ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Thermal Conductivity Estimation of Diverse Liquid Aliphatic Oxygen-Containing Organic Compounds Using the Quantitative Structure−Property Relationship Method

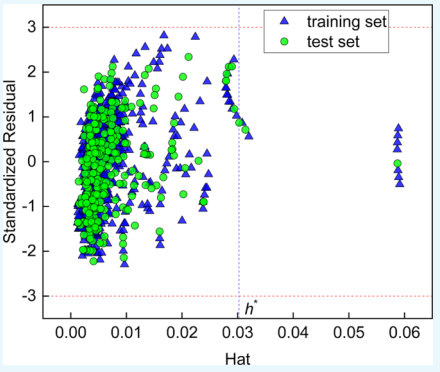
Haixia Lu, Wanqiang Liu,* Fan Yang, Hu Zhou, Fengping Liu, Hua Yuan, Guanfan Chen, and Yinchun Jiao*

Cite This: https://dx.doi.org/10.1021/acsomega.9b04190

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Thermal conductivity is an essential thermodynamic data in chemical engineering applications. Liquid aliphatic oxygen-containing organic compounds are important organic intermediates and raw materials. As a result, estimating thermal conductivity of liquid aliphatic oxygen-containing organic compounds is of significance in industry production. In this study, the genetic function approximation method was applied to screen descriptors and develop a 6-descriptor linear quantitative structure−property relationship model. The entire data set of these compounds covering 1064 thermal conductivity values was divided into 694-member training set, 298-member test set, and 72-member prediction set. The average absolute relative deviation of the training set, test set, and prediction set were 4.14, 4.41, and 4.16%, respectively. Model validation and Y-randomization test proved that the developed model has goodness-of-fit, predictive power, and robustness. In addition, the applicability domain of the developed model was visualized by the Williams plot. This study can provide a convenient method to estimate the thermal conductivity for researchers in chemical engineering production.



1. INTRODUCTION

Thermal conductivity acts as an important physical property that can reflect the ability of heat transfer of a substance. In chemical industry, petroleum industry, and energy engineering, thermal conductivity is one of the essential basic thermodynamic data for heat transfer design.

From the middle of the 18th century, many researchers have made a lot of exploration on thermal conductivity measuring methods.[1−6] However, these methods are time-consuming, costly, and technique limited, and sometimes, there is a certain error in the measured thermal conductivity value. The main reason is that the heat loss caused by convection and radiation cannot be controlled during the measurement. As a result, developing a theoretical method for estimating thermal conductivity is of theoretical significance.

Some estimation methods have been proposed to predict thermal conductivity of aliphatic oxygen-containing organic compounds, which were based on the liquid molecular motion theory model and knowledge of the liquid heat conduction mechanism. However, it should be aware that these methods need be supported by empirical theory and experimental data. As a result, these methods are empirical with big errors, generally between 2 and 10%, as shown in Table 1.

The quantitative structure−property/activity relationship (QSPR/QSAR) was defined as a mathematical relationship connecting the chemical structure with compound properties in a quantitative manner.[10] It can reliably predict the physicochemical, biological, and pharmacological properties of compounds from the molecular structures of compounds.

Table 1. Results of Aliphatic Oxygen-Containing Organic Compound Calculations

	Rodenbush et al.		Nagvekar and Daubert		Baroncini et al.	
classes	no. of data	MRD[a] /%	no. of data	MRD/%	no. of data	MRD/%
alcohols	267	2.7	634	6.3	592	7.7
acids	68	3.2	787	6.5	236	4.1
esters	92	2.8	243	9.7	197	5.7
ketones	48	3.2	68	6.3	72	8.3
ethers	18	3.1	75	7.3	60	3.2
aldehydes	28	2.1	43	8.3	44	6.4
refs	7		8		9	

[a]MRD stands for mean relative deviation.

Received: December 7, 2019
Accepted: March 31, 2020

Table 2. Retrospect of Works for Thermal Conductivity Prediction

methods	authors	compound class	parameters	N	R^2	results	refs
MLR	Gao and Cao	alkanes	4	155	0.9510	s^a = 0.0033	14
MLR	Kauffman and Jurs	organic solvents	9	213	0.953	RMSE[b] = 0.0136	15
GFA[c]	Khajeh and Modarress	alcohols	5	116	0.9438	RMSE = 0.0474	16
MLR	Liu et al.	alcohols	4	139	0.9738	RMSE = 0.0029	17
GFA	Liu et al.	alkyl halides	6	410	0.9745	RMSE = 0.0035	18

[a]s stands for the standard deviation. [b]RMSE stands for root-mean-square error. [c]GFA stands for genetic function approximation.

Table 3. Regression Statistics of Parameters Involved in the QSPR Model

parameters	type	coefficients	standardized coefficients	t	p	VIF
SM2_B(s)	2D matrix-based descriptors	0.0711	1.0443	47.33	0.000	3.88
SIC0	Information indices	0.2960	0.9626	58.79	0.000	2.14
IC1	Information indices	−0.0511	−0.5876	−36.67	0.000	2.04
T	Temperature	−0.0002	−0.5853	−49.16	0.000	1.16
Eta_F	ETA indices	−0.0127	−0.5579	−27.77	0.000	3.22

This method became prevalent when it is not possible to obtain accurate values in the experiment limited to economy, time, or technology.[11−13] As for thermal conductivity, few works are reported with QSPR. Results of some works are listed in Table 2.

Aliphatic alcohols, ethers, aldehydes, ketones, acids, and esters are diverse important chemical products and intermediates, which are widely used as industrial raw materials, lubricants, solvents, medicines, daily necessities, food additives, and so forth. Thermal conductivity is an important thermodynamic property for these organic compounds in industrial application. Therefore, how to accurately predict the thermal conductivity of these diverse aliphatic oxygen-containing organic compounds has important significance for engineering application.

Structures of these compounds have some certain similarities, which satisfies the basic conditions for the QSPR study.[19−21] This work intends to (1) collect thermal conductivity data of aliphatic alcohols, ethers, aldehydes, ketones, acids, and esters; (2) extract molecular descriptors from molecules; (3) develop a QSPR model and validate the model; and (4) predict the thermal conductivity of diverse aliphatic oxygen-containing organic compounds using the model.

2. RESULTS AND DISCUSSION

Using the GFA (100 initial population size, 500 the maximum generations, 10% the mutation probability, and the smooth parameter of LOF $\alpha = 0.5$), a 6-descriptor linear QSPR model was developed:

$$\lambda = -0.0793 + 0.0711 \times SM2\_B(s) + 0.296 \times SIC0 - 0.0511 \times IC1 - 0.0002 \times T - 0.0127 \times Eta\_F - 0.0587 \times MATS2m$$

$$n_{tr} = 694,\ R_{tr}^2 = 0.9138,\ \mathrm{RMSE} = 0.0067,\ F = 1214.35,\ s = 0.0067,\ \mathrm{AARD\ \%} = 4.14\%,$$

$$Q_{CV}^2 = 0.9118,\ \mathrm{RMSE_{CV}} = 0.0072,\ n_{test} = 298,\ R_{test}^2 = 0.8922,\ \mathrm{AARD\ \%} = 4.41\%,$$

$$\mathrm{RMSE_{test}} = 0.0071,\ Q_{ext\text{-}F1}^2 = 0.8917,\ Q_{ect\text{-}F2}^2 = 0.8916,\ Q_{ect\text{-}F3}^2 = 0.9065,\ r_m^2 = 0.8466,$$

$$\mathrm{CCC} = 0.9437,\ n_{pred} = 72,\ R_{pred}^2 = 0.8816,\ \mathrm{RMSE_{pred}} = 0.0064,\ \mathrm{AARD\ \%} = 4.16\%$$

where *SM2_B(s)* is the spectral moment of order 2 from the Burden matrix weighted by I-state, *SIC0* is the structural information content index (neighborhood symmetry of 0-order), *IC1* is information content index (neighborhood symmetry of 1-order), *T* is the temperature, *Eta_F* is the eta functionality index, and *MATS2m* is Moran autocorrelation of lag 2 weighted by mass. The regression statistics of these involved parameters are given in Table 3.

2.1. Model Validation. As can be seen from the absolute values of standardized coefficients in Table 3, the most significant influence of descriptors on thermal conductivity in turn is *SM2_B(s)*, *SIC0*, *IC1*, *T*, *Eta_F*, and *MATS2m*. The variance inflation factor (VIF) values of these descriptors lower than 10 together with the intercorrelation matrix (Table 4) manifest the absence of multicollinearities among the descriptors.[22] In addition, the confidence level p values of the descriptors all far less than 0.0001 and $F = 1214.35$ prove the statistical significances of the descriptors and the developed model, respectively.

Table 4. Correlation Matrix of the Involved Descriptors

	SM2_B(s)	*SIC0*	*IC1*	*T*	*Eta_F*	*MATS2m*
SM2_B(s)	1.000					
SIC0	−0.366	1.000				
IC1	0.114	0.575	1.000			
T	0.206	−0.277	−0.250	1.000		
Eta_F	0.810	−0.231	0.093	0.106	1.000	
MATS2m	0.119	0.193	0.349	−0.168	0.239	1.000

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

The squared correlation coefficient of training set R_{tr}^2 = 0.9138 and root-mean-square error (RMSE) = 0.0067 indicated that the developed model was acceptable with goodness-of-fit. Using this model to predict the test set and the prediction set, the squared correlation coefficients were 0.8922 and 0.8816, which suggested that the developed model had predictive ability. The correlation between experimental thermal conductivity and calculated values was plotted in Figure 1 (detailed information are given in Supporting Information, Table S2) which shows that the experiment thermal conductivity were highly in agreement with the calculated values.

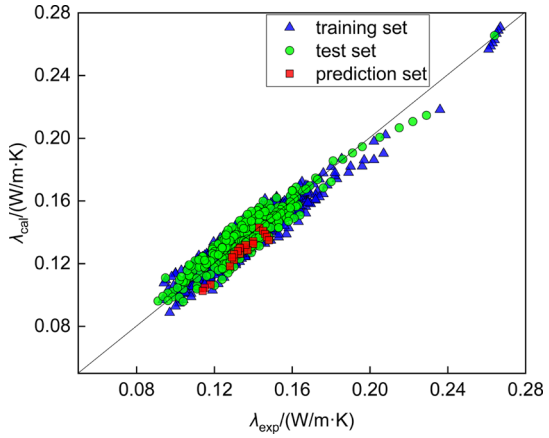


Figure 1. Experimental thermal conductivity vs calculated values.

In addition, predicted thermal conductivity values of different chemical classes versus temperature are shown in Figure 2. It can be found that the tendency of predicted

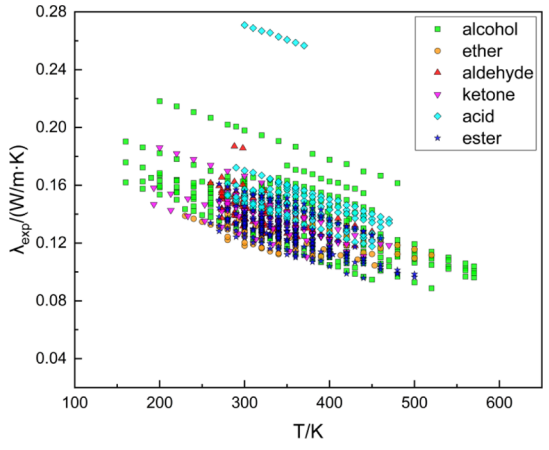


Figure 2. Predicted thermal conductivity values of different chemical classes vs temperature.

thermal conductivity variation with temperature is basically the same with the experimental thermal conductivity variation with temperature. As for the leave-one-out cross validation, Q_{CV}^2 = 0.9118 and $RMSE_{CV}$ = 0.0072 manifested the stability of the developed model.

The Y-randomization test was carried out to testify the absence of chance correlation in this study. Random shuffles of the thermal conductivity were conducted ten times, and the results are shown in Table 5. The low R_{rand}^2 and Q_{rand}^2 values lower than the original model suggests that good stability and

Table 5. R^2 and Q^2 Values after Several Y-Randomization Tests

iteration	R^2	Q^2
1	0.01	0.01
2	0.01	0.01
3	0.01	0.01
4	0.01	0.02
5	0.01	0.02
6	0.01	0.01
7	0.01	0.01
8	0.01	0.01
9	0.02	0.00
10	0.00	0.02

predictive performance of the model are not based on the chance correlation.[23] In addition, $^{C}R_p^2$ equals to 0.9094, which is close to the value of R_{tr}^2, manifesting the robustness of the developed model.

For the training set, test set, and prediction set, the average absolute relative deviation (AARD) are 4.14, 4.41, and 4.16%, respectively, resulting in an AARD of 4.23% for the whole dataset, and for the entire data set, the squared correlation coefficient and the RMSE are 0.9038 and 0.0067, respectively. The calculated statistical metrics recommended by Tropsha et al.[24,25] for the test set are as follows, which testify the predictive ability of the developed model

$$R_{CV,ext}^2 = 0.8878 > 0.5$$

$$r^2 = 0.8753 > 0.6$$

$$\frac{(r^2 - r_0^2)}{r^2} = (0.8753 - 0.8746)/0.8753 < 0.1$$

or

$$\frac{(r^2 - r_0'^2)}{r^2} = (0.8753 - 0.8644)/0.8753 < 0.1$$

$$0.85 \le k = 0.9900 \le 1.15 \text{ or } 0.85 \le k' = 1.0076 \le 1.15$$

According to the ref 26, both experimental and calculated endpoint values were taken in the log scale, and MAE values were calculated from long transformed values. In this study, the MAE, training set range, and σ of the multiple linear regression (MLR) model are 0.0174, 7.820 and 0.0113, respectively. As a result, it can be found that two conditions are satisfying the "GOOD model" criteria as the following

$$\text{MAE} \le 0.1 \times \text{training set range and MAE} + 3 \times \sigma \le 0.2 \times \text{training set range}$$

By using the "Prediction Reliability Indicator" tool, it can be found that none of the external compounds is "bad". Detailed information can be seen in Supporting Information (Table S2).

Based on these results, it can be concluded that the GFA–MLR model has high accuracy, robustness, and good predictive ability.

2.2. Applicability Domain. As shown in Figure 3, the blue vertical line stands for the threshold value h^* and the two horizontal red lines stand for ±3 standard deviation units. The AD is located in the region of $0 \le h \le 0.0303$ and $-3 \le R \le 3$. It can be found that a majority of data points fell within the

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

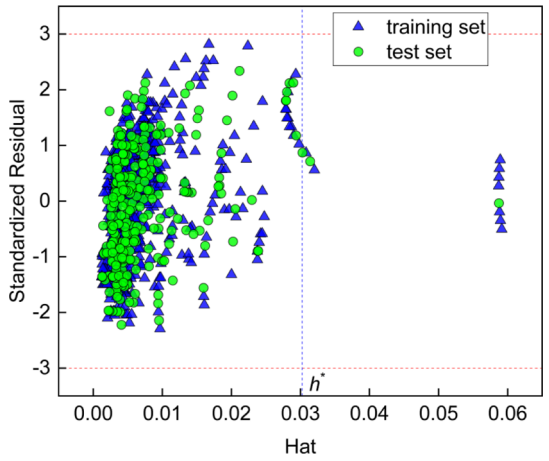


Figure 3. AD of the developed model.

AD, which further indicates the robustness and validity of the developed model.

Meanwhile, data points (thermal conductivity of formic acid at 300, 310, 320, 330, 340, 350, 360, and 370 K, and thermal conductivity of isoheptanol at 230, 240, 250, and 260 K) are located in the region of $h > 0.0303$ and $-3 \leq R \leq 3$, which are so-called as X outliers (good influence points). These compounds were well predicted by the developed model, which can stabilize it and make it more precise.[27] There is no Y outlier, which proves the reliability of the developed model.

2.3. Descriptor Interpretation. As discussed previously, the most significant influence of descriptors on thermal conductivity is *SM2_B(s)*, *SIC0*, *IC1*, *T*, *Eta_F*, and *MATS2m*, in turn.

SM2_B(s)[28] is one of the 2D matrix-based descriptors. It can be interpreted as a spectral moment of order 2 from the Burden matrix weighted by I-state, which was proposed to evaluate the spectral moment from the Burden matrix and is expressed as

$$SM2\_B(s) = \ln\left(1 + \sum_{i=2}^{n_{SK}} \lambda_i{}^2\right)$$

where n_{SK} represents the number of graph vertices and λ_i represents the eigenvalues of the Burden matrix. The Burden matrix is the augmented adjacency matrix derived from the H-depleted molecular graph. The adjacency matrix can describe the relationship of each vertex. Thus, *SM2_B(s)* can be used to investigate the effect of different molecule group interaction on thermal conductivity. The positive correlation coefficient for *SM2_B(s)* indicates that the stronger the molecule group interaction is, the larger the thermal conductivity is.

IC1[28,29] is the information content index (neighborhood symmetry of 1-order), one of the information indices. The general form of the neighborhood information content (*ICk*) is calculated as follows

$$IC1 = -\sum_{g=1}^{G} \frac{A_g}{n_{AT}} \cdot \log_2\left(\frac{A_g}{n_{AT}}\right)$$

where g runs over the equivalence classes, A_g represents the cardinality of the gth equivalence class, and n_{AT} represents the number of molecule atoms. *IC1* can be used to evaluate and describe the effect of molecules size, shape, and branching information on thermal conductivity. From the negative correlation coefficient, it can be found that the more the molecular branches are, the shorter the main chain is, and the lower the thermal conductivity is.

SIC0 (structural information content index (neighborhood symmetry of 0-order)) is a type of information indices. It is calculated as the information content *IC0* normalized form to delete the graph size influence

$$SIC0 = \frac{ICk}{\log_2 n_{AT}}$$

It is a topological descriptor that can be used to measure the degree of atom diversity in a molecule and describe its shape. As for aliphatic oxygen-containing organic compounds, deleting the influence of graph size, it can be regarded as a measurement of number of oxygen atoms. The greater the number of oxygen atoms is, the more the charge transfer is, and the larger the thermal conductivity is. Thus, the *SIC0* is positively related with thermal conductivity.

In addition, T (temperature) also has a remarkable effect on thermal conductivity. The positive value of coefficient means that T is negatively related to thermal conductivity. As the temperature rises, the liquid molecules move more chaotic, and the molecular directional movement from the high-energy region to the low-energy region weakens, which leads to the decrease of heat conduction.[30]

Eta_F is the abbreviation of eta functionality index, a type of ETA indices, which was proposed to evaluate the molecule functionality, here quantifying the presence of heteroatoms and multiple bonds.[28,31] For aliphatic oxygen-containing organic compounds, the larger the *Eta_F* value is, the more molecular branches are, and the shorter the main chain is. Accordingly, heat transfer along the main chain is reduced, and so is thermal conductivity. For example, the *Eta_F* values of n-butanol, isobutyl alcohol, and tertbutyl alcohol are 0.669, 0.722, and 0.772, respectively, and at 340 K, the thermal conductivity values of them are 0.143, 0.128, and 0.11.

The Moran autocorrelation of lag 2 weighted by mass *MATS2m* is one of the 2D autocorrelations. The general form of Moran autocorrelation is represented by using Moran coefficient to the molecular graph

Table 6. Comparisons with Previous Studies

author	methods	no	compounds class	N	R^2	results
Gao and Cao[14]	MLR	155	alkanes	4	0.9510	s = 0.0033
Kauffman and Jurs[15]	MLR	213	organic solvents	9	0.953	RMSEP = 0.0136
Khajeh and Modarress[16]	GFA−MLR	116	alcohols	5	0.9521	RMSEP = 0.0474
Liu et al.	MLR	139	alcohols	4	0.9738	RMSEP = 0.0029
this study	GFA−MLR	1064	aliphatic oxygen-containing organic compounds	6	0.8922	AARD % = 4.41%
						$RMSE_{test}$ = 0.0071

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

$$MATS2m = \frac{\frac{1}{\Delta}\cdot\sum_{i=1}^{n_{AT}}\sum_{j=1}^{n_{AT}}\delta_{ij}\cdot(w_i - \overline{w})\cdot(w_j - \overline{w})}{\frac{1}{n_{AT}}\cdot\sum_{i=1}^{n_{AT}}(w_i - \overline{w})^2}$$

where w_i represents any atomic property, $\overline{w}$ represents the w_i average value on the molecule, δ_{ij} represents the Kronecker delta, and Δ represents the sum of the Kronecker deltas. The Moran coefficient usually takes value in the interval $[-1, +1]$. Positive spatial autocorrelation corresponds to positive coefficient values, whereas negative spatial autocorrelation corresponds to negative coefficient values. The molecules with positive autocorrelation have more branch chains than molecules with linear chain, of which the main chain is shorter and thermal conductivity value reduces. As a result, the *MATS2m* is negatively related with thermal conductivity.

All these six parameters have significant contribution to the thermal conductivity of oxygen-containing organic compounds.

2.4. Comparison between the Previous Studies and the Present Study. Comparisons with previous studies are shown in Table 6. Compared with the previous studies, the present study has been further improved. Under the premise of ensuring high fitting ability and predictive power, the present study has fully validated the developed models and analyzed the applicability domains (ADs).

3. CONCLUSIONS

In the present study, a 6-descriptor linear QSPR model was developed for predicting the thermal conductivity of diverse aliphatic oxygen-containing organic compounds (aliphatic alcohols, ethers, aldehydes, ketones, acids, and esters). The molecular descriptors were calculated based on the optimized structures of aliphatic oxygen-containing organic compounds and were screened using the genetic function approximation (GFA) method. The predictive ability, model validation, and the AD of the developed model indicate that the GFA−MLR model has good predictive reliability and robustness. The GFA−MLR model provides a transparent output, with an AARD of 4.23% for the whole data set. Meanwhile, it can be concluded that molecule group interaction, molecular branches, number of oxygen atoms, temperature, and main chain of molecules have great effects on the thermal conductivity of aliphatic oxygen-containing organic compounds. The present study is of great significance not only by providing a robust model for predicting thermal conductivity of diverse aliphatic oxygen-containing organic compounds but also by shedding light on other thermodynamic data estimation.

4. MATERIALS AND METHODS

4.1. Data set Collection. Thermal conductivity of the substance is mainly affected by the molecular structure and the temperature.[32] Therefore, in this study, at first, 992 thermal conductivity data of liquid aliphatic alcohols, ethers, aldehydes, ketones, acids, and esters at different temperatures were collected from the *Handbook of Thermal Conductivity of Liquids and Gases*.[33] These compounds covered 112 aliphatic oxygen-containing compounds. There are 290 thermal conductivity values of alcohols (29%), 78 thermal conductivity values of ethers (8%), 100 thermal conductivity values of aldehydes (10%), 112 thermal conductivity values of ketones (11%), 140 thermal conductivity values of acids (14%), and 272 thermal conductivity values of esters (28%). Thermal conductivity values of different chemical classes versus temperature are shown in Figure 4.

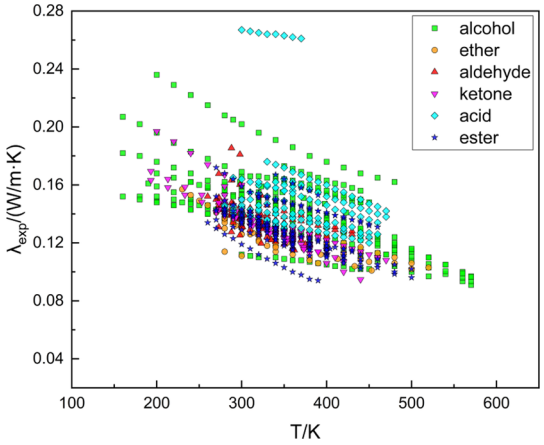


Figure 4. Thermal conductivity values of different chemical classes vs temperature.

In this study, Kennard Stone algorithm was performed to split the data set, which was developed based on the different Euclidean distance among all data. At first, the samples with larger difference were selected into the training set. Then, the remaining samples which are closer to the training data enter the test set. Under the circumstance, all representative samples can be divided into the training set. To some extent, the unevenness of the training set can be avoided.[34] In the end, the entire data set was divided into 694-member training set and 298-member test set with Kennard−Stone algorithm. The QSPR model would be developed based on the training set, and the test set was used to evaluate the performance and robustness of the developed model.[35]

In addition, to further testify the predictive power of the developed model, in our laboratory, 72 experimental thermal conductivity data of 15 compounds were measured as the prediction set. These data were measured using the DRE-2A thermal conductivity instrument with nonequilibrium state transient hot-wire method.[17] The transient hot wire method is a method for measuring the thermal conductivity of liquids and has the characteristics of high speed and high precision. According to the relationship between temperature and time, thermal conductivity was calculated automatically by the instrument. Each sample was measured three times, and the final data is an averaged value. The error of measurement is <3%.

The detailed information of the involved compounds and the whole data set which contains 1064 thermal conductivity data are shown in Supporting Information (Tables S1 and S2).

4.2. Molecular Descriptors. All molecule structures were constructed in GaussView graphical interface software package.[36] Then, the structures were output as the Gaussian input files. After modifying the input files, using the Keywords "Opt freq B3LYP/6-31G(d)", the molecule structures were optimized with the Gaussian 09W.[37] Confirming that there is no virtual frequency in the convergence of the optimization result, the optimized structures were saved as .sdf format files. In the end, 4885 molecular descriptors for each optimized molecule were calculated in Dragon 6.0 software[38] (https://chm.kode-solutions.net/products_dragon.php) covering most of various theoretical approaches. The list of descriptors includes the

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

simplest atom types, functional groups, fragment counts, topological and geometrical descriptors, three-dimensional descriptors, and so forth.

The calculated descriptors with constant and near-constant values, descriptors with at least one missing value (some descriptors of some molecules can't be calculated), and descriptors with pair correlation larger than or equal to 0.90 were excluded in Dragon 6.0. As a result, 363 descriptors were retained.

4.3. Descriptor Screening and Model Development. Generally, the multiple linear regression method[39] is applied to describe the linear quantitative relationship between a dependent variable and multiple independent variables. It is used commonly in QSPR research on account of its simplicity, easy interpretability, and transparency.[40,41]

The GFA method can simulate biological evolution to generate statistical models, which was derived from the combination of genetic algorithm and multivariate adaptive regression splines.[42] In the present study, this method was used to screen molecular descriptors and provide several MLR models for selecting the best regression model. A flowchart of the GFA process can be seen in Supporting Information (Figure S1). Model development was carried out in the QSAR module of Material Studio 8.0 software. At first, multiple equations are randomly established on the basis of the population of descriptors. From the developed equations, the "parents" are selected according to the probability ratio of fitting, and the next generation equation is generated. After a crossover of each generation, mutations were made.[43] These operations will be repeated continuously for the specified number of iterations unless the convergence criteria are met. The fitting criteria for a GFA model can be evaluated using different scoring functions during evolution. In the present study, lack-of-fit was used as the scoring function because it can determine the most proper number of variables, avoid overfitting, and make the smoothness of the fitting under control.[44,45]

4.4. Model Validation. Model validation was performed to evaluate the goodness-of-fit statistic, robustness, and predictive power of the developed model.

4.4.1. Internal Validation. The role of the internal validation is to monitor the accuracy of the developed model and confirm the presence of overfitting. Generally, the commonly applied parameters to measure the goodness-of-fit statistic is the squared correlation coefficient (R_{tr}^2), the RMSE, and AARD (%) of the training set. Moreover, for MLR model, there must be the test of multicollinearities among the descriptors to reduce the redundant parameters. Thus, the VIF was calculated. If VIF values of each descriptor in the developed model are lower than 10, it can be concluded that no multicollinearity exist among the descriptors.[22] In addition, leave-one-out cross validation (LOO cross validation) was applied to further evaluate the possibility of overfitting with Q_{CV}^2, $RMSE_{CV}$.

4.4.2. External Validation. External validation was performed to assess the generalization ability of the developed model. Validation metrics such as R_{test}^2, $RMSE_{test}$, AARD %,[46,47] $Q_{ext\text{-}F1}^2$, $Q_{ect\text{-}F2}^2$, $Q_{ect\text{-}F3}^2$, r_m^2, and CCC[48] were applied to evaluate the model predicting performance. Tropsha et al.[24,25] proposed that whether the developed QSPR model successful depends on the following criteria of the test set

$$R_{CV,ext}^2 > 0.5$$

$$r^2 > 0.6$$

$$\frac{(r^2 - r_0^2)}{r^2} < 0.1 \text{ or } \frac{(r^2 - r_0'^2)}{r^2} < 0.1$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15$$

The parameters involved in above conditions are presented clearly in the Reference of Tropsha et al.[49]

In addition, in order to further illustrate the stability and robustness of the developed model, Roy et al.[26] proposed the MAE based criteria. If the MAE of the developed model satisfies the below conditions,[50] it can be concluded that the model is acceptable

$$\text{MAE} \le 0.1 \times \text{training set range and MAE} + 3 \times \sigma \le 0.2 \times \text{training set range}$$

where the σ value represents the standard deviation of the absolute error values for the test set data.

Furthermore, in this study, the quality of predictions for the external compounds was checked with the following tool: "prediction reliability indicator", which is proved very valid for multiple linear regression models.[50] This tool categorized the quality of predictions for the test set into three groups (good, moderate, and bad) based on absolute prediction errors.

4.4.3. Y-Randomization Test. In addition, Y-Randomization was performed to confirm the presence of chance correlation between the dependent variables and the independent variables.[24,51] For Y-Randomization test, low R_{rand}^2 and Q_{rand}^2 compared to the original R^2 and Q^2 of the resulting model are expected.[52] Meanwhile, the ${}^{C}R_p^2$ was applied to evaluate the chance correlation, which is expected to be close to the value of R_{tr}^2.[53] The Y-randomization test was performed in the program package MLR Y-randomization 1.2 (http://dtclab.webs.com/software-tools).

4.5. Applicability Domain. It is essential to define a domain of applicability for the developed model, namely, the AD. The AD is determined by the response (property) of the compounds and the descriptors in the developed QSPR model.[54] The Williams plot with standardized cross validated residuals (R) versus leverage (Hat diagonal) values (h) is highly recommended, which is widely used for visualizing the AD.[55−57] It should be noted that if the h value of a compound in the training set is greater than the threshold value h^* ($h^* = 3p/n$, where p is the number of model variables plus one and n is the number of the training set data), the structure of this compound reinforces the developed model.[10,58] If the most data points are located in the region of $0 \le h \le h^*$ and $-3 \le R \le 3$, the developed model can be considered statistically acceptable and valid.

If cross validated standardized residuals of a compounds is greater than three standard deviation units ($R > 3\sigma$) while the leverage value is lower than the threshold value ($h_i < h^*$), this compound could be judged as the response outlier (Y outlier). If the leverage value of a compound is greater than threshold value ($h_i > h^*$) while the value of R is lower than the 3 standard deviation units, this compound could be judged as the structurally influential compounds (X outlier).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.9b04190.

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

Chemicals involved in this work and values of molecular descriptors and thermal conductivity (PDF)

AUTHOR INFORMATION

Corresponding Authors

Wanqiang Liu − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*; orcid.org/0000-0001-5561-0867; Email: wanqiangliu@hnust.edu.cn

Yinchun Jiao − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*; Email: yinchunjiao@hnust.edu.cn

Authors

Haixia Lu − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Fan Yang − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Hu Zhou − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Fengping Liu − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Hua Yuan − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Guanfan Chen − *School of Chemistry and Chemical Engineering, Key Laboratory of Theoretical Organic Chemistry and Function Molecule of Ministry of Education, Hunan Province College Key Laboratory of QSAR/QSPR, Hunan Provincial Key Laboratory of Controllable Preparation and Functional Application of Fine Polymers, Hunan University of Science and Technology, Xiangtan 411201, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.9b04190

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work was financially supported by the National Natural Science Foundation of China (no. 21776067), the Hunan Provincial Natural Science Foundation of China (no. 2019JJ50161), and the Scientific Research Fund of Hunan Provincial Education Department (no.19K031).

ABBREVIATIONS

Roman symbols

R, the correlation coefficient; *T*, temperature

Greek symbols

λ, thermal conductivity

Sub- and superscripts

cal, calculated property; exp, experimental property; QSPR, quantitive structure−property relationship; *SM2_B(s)*, spectral moment of order 2 from Burden matrix weighted by I-state; *SIC0*, structural information content index (neighborhood symmetry of 0-order); *IC1*, information content index (neighborhood symmetry of 1-order); *Eta_F*, eta functionality index; *MATS2m*, Moran autocorrelation of lag 2 weighted by mass; MLR, multiple linear regression; GFA, genetic function approximation; AARD, average absolute relative deviation; RMSE, root-mean-square error; VIF, variance inflation factors; AD, applicability domain

REFERENCES

(1) Yang, I.; Kim, D.; Lee, S. Construction and preliminary testing of a guarded hot plate apparatus for thermal conductivity measurements at high temperatures. *Int. J. Heat Mass Transfer* **2018**, *122*, 1343−1352.

(2) Gustafsson, S. E.; Karawacki, E.; Khan, M. N. Transient hot-strip method for simultaneously measuring thermal conductivity and thermal diffusivity of solids and fluids. *J. Phys. D: Appl. Phys.* **1979**, *12*, 1411.

(3) Parker, W. J.; Jenkins, R. J.; Butler, C. P.; Abbott, G. L. Flash method of determining thermal diffusivity, heat capacity, and thermal conductivity. *J. Appl. Phys.* **1961**, *32*, 1679−1684.

(4) Gustafsson, S. E. Transient plane source techniques for thermal conductivity and thermal diffusivity measurements of solid materials. *Rev. Sci. Instrum.* **1991**, *62*, 797−804.

(5) Tada, Y.; Harada, M.; Tanigaki, M.; Eguchi, W. Laser flash method for measuring thermal conductivity of liquids-application to low thermal conductivity liquids. *Rev. Sci. Instrum.* **1978**, *49*, 1305−1314.

(6) Watanabe, H.; Kato, H. Thermal Conductivity and Thermal Diffusivity of Twenty-Nine Liquids: Alkenes, Cyclic (Alkanes, Alkenes, Alkadienes, Aromatics), and Deuterated Hydrocarbons. *J. Chem. Eng. Data* **2004**, *49*, 809−825.

(7) Rodenbush, C. M.; Viswanath, D. S.; Hsieh, F.-h. A Group Contribution Method for the Prediction of Thermal Conductivity of Liquids and Its Application to the Prandtl Number for Vegetable Oils. *Ind. Eng. Chem. Res.* **1999**, *38*, 4513−4519.

(8) Nagvekar, M.; Daubert, T. E. A group contribution method for liquid thermal conductivity. *Ind. Eng. Chem. Res.* **1987**, *26*, 1362−1365.

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

(9) Baroncini, C.; Filippo, P. D.; Latini, G. Thermal conductivity estimation of the organic and inorganic refrigerants in the saturated liquid state. *Int. J. Refrig.* **1983**, *6*, 60−62.
(10) Dearden, J. C. The History and Development of Quantitative Structure-Activity Relationships (QSARs). *Int. J. Quant. Struct.-Prop. Relat.* **2016**, *1*, 44.
(11) Liu, W. Q.; Cao, C. Z. QSPR Study of Normal Boiling Point of Aliphatic Esters Based on the Polarizability Effect Index. *Acta Chim. Sin* **2010**, *68*, 2401−2408.
(12) Gebreyohannes, S.; Dadmohammadi, Y.; Neely, B. J.; Gasem, K. A. M. A Comparative Study of QSPR Generalized Activity Coefficient Model Parameters for Vapor-Liquid Equilibrium Mixtures. *Ind. Eng. Chem. Res.* **2016**, *55*, 1102−1116.
(13) Yuan, S.; Jiao, Z.; Quddus, N.; Kwon, J. S.-I.; Mashuga, C. V. Developing Quantitative Structure-Property Relationship Models To Predict the Upper Flammability Limit Using Machine Learning. *Ind. Eng. Chem. Res.* **2019**, *58*, 3531−3537.
(14) Gao, S.; Cao, C. Z. A Topological-Quantum Method for the Estimation of the Thermal Conductivity of Liquid Alkanes. *Acta Phys.-Chim. Sin.* **2006**, *22*, 1478−1483.
(15) Kauffman, G. W.; Jurs, P. C. Prediction of Surface Tension, Viscosity, and Thermal Conductivity for Common Organic Solvents Using Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 408−418.
(16) Khajeh, A.; Modarress, H. Quantitative structure-property relationship prediction of liquid thermal conductivity for some alcohols. *Struct. Chem.* **2011**, *22*, 1315−1323.
(17) Liu, W. Q.; Lu, H. X.; Cao, C. Z.; Jiao, Y. C.; Chen, G. F. An Improved Quantitative Structure Property Relationship Model for Predicting Thermal Conductivity of Liquid Aliphatic Alcohols. *J. Chem. Eng. Data* **2018**, *63*, 4735−4740.
(18) Lu, H.; Yang, F.; Liu, W.; Yuan, H.; Jiao, Y. A Robust Model for Estimating Thermal Conductivity of Liquid Alkyl Halides. *SAR QSAR Environ. Res.* **2020**, *31*, 73−85.
(19) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* **2010**, *110*, 5714−5789.
(20) Wang, B.; Zhou, L.; Xu, K.; Wang, Q. Prediction of Minimum Ignition Energy from Molecular Structure Using Quantitative Structure-Property Relationship (QSPR) Models. *Ind. Eng. Chem. Res.* **2017**, *56*, 47−51.
(21) Nguyen, L. H.; Truong, T. N. Quantitative Structure-Property Relationships for the Electronic Properties of Polycyclic Aromatic Hydrocarbons. *ACS Omega* **2018**, *3*, 8913−8922.
(22) Famini, G. R.; Penski, C. A.; Wilson, L. Y. Using theoretical descriptors in quantitative structure activity relationships: Some physicochemical properties. *J. Phys. Org. Chem.* **1992**, *5*, 395−408.
(23) Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P. A.; Markopoulos, J.; Igglessi-Markopoulou, O. A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. *Bioorg. Med. Chem.* **2006**, *14*, 6686−6694.
(24) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(25) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.
(26) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18−33.
(27) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *Altern. Lab. Anim.* **2005**, *33*, 445−459.
(28) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2000.
(29) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379−423.
(30) Lienhard, J. H. I.; Lienhard, J. H. V. *A Heat Transfer Textbook*; Phlogyston Press: Cambridge: Massachusetts, 2011.
(31) Roy, K.; Ghosh, G. Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. *Int. Electron J. Mole. Design* **2003**, *2*, 599−620.
(32) Shi, J.; Chen, L.; Chen, W.; Shi, N.; Yang, H.; Xu, W. Prediction of the Thermal Conductivity of Organic Compounds Using Heuristic and Support Vector Machine Methods. *Acta Phys.-Chim. Sin.* **2012**, *28*, 2790−2796.
(33) Vargaftik, N. B. *Handbook of Thermal Conductivity of Liquids and Gases*; CRC Press: Boca Raton FL, 1994.
(34) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11*, 137−148.
(35) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.
(36) Frisch, H. P. H. A.; Dennington, R. D.; Keith, T. A. M. J.; Nielsen, A. J. H. A.; Hiscocks, J. *GaussView*, version 5.08; Semichem. Inc.: Wallingford, CT, 2010.
(37) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E. *Gaussian 09*; Gaussian, Inc.: Wallingford CT, 2009.
(38) Srl, Talete. *Dragon Software for Molecular Descriptor Calculation, 6.0*; Srl, Talete: Milano, Italy, 2014.
(39) Brereton, R. *Applied Chemometrics for Scientists*; John Wiley & Sons, Ltd.: Chichester, U.K., 2007.
(40) Roy, K.; Kar, S.; Das, R. N., *Statistical Methods in QSAR/QSPR*; Springer International Publishing: Cham, 2015.
(41) Lu, Y.; Ng, D.; Mannan, M. S. Prediction of the Reactivity Hazards for Organic Peroxides Using the QSPR Approach. *Ind. Eng. Chem. Res.* **2011**, *50*, 1515−1522.
(42) Roy, K.; Pratim Roy, P. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur. J. Med. Chem.* **2009**, *44*, 2913−2922.
(43) Khajeh, A.; Modarress, H. QSPR prediction of flash point of esters by means of GFA and ANFIS. *J. Hazard. Mater.* **2010**, *179*, 715−720.
(44) Mirkhani, S. A.; Gharagheizi, F.; Ilani-Kashkouli, P.; Farahani, N. An accurate model for the prediction of the glass transition temperature of ammonium based ionic liquids: A QSPR approach. *Fluid Phase Equilib.* **2012**, *324*, 50−63.
(45) Mirkhani, S. A.; Gharagheizi, F. Predictive Quantitative Structure-Property Relationship Model for the Estimation of Ionic Liquid Viscosity. *Ind. Eng. Chem. Res.* **2012**, *51*, 2470−2477.
(46) Consonni, V.; Todeschini, R.; Ballabio, D.; Grisoni, F. On the Misleading Use of Q^2_{F3} for QSAR Model comparison. *Mol. Inf.* **2018**, *38*, 1800029.
(47) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26*, 694−701.
(48) Roy, K.; Mitra, I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screening* **2011**, *14*, 450−474.
(49) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX

(50) Roy, K.; Ambure, P.; Kar, S. How Precise Are Our Quantitative Structure-Activity Relationship Derived Predictions for New Query Chemicals? *ACS Omega* **2018**, *3*, 11392−11406.
(51) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.
(52) Ma, S.; Lv, M.; Deng, F.; Zhang, X.; Zhai, H.; Lv, W. Predicting the ecotoxicity of ionic liquids towards Vibrio fischeri using genetic function approximation and least squares support vector machine. *J. Hazard. Mater.* **2015**, *283*, 591−598.
(53) Mitra, I.; Saha, A.; Roy, K. Exploring quantitative structure-activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol. Simul.* **2010**, *36*, 1067−1079.
(54) Klingspohn, W.; Mathea, M.; ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of different measures for defining the applicability domain of classification models. *J. Cheminf.* **2017**, *9*, 44.
(55) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791−4810.
(56) Mansouri, K.; Cariello, N. F.; Korotcov, A.; Tkachenko, V.; Grulke, C. M.; Sprankle, C. S.; Allen, D.; Casey, W. M.; Kleinstreuer, N. C.; Williams, A. J. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J. Cheminf.* **2019**, *11*, 60.
(57) Cao, L.; Zhu, P.; Zhao, Y.; Zhao, J. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. *J. Hazard. Mater.* **2018**, *352*, 17−26.
(58) Soro, D.; Ekou, L.; Ouattara, B.; Kone, M. G.-R.; Ekou, T.; Ziao, N. DFT Study, Linear and Nonlinear Multiple Regression in the Prediction of HDAC7 Inhibitory Activities on a Series of Hydroxamic Acids. *Comput. Mol. Biosci.* **2019**, *09*, 63−80.

https://dx.doi.org/10.1021/acsomega.9b04190
ACS Omega XXXX, XXX, XXX−XXX