Introduction

Hexadecane/air partition coefficient for the organic compound ($K_{hexadecane/air}$) is a key behavioral parameter for characterizing their migration between organic phase and air (Brown, 2014; Stenzel et al., 2012; Bronner et al., 2010; Poole et al., 2009), which is one of the foundations for comprehensively understanding their environmental partitioning behavior and fate. Moreover, the logarithmic form of $K_{hexadecane/air}$, i.e., *L*, is also an important descriptor characterizing the non-specific intermolecular interactions being involved in different partition processes, and it plays an influential role in the polyparameter linear free energy relationships (pp-LFERs) (Endo and Goss, 2014a, b; Stenzel et al., 2013;

* Corresponding author.
E-mail: lijunhua@tsinghua.edu.cn (J. Li).

https://doi.org/10.1016/j.jes.2021.10.033
1001-0742/

Wang et al., 2017; Yaman et al., 2020; Zhu et al., 2020, 2019; Zhao et al., 2018).Especially for environmental modeling with pp-LFERs to predict the partition behavior of compounds, the *L* value is indispensable. Therefore, obtaining *L* values is of great significance not only for developing environmental prediction models based on pp-LFERs but also for comprehending the environmental behavior and assessing the environmental risks of organic compounds.

As is known to all, *L* is defined as $\log K_{hexadecane/air}$, where $K_{hexadecane/air} = C_{hexadecane}/C_{air}$, $C_{hexadecane}$ represents the equilibrium concentration of solute in hexadecane and C_{air} is the equilibrium concentration of solute in air. Generally, it can be determined by using gas-liquid chromatography or headspace method (Poole et al., 2009; Abraham et al., 1987).However, these conventional experimental methods only can be applied to the measurement of *L* values with a relatively smaller range for volatile organic compounds. Recently, new experimental methods have been established (Brown, 2014), so that more and more compounds can be determined and the range of *L* value has been broaden. Up to date, the number for available *L* values exceeds 1000. Considering more than 140,000 commercially used chemicals, it is tedious and unreasonable to measure their *L* values with experimental methods seriatim. Hence, it is necessary for developing prediction methods to get *L* values highly efficiently.

Several prediction methods including prevalent quantitative structure-activity relationship (QSAR) have been established for obtaining *L* values of organic compounds (Brown, 2014; Stenzel et al., 2012; Bronner et al., 2010). With reference to the Organization for Economic Cooperation and Development (OECD) principles (OECD, 2007), i.e., (1) a defined endpoint; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measure of goodness-of-fit, robustness and predictivity and (5) a mechanistic interpretation, if possible, it is not difficult to find that the evaluation parameters for the previous prediction methods are incomplete, and there is a lack of mechanistic interpretation for the established QSAR model. In addition, for the emerging environmental contaminants organosilicon compounds (Alton and Browne, 2020; Bzdek et al., 2014; Fairbrother and Woodburn, 2016; Grabitz et al., 2020; Krogseth et al., 2016; Panagopoulos et al., 2017; Rücker and Kümmerer, 2015; Tang et al., 2015; Wu and Johnston, 2017; Zhi et al., 2021), which have received increasing concerns, the prediction for their *L* values is also highly meaningful. The applicability domain (AD) of the previous prediction model for *L* values need to be further amplified, so that it can cover more kinds of organosilicon compounds. In terms of the algorithm being utilized for the model establishment, multiple linear regression (MLR) (Galán-Madruga et al., 2022; Cheng et al., 2020; Zhang et al., 2020; Li et al., 2020; Cho et al., 2018; Lan et al., 2018; Zhao et al., 2018; Liu et al., 2017, 2016)and support vector machine (SVM) (Wang et al., 2009; Wang et al., 2019; Goudarzi and Goodarzi, 2008) are popular in environmental modeling for predicting environmental behavioral parameters of organic compounds. Nevertheless, the SVM algorithm has not been applied for predicting the *L* values of organic compounds yet, and the differences between the MLR and SVM models for estimating the *L* values are still unclear.

To address the above mentioned issues, *L* values for 1272 organic compounds with various molecular structures including organosilicon compounds were collected for establishing QSAR models in the present study. The MLR and SVM algorithms were used for developing models to estimate *L* values respectively. Furthermore, we evaluated the goodness of fit, robustness and predictive ability for the developed MLR and SVM models according to the OECD guidelines for QSAR development and validation. Besides, the comparison was implemented for revealing the differences between the MLR and SVM models. The mechanistic interpretation was also provided in this study.

1. Materials and methods

1.1. *Data set*

The experimental *L* values at 25°C for 1476 chemicals were obtained from the UFZ-LSER database (UFZ-preselected published values) (UFZ-LSER database v 3.2.1). After deleting the inorganic compounds and the organic compound whose structure being far away from the structural domain, 1272 organic compounds were retained in the total data set. The details for names, CAS numbers, experimental and predicted *L* values from the established models were provided in the Supplementary material (**Table S1**). The total data set was randomly split into a training set and a validation set with a ratio of 4:1. The training set consists of 1018 organic compounds and the validation set covers 254 different compounds. The *L* values in the training set are in the range of –0.817∼17.740, and the *L* values in the validation set are in the range of –0.403 ∼ 13.980. The distribution for the *L* values with mean of 5.552 and standard deviation (SD) of 3.023 in the total data set was performed in **Fig. S1**. Besides, the distribution for the *L* values in the training and validation set (**Fig. S2**) was also shown in the Supplementary material.

1.2. *Molecular structure descriptors*

Three dimensional molecular structures for the 1272 organic compounds were obtained from the website of Chemical Book (https://www.chemicalbook.com/ProductIndex.aspx). We pre-optimized these molecular structures with MM2 method (Schnur et al., 1991)by using ChemBio3D Ultra (Version 12.0) (http://www.cambridgesoft.com.). Afterwards, these optimized structures were further optimized with PM7 method (Stewart, 2013) in the MOPAC 2016 program (http://openmopac.net/MOPAC2016.html) for getting the most stable structures. The gradient norm for the structural optimization was set as 0.001 kcal/mol/Angstrom. 20 molecular structural descriptors (**Table S2**) were collected from the quantum chemical calculations for establishing prediction models. Besides, based on the optimized molecular structures, values for 372 molecular structural descriptors were calculated with Dragon software (Version 6.0) (Talete srl, 2012.). After removing the descriptors with constant and near-constant values, 225 Dragon descriptors were obtained.

Table 1 – Depiction for the predictor variables and the corresponding standardized coefficients, t, *p* values.

Descriptors	Depiction	Standardized coefficients	t*	p*
S_A	conductor-like screening model (COSMO) area	0.629	28.416	< 0.001
Mi	mean first ionization potential (scaled on carbon atom)	–0.282	–33.880	< 0.001
SCBO	sum of conventional bond orders (H-depleted)	0.171	6.131	< 0.001
nH	number of hydrogen atoms	–0.097	–9.120	< 0.001
*n*CIC	number of rings (cyclomatic number)	0.178	9.982	< 0.001
Hy	hydrophilic factor	0.112	16.363	< 0.001

* *t* statistic is obtained by *t*-test; *p* is the significance level of *t*-test.

1.3. *Model development*

1.3.1. *Multiple linear regression (MLR)*

The *L* values for 1018 organic compounds in the training set were used as dependent variable and the molecular structural descriptors were used as predictive variables. Stepwise multiple linear regression (MLR) in the software package SPSS (Version 22.0) was applied for selecting variables and establishing prediction models. The model with the maximum determination coefficient (R^2) and the least number of variables was chosen as the optimal prediction model.

1.3.2. *Support vector machine (SVM)*

Based on the same descriptors with those in the optimal MLR model, support vector machine (SVM) prediction model was developed. The SVM model can be expressed as the following equation,

$$y(x)_{pre} = \sum_{i=1}^{n} \alpha_i K\left(x_i, x_j\right) + b \quad (1)$$

where, $y(x)_{pre}$ denotes the predicted values; α_i is Langrange multiplier; $K(x_i, x_j)$ represents kernel function; b denotes a threshold parameter. In this study, Gaussian radial basis function (RBF) [$\exp(-\gamma\|x_i - x_j\|^2)$], a commonly used kernel function, was utilized for producing higher precision prediction. In terms of Gaussian RBF, x_i and x_j are two independent molecular structural descriptors, and γ is a regularization parameter. In order to get the best SVM model, we should minimize the following function,

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad (2)$$

where $1/2\|w\|^2$ denotes the margin, $\Sigma\xi_i$ represents the sum of training errors, and C is a tuning parameter.

1.4. *Model evaluation and characterization of ADs*

According to the OECD principles on the development and validation of QSAR models (OECD, 2007),the goodness of fit, robustness and predictive ability for the developed MLR and SVM models were evaluated with these parameters, i.e., coefficients of determination (R^2), root mean square error for training set and validation set ($RMSE_t$ and $RMSE_v$), leave-one-out cross-validated Q^2 (Q^2_{LOO}) and external explained variance (Q^2_v). In addition, Williams plots of standardized residuals (δ^*) versus leverage values (h) (Gramatica, 2007) were applied for exhibiting the ADs of the developed MLR and SVM models.

2. Results and discussion

2.1. *Development and evaluation for the MLR and SVM models*

2.1.1. *MLR model*

The optimum MLR model is,

$$L = 16.309 + 0.025S_A - 14.651Mi + 0.064SCBO - 0.037nH + 0.561nCIC + 0.681Hy \quad (3)$$

$n_t = 1018$, $R^2 = 0.958$, $RMSE_t = 0.620$,
$F = 3819.850$, $p < 0.001$, $Q^2_{LOO} = 0.957$,
$n_v = 254$, $Q^2_v = 0.961$, $RMSE_v = 0.604$

where, n_t and n_v denote the number for the organic compounds in the training and validation data sets, respectively. According to the criteria $R^2 > 0.70$ and $Q^2 > 0.60$ (Chirico and Gramatica, 2011), we can know that the developed MLR model has good goodness-of-fit, robustness and predictive ability. In recent years, the statistical parameters, i.e., Q^2_{F1}, Q^2_{F2} and Q^2_{F3}, have been proposed for evaluating the model's external predictivity (Gramatica, 2020). Herein, we further calculated the Q^2_{F1}, Q^2_{F2} and Q^2_{F3} for the developed MLR model, and the results ($Q^2_{F1} = 0.961$, $Q^2_{F2} = 0.961$ and $Q^2_{F3} = 0.960$) also showed that the MLR model has satisfactory external predictivity. The definition, standardized coefficients, *t* and *p* values for the predictor variables in the MLR model are shown in Table 1. As displayed in Fig. 1, the predicted *L* values from the MLR model are consistent with the experimental ones.

2.1.2. *SVM model*

SVM analysis was carried out for developing a nonlinear model to estimate the *L* values, by using the dependent variable and predictive variables which are same with those utilized for the development of the MLR model. Two important parameters, i.e., the tuning parameter *C* and regularization parameter γ, were searched for finding the optimal combination of *C* and γ, so that the developed SVM model can perform the best predictive ability. As showned in Fig. 2, the optimal parameters *C* and γ are in the deep red zone. In terms of the

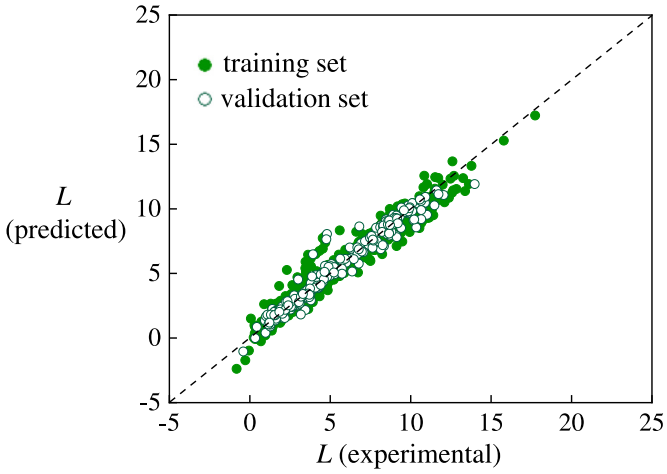

Fig. 1 – Predicted *L* values from the MLR model versus the experimental ones.

Table 2 – Assessment for different prediction tools of *L* values.

Model	k	n_t	R^2	Q^2_{Loo}	$RMSE_t$	n_v	Q^2_v	$RMSE_v$
CI(1) (Stenzel et al., 2012)	3	387			1.55			
CI(2) (Stenzel et al., 2012)	3	387			1.69			
SPARC (Stenzel et al., 2012)		365			1.28			
COSMOthermX (Stenzel et al., 2012)		374			0.94			
ABSOLV (Stenzel et al., 2012)		387			0.99			
IFS (Brown, 2014)	68	610	0.988		0.286	1219		0.300
MLR (This study)	6	1018	0.958	0.957	0.620	254	0.961	0.604
SVM (This study)	6	1018	0.985	0.973	0.366	254	0.980	0.436

*CI represents connectivity indices; IFS denotes iterative fragment selection; MLR is multiple linear regression; SVM is support vector machine; k represents the number of descriptors; n_t and n_v denote the number of compounds from the training and validation sets.

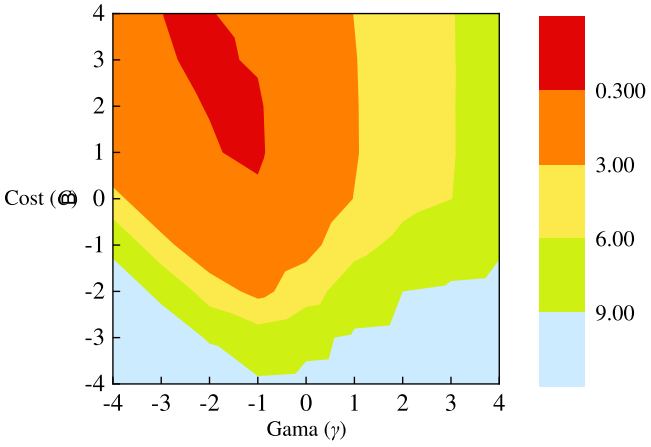

Fig. 2 – Contour plot obtained by searching the optimal combination of *C* and *γ* (logarithmic scale) for the SVM model.

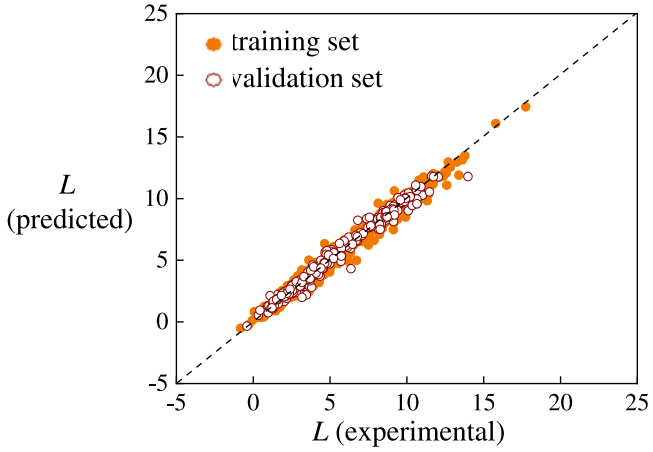

Fig. 3 – Plot of predicted *L* values from the SVM model versus experimental ones.

best SVM model, $C = 10$ and $\gamma = 0.1$ were utilized. The evaluation parameter values for the established SVM model are listed in Table 2, implying that the SVM model performs well for the goodness-of-fit, robustness and predictive ability. Besides, the values, 0.980 (Q^2_{F1}), 0.980 (Q^2_{F2}) and 0.979 (Q^2_{F3}) also implied that the SVM model has good external predictivity. Fig. 3 shows that the predicted *L* values from the SVM model agree well with the experimental ones.

2.2. *Applicability domains for the MLR and SVM models*

The applicability domains of the developed MLR and SVM models were characterized with Williams plots in Fig. 4. It can

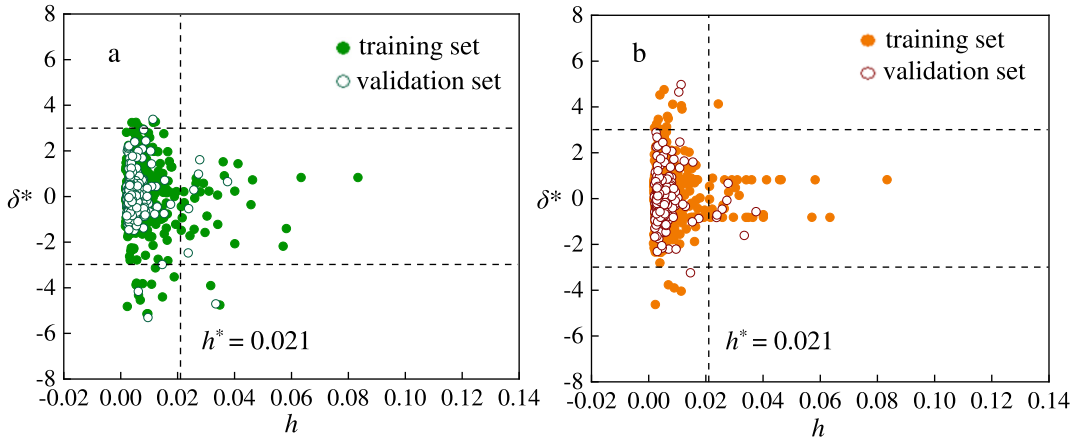

Fig. 4 – Williams plots of standardized residuals (δ*) versus leverage values (h) for (a) MLR and (b) SVM models (h* is the warning leverage value).

be found that more than 97% for the compounds from the total data set are located in the applicability domains. There are 28 organic compounds in the training set for the MLR model and 29 compounds in the training set for the SVM model having $|\delta^*| < 3$ and $h > h^*$ ($h^* = 0.021$), implying that these compounds from the training set are greatly influential on the developed models. In addition, six organic compounds in the validation set for the MLR model and seven organic compounds in the validation set for the SVM model were also found with $|\delta^*| < 3$ and $h > h^*$ ($h^* = 0.021$). It indicates that these compounds from the validation set are structurally distant from the organic compounds in the training set, and their *L* values from prediction using the developed models are extrapolation values.

In terms of the MLR model (Fig. 4**a**), 19 compounds in the training set and 4 compounds in the validation set are identified as outliers for their $|\delta^*|$ values higher than 3. While for the SVM model (Fig. 4**b**), there are 18 outlier compounds in the training set and 3 outlier compounds in the validation set. More details for the outliers can be found in **Table S3**. It should be noted that the ADs for one prediction model depends on the compounds used for the model development. The developed SVM model utilized the same compounds with those applied for the establishment of the MLR model. Therefore, the ADs for the SVM model is the same with that for MLR model, and it has been extended to include more kinds of oraganosilicon compounds (e.g., silane, silazane, silicon sulfide, siloxane and chlorinated siloxane) than previous prediction methods. The developed MLR and SVM models can be used for estimating *L* values of various organic compounds with the following chemical formula: RH, $R_2C{=}CR_2$', RC≡CR', RC_6H_5, $RCH_2C_6H_5$, RX (X = F, Cl, Br, I), ROH, RCOR', RCHO, RCOOH, ROR', RCOOR', RNO_2, $RCONH_2$, RNH_2, R_2NH, R_3N, RN_2R', RCN, RC_5H_4N, R_3P, $ROP(=O)(OR')_2$, RSR', RSSR', RSH, SiR_4, $R_3SiOSiR_3$', $R_3SiNHSiR_3$' and $R_3SiSSiR_3$'.

2.3. *Mechanistic interpretation*

As shown in Eq. (3), six molecular structural descriptors were selected for the MLR model. Among these descriptors, S_A has the most significant effects on *L* values for the organic compounds according to its standardized coefficient being listed in Table 1. S_A represents the conductor-like screening model (COSMO) area, which is calculated at the van der Waals' distance. The coefficient for S_A is positive, indicating that it is positively correlated with the *L* values of organic compounds. The *L* value for the organic compound increases with the increase of its S_A value. The reason may be that one compound with a larger COSMO area can have more vdW contacts with hexadecane, which can promote its distribution in hexadecane phase. Besides, the descriptor *Mi* (De and Roy, 2018; Chavan et al., 2014), which represents the mean first ionization potential (scaled on carbon atom), has a negative coefficient in the MLR model. It indicates that the compound with a lower *Mi* value will have a higher *L* value. In general, one compound with lower mean first ionization potential has larger molecular size (Sæthre et al., 2011). The increase of molecular size will increase the dispersion interaction between the organic compound and hexadecane (Schüürmann et al., 2006; Torres et al., 2018), thereby increasing its concentration in the hexadecane phase. Hence, the descriptors, i.e., S_A and *Mi*, depicts the important role of van der Waals interactions in the partition processes for the compounds between hexadecane and air jointly.

The descriptor *SCBO* (Azimi et al., 2012), denoting the sum of conventional bond orders (H-depleted), has a positive contribution to the *L* value, implying that the *L* value for one compound increases with the increase of its *SCBO* value. In addition, the ring descriptor, *nCIC* (Nantasenamat, 2013), also has a positive coefficient. *nCIC* represents the number of rings (cyclomatic number). The descriptor *nH* is the number of hydrogen atoms. The negative coefficient for *nH* shows that the *nH* value for the organic compound has a negative contribution to the *L* value. These three descriptors, SCBO, *n*CIC and *nH* describe the effects of the structural complexity for the compound on the *L* value.

Note that the hydrophilic factor *Hy* (Liu et al., 2017) is positively correlated with the *L* values for the organic compounds. Thus, the increase of hydrophilic factor can result in the increase of the *L* value slightly.

As discussed above, the *L* value is greatly influenced by the van der Waals interactions between the compound and the

hexadecane. Besides, the structural complexity for the compound (conventional bond orders, the number of rings and the number of hydrogen atoms) and hydrophilic factor also have effects on the *L* value.

2.4. *Comparison for the prediction tools*

Up to date, several prediction tools can be used to estimate the *L* values of organic compounds. The MLR and SVM models developed in current study are compared with the previous prediction tools in **Table 2**. Among the prediction tools, i.e., connectivity indices (CI) (1) (Stenzel et al., 2012), CI (2) (Stenzel et al., 2012), SPARC (Stenzel et al., 2012), COSMOthermX (Stenzel et al., 2012) and ABSOLV (Stenzel et al., 2012), the present MLR and SVM models perform best. In comparison with the previous IFS (Brown, 2014) model, the predictive ability for the SVM model with less number of descriptors is comparable with that for the IFS model, while the root mean square errors for MLR model is slightly larger than those for SVM and the previous IFS models. Besides, the current MLR and SVM models have a broader ADs than the IFS model in terms of the organosilicon compounds.

3. Conclusions

The MLR and SVM models for predicting *L* values of organic compounds were developed and evaluated according to the OECD guidelines. The established QSAR models have satisfactory goodness of fit, robustness and predictive ability. The SVM algorithm was first utilized for predicting *L* values, and the established SVM model performs better to some extent than the MLR model. The applicability domain of the developed models covers different classes of compounds including emerging environmental pollutants, i.e., oraganosilicon compounds (e.g., silane, silazane, silicon sulfide, siloxane and chlorinated siloxane). The gap of mechanistic interpretation has been filled, and the results show that the van der Waals interactions paly dominant roles in the partition processes between the hexadecane and air for organic compounds. The developed models in this study can serve as novel *in silico* tools to obtain the *L* values of organic compounds highly efficiently.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 21936005).

Appendix A Supplementary data

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jes.2021.10.033.

REFERENCES

Abraham, M.H., Grellier, P.L., Mcgill, R.A., 1987. Determination of olive oil-gas and hexadecane-gas partition coefficients, and calculation of the corresponding olive oil-water and hexadecane-water partition coefficients. Perkin Trans. 2, 797–803.

Alton, M.W., Browne, E.C., 2020. Atmospheric chemistry of volatile methyl siloxanes: kinetics and products of oxidation by OH radicals and Cl atoms. Environ. Sci. Technol. 54, 5992–5999.

Azimi, G., Afiuni-Zadeh, S., Karami, A., 2012. A QSAR study for modeling of thyroid receptors β1 selective ligands by application of adaptive neuro-fuzzy inference system and radial basis function. J. Chemometr. 26, 135–142.

Bronner, G., Fenner, K., Goss, K., 2010. Hexadecane/air partitioning coefficients of multifunctional compounds: Experimental data and modeling. Fluid Phase Equilibr 299, 207–215.

Brown, T.N., 2014. Predicting hexadecane-air equilibrium partition coefficients (*L*) using a group contribution approach constructed from high quality data. SAR QSAR Environ. Res. 25, 51–71.

Bzdek, B.R., Horan, A.J., Pennington, M.R., Janechek, N.J., Baek, J., Stanier, C.O., et al., 2014. Silicon is a frequent component of atmospheric nanoparticles. Environ. Sci. Technol. 48, 11137–11145.

Chavan, S., Nicholls, I., Karlsson, B., Rosengren, A., Ballabio, D., Consonni, V., et al., 2014. Towards global QSAR model building for acute toxicity: Munro database case study. Int. J. Mol. Sci. 15, 18162–18174.

Cheng, Z., Chen, Q., Cervantes, S., Tang, Q., Gao, X., Tan, Y., et al., 2020. Two-dimensional and three-dimensional quantitative structure-activity relationship models for the degradation of organophosphate flame retardants during supercritical water oxidation. J. Hazard. Mater. 394, 121811.

Chirico, N., Gramatica, P., 2011. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. 51, 2320–2335.

Cho, C., Stolte, S., Yun, Y., 2018. Validation and updating of QSAR models for partitioning coefficients of ionic liquids in octanol-water and development of a new LFER model. Sci. Total Environ. 633, 920–928.

De, P., Roy, K., 2018. Greener chemicals for the future: QSAR modelling of the PBT index using ETA descriptors. SAR QSAR Environ. Res. 29, 319–337.

Endo, S., Goss, K., 2014. Applications of polyparameter linear free energy relationships in environmental chemistry. Environ. Sci. Technol. 48, 12477–12491.

Endo, S., Goss, K., 2014. Predicting partition coefficients of polyfluorinated and organosilicon compounds using polyparameter linear free energy relationships (PP-LFERs). Environ. Sci. Technol. 48, 2776–2784.

Fairbrother, A., Woodburn, K.B., 2016. Assessing the aquatic risks of the cyclic volatile methyl siloxane D4. Environ. Sci. Tech. Let. 3, 359–363.

Goudarzi, N., Goodarzi, M., 2008. Prediction of the logarithmic of partition coefficients (log P) of some organic compounds by least square-support vector machine (LS-SVM). Mol. Phys. 106, 2525–2535.

Grabitz, E., Olsson, O., Amsel, A., Rummel, B., Mitzel, N.W., Kümmerer, K., 2020. Abiotic and biotic degradation of five aromatic organosilicon compounds in aqueous media—structure degradability relationships. J. Hazard. Mater. 392, 122429.

Gramatica, P., 2007. Principles of QSAR models validation: Internal and external. QSAR Combust. Sci. 26, 694–701.

Gramatica, P., 2020. Principles of QSAR modeling: Comments and suggestions from personal experience. Int. J. Quant. Struct. Prop. Relat. 5, 1–37.

Krogseth, I.S., Whelan, M.J., Christensen, G.N., Breivik, K., Evenset, A., Warner, N.A., 2016. Understanding of cyclic

volatile methyl siloxane fate in a high latitude lake is constrained by uncertainty in organic carbon–water partitioning. Environ. Sci. Technol. 51, 401–409.
Lan, Z., Zhang, B., Huang, X., Zhu, Q., Yuan, J., Zeng, L., et al., 2018. Source apportionment of $PM_{2.5}$ light extinction in an urban atmosphere in China. J. Environ. Sci. 63, 277–284.
Li, T., Huang, Y., Wei, G., Zhang, Y., Zhao, Y., Crittenden, J.C., et al., 2020. Quantitative structure-activity relationship models for predicting singlet oxygen reaction rate constants of dissociating organic compounds. Sci. Total Environ. 735, 139498.
Liu, H., Wei, M., Yang, X., Yin, C., He, X., 2017. Development of TLSER model and QSAR model for predicting partition coefficients of hydrophobic organic chemicals between low density polyethylene film and water. Sci. Total Environ. 574, 1371–1378.
Liu, H., Yang, X., Lu, R., 2016. Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin. Chemosphere 156, 1–7.
Galán-Madruga, D., García-Cambero, J.P., 2022. An optimized approach for estimating benzene in ambient air within an air quality monitoring network. J. Environ. Sci. 111, 164–174.
MOPAC, 2016. http://openmopac.net/MOPAC2016.html
Nantasenamat, C., Worachartcheewan, A., Prachayasittikul, S., Isarankura-Na-Ayudhya, C., Prachayasittikul, V., 2013. QSAR modeling of aromatase inhibitory activity of 1-substituted 1,2,3-triazole analogs of letrozole. Eur. J. Med. Chem. 69, 99–114.
OECD, 2007. Guidance document on the validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] models.
Panagopoulos, D., Jahnke, A., Kierkegaard, A., MacLeod, M., 2017. Temperature dependence of the organic carbon/water partition ratios (K_{OC}) of volatile methylsiloxanes. Environ. Sci. Tech. Let. 4, 240–245.
Poole, C.F., Atapattu, S.N., Poole, S.K., Bell, A.K., 2009. Determination of solute descriptors by chromatographic methods. Anal. Chim. Acta 652, 32–53.
Rücker, C., Kümmerer, K., 2015. Environmental chemistry of organosiloxanes. Chem. Rev. 115, 466–524.
Sæthre, L.J., Børve, K.J., Thomas, T.D., 2011. Chemical shifts of carbon 1s ionization energies. J. Electron. Spectros. Relat. Phenomena. 183, 2–9.
Schnur, D.M., Grieshaber, M.V., Bowen, J.P., 1991. Development of an internal searching algorithm for parameterization of the MM2/MM3 force fields. J. Comput. Chem. 12, 844–849.
Schüürmann, G., Ebert, R.U., Kühne, R., 2006. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. Environ. Sci. Tech. 40, 7005–7011.
Stenzel, A., Endo, S., Goss, K., 2012. Measurements and predictions of hexadecane/air partition coefficients for 387 environmentally relevant compounds. J. Chromatogr. A 1220, 132–142.
Stenzel, A., Goss, K., Endo, S., 2013. Determination of polyparameter linear free energy relationship (pp-LFER) substance descriptors for established and alternative flame retardants. Environ. Sci. Technol. 47, 1399–1406.
Stewart, J.J.P., 2013. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. J. Mol. Model. 19, 1–32.
Talete srl, 2012. Dragon (Software for Molecular Descriptor Calculation) Version 6.0 <http://www.talete.mi.it/>.
Tang, X., Misztal, P.K., Nazaroff, W.W., Goldstein, A.H., 2015. Siloxanes are the most abundant volatile organic compound emitted from engineering students in a classroom. Environ. Sci. Tech. Lett. 2, 303–307.
Torres, A., Suárez, J.A., Remesal, E.R., Márquez, A.M., Sanz, J.F., Cañibano, C.R., 2017. Adsorption of prototypical asphaltenes on silica: First-principles DFT simulations including dispersion corrections. J. Phys. Chem. B 122, 618–624.
Wang, B., Chen, J.W., Li, X.H., Wang, Y.N., Chen, L., Zhu, M., et al., 2009. Estimation of soil organic carbon normalized sorption coefficient (K_{OC}) using Least Squares-Support Vector Machine. QSAR Combust. Sci. 28, 561–567.
Wang, Y., Chen, J.W., Wei, X.X., Hernandez Maldonado, A.J., Chen, Z.F, 2017. Unveiling adsorption mechanisms of organic pollutants onto carbon nanomaterials by density functional theory computations and linear free energy relationship modeling. Environ. Sci. Technol. 51, 11820–11828.
Wang, Y., Chen, J., Tang, W., Xia, D., Liang, Y., Li, X., 2019. Modeling adsorption of organic pollutants onto single-walled carbon nanotubes with theoretical molecular descriptors using MLR and SVM algorithms. Chemosphere 214, 79–84.
Wu, Y., Johnston, M.V., 2017. Aerosol formation from OH oxidation of the volatile cyclic methyl siloxane (cVMS) decamethylcyclopentasiloxane. Environ. Sci. Technol. 51, 4445–4451.
Yaman, B., Dumanoglu, Y., Odabasi, M., 2020. Measurement and modeling the phase partitioning of organophosphate esters using their temperature-dependent octanol–air partition coefficients and vapor pressures. Environ. Sci. Technol. 54, 8133–8143.
Zhang, G., Zhang, S., 2020. Quantitative structure-activity relationship in the photodegradation of azo dyes. J. Environ. Sci. 90, 41–50.
Zhao, S., Jones, K.C., Sweetman, A.J., 2018. Can poly-parameter linear-free energy relationships (pp-LFERs) improve modelling bioaccumulation in fish? Chemosphere 191, 235–244.
Zhao, Y., Choi, J., Bediako, J.K., Song, M., Lin, S., Cho, C., et al., 2018. Adsorptive interaction of cationic pharmaceuticals on activated charcoal: Experimental determination and QSAR modelling. J. Hazard. Mater. 360, 529–535.
Zhi, L., Sun, H., Xu, L., Cai, Y., 2021. Distribution and elimination of trifluoropropylmethylsiloxane oligomers in both biosolid-amended soils and earthworms. Environ. Sci. Technol. 55, 985–993.
Zhu, T., Chen, W., Cheng, H., Wang, Y., Singh, R.P., 2019. Prediction of polydimethylsiloxane-water partition coefficients based on the pp-LFER and QSAR models. Ecotoxicol. Environ. Saf. 182, 109374.
Zhu, T., Jiang, Y., Cheng, H., Singh, R.P., Yan, B., 2020. Development of pp-LFER and QSPR models for predicting the diffusion coefficients of hydrophobic organic compounds in LDPE. Ecotoxicol. Environ. Saf. 190, 110179.