ACADEMIA ROMÂNĂ
Revue Roumaine de Chimie
http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2018, *63*(2), 111-121

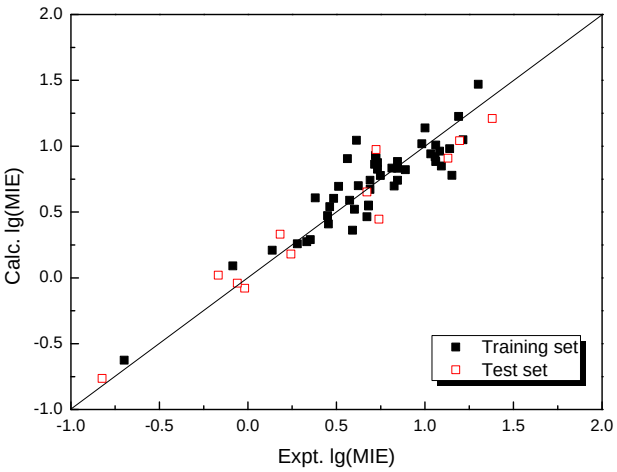
VALIDATED QSPR MODELS FOR THE PREDICTION OF MINIMUM IGNITION ENERGY

Shan LI,[a] Chengzhuo WEI,[b] Lingling FAN[b] and Jie XU[b,]

[a] Information and Engineering School, Wuchang University of Techonology, 430223, Wuhan, China
[b] College of Materials Science & Engineering, State Key Laboratory of New Textile Materials & Advanced Processing Technology, Wuhan Textile University, 430200, Wuhan, China

Received April 25, 2017

A quantitative structure-property relationship (QSPR) study was reported to predict the minimum ignition energy based on a training set of 48 chemicals and a test set of 12 chemicals divided by DUPLEX algorithm. The QSPR models were built by stepwise multiple linear regression (MLR) and nonlinear artificial neural network (ANN), respectively. The average absolute error provided by the MLR and ANN model was 0.126 and 0.116, respectively, indicating satisfactory predictive ability. The results from Y-randomization tests, leave-one-out cross-validations, and external validation through test set confirmed the reliability and robustness of the models. Moreover, the applicability domain of the developed models was assessed and visualized by the Williams plot.



INTRODUCTION

Minimum ignition energy (MIE) is of great importance to understanding and measuring ignition hazards of fuels.[1] A specific threshold of ignition energy (corresponding to the MIE) can be observed for each flammable mixture, and ignition sources with energy less than this threshold will never ignite the mixture.[2] The MIE is usually determined from the energy stored in a capacitor at a known voltage that is then discharged through a specified fixed electrode gap (*i.e.*, the ASTM E582).[2] However, the experimental determination of MIE for flammable and explosive chemicals is dangerous, costly and time-consuming along with high experimental uncertainties. Therefore, reliable theoretical models to estimate the MIE values are strongly required.

Corresponding author: xujieo@ustc.edu

Quantitative structure-property relationship (QSPR) approach is based on the assumption that the variation of the behavior of the compounds, as expressed by any measured physicochemical properties, can be correlated with numerical changes in structural features of all compounds.[3-20] The advantage of this approach lies in the fact that it requires only the knowledge of the chemical structure and is not dependent on any experimental properties. Once a correlation is established and validated, it can be applicable for the prediction of the property of new compounds that have not been synthesized or found. To support this development, Organization for Economic Co-operation and Development (OECD) has drawn up the following principles for the validation of QSPR models:[21] (1) a defined endpoint (including experimental protocol); (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures of goodness-of-fit, robustness and predictive power; (5) a mechanistic interpretation, when it is possible. Baati has tried to develop some models to predict MIE values but unfortunately none of them are qualified.[1] The proposed global model comprised as many as 16 parameters and the average deviation of this model was even larger than 1 000%. Among the three local models reported, the only satisfying one was the model built for the solid state materials. However, this model included as many as 27 parameters, which was very complex and unpractical.

More recently, Wang *et al.*[22] built two QSPR models by multiple linear regression (MLR) and support vector machine (SVM), with nine descriptors involved. These models were characterized by good determination coefficients (R^2), but they were not validated adequately and their respective applicability domains were not defined. These models did not fit all OECD standards and could not be used within the framework of the European Union regulation REACH (Registration, Evaluation, Authorization and Restriction of Chemicals).[23]

The aim of this work is to develop fully validated QSPR models to predict the MIE values respecting all OECD principles including the determination of their applicability domains. Linear and nonlinear models were built by using MLR and artificial neural network (ANN) methods. Both models were validated by a series of internal and external validation methods, including cross-validations to characterize robustness, Y-randomization technique to avoid chance correlations and external validation to estimate predictive power of models and taking into account the applicability domain (AD).

MATERIALS AND METHOD

1. Dataset

Table 1

Experimental and calculated lg (MIE) values for the training and test sets

No.	Compound	Expt.	Calc.		No.	Compound	Expt.	Calc.	
			MLR	ANN				MLR	ANN
1	Ethane	0.455	0.409	0.437	31	Acrolein	0.137	0.211	0.192
2	Propane	0.484	0.604	0.637	32	Propionaldehyde	0.512	0.696	0.732
3	Methane	0.672	0.465	0.504	33	Acetaldehyde	0.575	0.589	0.566
4	n-Pentane	0.69	0.742	0.743	34	Methyl ethyl ketone	0.724	0.910	0.949
5	Isobutane	0.716	0.863	0.853	35	Acetone	1.061	0.883	0.929
6	Isopentane	0.845	0.883	0.867	36	Methyl formate	0.602	0.521	0.530

8	Triptane	1	1.139	1.053	38	Ethyl acetate	1.152	0.780	0.805
9	Isooctane [a]	1.13	0.909	0.840	39	Dimethyl ether	0.462	0.541	0.580
10	2,2-Dimethylpropane [a]	1.196	1.042	0.991	40	Dimethoxymethane	0.623	0.700	0.728
11	2,2-Dimethylbutane	1.215	1.049	0.997	41	Diethyl ether	0.69	0.672	0.690
12	Acetylene	-0.699	-0.625	-0.693	42	Diisopropyl ether	1.057	0.920	0.869
13	Vinylacetylene	-0.085	0.091	0.055	43	Dimethyl sulfide	0.681	0.553	0.591
14	Ethylene [a]	-0.018	-0.079	-0.193	44	Di-tert-butyl peroxide	0.613	1.045	0.938
15	Methylacetylene [a]	0.182	0.332	0.290	45	Furan	0.352	0.290	0.419
16	1,3-Butadiene [a]	0.243	0.181	0.258	46	Thiophene	0.591	0.362	0.498
17	Propylene	0.45	0.473	0.501	47	Benzene [a]	0.74	0.446	0.656
18	1-Heptyne	0.748	0.778	0.770	48	Ethylene oxide [a]	-0.06	-0.040	-0.059
19	2-Pentene [a]	0.672	0.653	0.676	49	Propylene oxide	0.279	0.259	0.273
20	Diisobutylene	0.982	1.018	0.944	50	Cyclopropane	0.38	0.608	0.642
21	Methanol	0.332	0.275	0.264	51	Dihydropyran	0.562	0.906	0.919
22	Isopropyl mercaptan [a]	0.724	0.975	1.009	52	Ethylenimine	0.681	0.546	0.584
23	Isopropyl alcohol	0.813	0.834	0.862	53	Cyclohexene	0.72	0.929	0.910
24	Allyl chloride	0.889	0.821	0.829	54	Cyclopentane	0.732	0.875	0.856
25	n-Propyl chloride	1.033	0.942	0.974	55	Tetrahydrofuran	0.732	0.824	0.837
26	Triethylamine	1.061	1.009	0.926	56	Cyclopentadiene	0.826	0.697	0.730
27	n-Butyl chloride	1.093	0.850	0.888	57	Tetrahydropyran	1.083	0.962	0.947
28	Isopropyl chloride	1.19	1.226	1.139	58	Cyclohexane	1.14	0.982	0.933
29	Isopropylamine	1.301	1.471	1.352	59	Carbon disulfide [a]	-0.824	-0.763	-0.768
30	Ethylamine [a]	1.38	1.210	1.319	60	Hydrogen sulfide [a]	-0.167	0.020	-0.036

[a] Compounds in the test set.

The experimental MIE values for 60 chemicals (Table 1) were collected from Calcote e*t al.*,[24] which were measured with stoichiometric fuel-air mixtures at a pressure of 1 atm.

2. Descriptor generation

The chemical structure was drawn with the HYPERCHEM program[25] and preoptimized by the MM+ molecular mechanics method[26] (Polak-Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical AM1 method[27] at a restricted Hartree-Fock level with no configuration interaction, applying a gradient limit of 0.03 kcal·Å^{-1}·mol^{-1} as a stopping criterion. Then totally 1 664 molecular descriptors for each compound were calculated based on the optimized geometries with the DRAGON software.[28]

To reduce redundant and useless information, descriptors with constant or near constant values and descriptors found to be highly correlated pairwise (one of any two descriptors with a correlation greater than 0.99[29]) were removed. Finally, 678 descriptors remained.

3. Dataset splitting

The dataset was split into training and test sets by the DUPLEX algorithm[30] combined with principal component analysis (PCA). This algorithm proceeds as follows: first the two points which are furthest away from each other are selected for the training set;

from the remaining points, the two objects which are furthest away from each other are included in the test set; then the remaining point which is furthest away from the two previously selected for the training set is included in the training set. The procedure is repeated selecting a single point for the test set which is furthest from the existing points in that set. Following the same procedure, points are added alternately to each set. Finally, points representing both training and test set were distributed uniformly within the whole space which is occupied by the entire dataset. Therefore, it guarantees that the composition of the training set and the test set is representative, at the same time it avoids the unbalance of the two datasets. The PCA was first performed based on the 678 descriptors of the complete dataset, and the obtained principal components were put into the DUPLEX algorithm to select a training set of 48 compounds and a test set of 12 compounds.

4. Model development

Stepwise MLR analysis combined with leave-one-out (LOO) cross-validation was used to select the best subset of descriptors and to develop linear QSPR models based on the training set. *F*-to-enter and *F*-to-remove were 4 and 3, respectively. In addition, a variance inflation factor (VIF) was calculated to test if multicollinearities existed among the descriptors, which is defined as

$$\mathrm{VIF} = \frac{1}{1 - R_j^2} \qquad (1)$$

where R^2_{j} is the squared correlation coefficient between the *j*th coefficient regressed against all the other descriptors in the model. If VIF equals to one, no self-correlation exists for each descriptor; if VIF ranges from 1.0 to 5.0, the corresponding model is acceptable; if VIF is larger than 10.0, the corresponding model is unstable and re-check of the variance correlation is needed.[31]

Nonlinear models were then developed by submitting the selected descriptors from MLR to a three-layer, fully connected, feed-forward ANN. The number of input neurons was equal to that of the descriptors in the linear model. The number of hidden neurons was optimized by trial and error procedure on the training process. One output neuron was used to represent the experimental value. The network was trained using the quasi-Newton BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm.[32] To avoid overtraining, one tenth data from the training set were randomly selected as separate validation set to monitor the training process.

5. Model evaluation and validation

Model performance was evaluated by the following statistical parameters: the R^2, the adjusted R^2, the *s*, the *F* ratio values, and the significance level value *p*. The adjusted R^2 is calculated using the following formula:

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-m-1} \right) (1 - R^2) \right] \qquad (2)$$

where *n* is the number of members of the training set and *m* is the number of descriptors involved in the model.

The predictive performance of the models was measured by average absolute error (AAE), defined as:

$$AAE = \frac{1}{n} \sum \left| y_i - \tilde{y}_i \right| \qquad (3)$$

where y_i and $\tilde{y}_i$ are the experimental and the calculated response values, respectively.

Y-randomization tests were carried out to prove the possible existence of chance correlation during the MLR model development. The resulting models obtained on the training set with the randomized response values should be of significantly poor quality when compared with the

proposed one because the relationship between the structure and response is broken. This is a proof of the proposed model's validity as it can be reasonably excluded that the originally proposed model was obtained by chance correlation.

Validation of the models was further performed by using the external test set. The external Q^2_{F1} for the test set is determined with the following equation:

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{tra})^2} \quad (4)$$

where $\bar{y}_{tra}$ is the averaged value for the response variable of the training set. According to Golbraikh and Tropsha,[33] a QSPR model is successful if it satisfies the following criteria:

$$Q^2_{F1} > 0.5, \quad r^2 > 0.6 \quad (5a)$$

$$(r^2 - r_0^2)/r^2 < 0.1 \text{ or } (r^2 - r'^2_0)/r^2 < 0.1 \quad (5b)$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \quad (5c)$$

Here:

$$r = \frac{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2 \sum_{i=1}^{n_{ext}} (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (6a)$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^{n_{ext}} (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum_{i=1}^{n_{ext}} (\tilde{y}_i - \bar{\tilde{y}})^2},$$

$$r'^2_0 = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - y_i^{r_0})^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2} \quad (6b)$$

$$k = \frac{\sum_{i=1}^{n_{ext}} y_i \tilde{y}_i}{\sum_{i=1}^{n_{ext}} y_i^2}, \quad k' = \frac{\sum_{i=1}^{n_{ext}} y_i \tilde{y}_i}{\sum_{i=1}^{n_{ext}} \tilde{y}_i^2 \sum \tilde{y}_i^2} \quad (6c)$$

where $\tilde{y}^{r_0}$ and y^{r_0} are defined as $\tilde{y}^{r_0} = ky$ and $y^{r_0} = k'\tilde{y}$, respectively.

Roy and coworkers[34, 35] proposed a set of r_m^2 metrics for the validation of QSPR models, which are defined as follows:

$$r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2}) \quad (7a)$$

$$\overline{r_m^2} = \frac{r_m^2 + r'^2_m}{2} \quad (7b)$$

$$\Delta r_m^2 = \left| r_m^2 - r'^2_m \right| \quad (7c)$$

For a successful QSPR model, $\overline{r_m^2}$ value should be larger than 0.5 and Δr_m^2 should be less than 0.2.

6. Applicability domain analysis

The applicability domain (AD)[36, 37] of the QSPR models was verified by the leverage approach. The leverage h_i [37] is defined as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (8)$$

where x_i is the descriptor row-vector of the *i*-th compound, x_i^T is the transpose of x_i, X is the descriptor matrix, and X^T is the transpose of X. The warning leverage h^* is generally fixed at $3(m+1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The Williams plot, the plot of leverage values *versus* standardized residuals, was used to give a graphical detection of both the response outliers (Y outliers) and the structurally influential compounds

(X outliers). In this plot, the two horizontal lines indicate the limit of normal values for Y outliers (*i.e.* samples with standardized residuals greater than 3.0 standard deviation units, ±3.0 *s*); the vertical straight lines indicate the limits of normal values for X outliers (*i.e.* samples with leverage values greater than the threshold value, $h > h^*$). For a sample in the external test set whose leverage value is greater than h^*, its prediction is considered unreliable, because the prediction is the result of a substantial extrapolation of the model. On the contrary, the accordance probability between the calculated and experimental values is as high as that for the samples in the training set when the leverage value of a compound is lower than h^*. It is noteworthy that the response outliers can be highlighted only for compounds with known responses and the possibility of a compound to be out of the structural AD of a model can be verified for every new compound.

RESULTS AND DISCUSSION

1. Results of the MLR model

Stepwise MLR analysis was employed to find the best linear model based on the training set. It can be found that MIE is not linearly correlated with any of the descriptors since none of the one-descriptor model was found to be significant. The six-descriptor model was selected as the best model:

$$\lg(\text{MIE}) = -0.294[\text{P1s}] + 0.264[\text{HIC}] + 0.335[\text{H0u}] + 2.377[\text{HATS3m}] + 0.638[\text{nRNH2}] - 0.466[\text{nOxiranes}] \quad (9)$$

n = 48, $R^2 = 0.872$, $R_{adj}^2 = 0.858$, $R_{CV}^2 = 0.831$, $s = 0.167$, $F = 60.39$, $p < 0.00001$

where P1s is the 1st component shape directional WHIM index/weighted by atomic electrotopological states; HIC is the mean information content on the leverage magnitude; H0u is the H autocorrelation of lag 0/unweighted; HATS3m is the leverage-weighted autocorrelation of lag 3/weighted by atomic masses; nRNH2 is the number of primary amines (aliphatic); nOxiranes is the number of Oxiranes. More detailed information about these descriptors can be found in Dragon software user's guide[28] and the references therein.

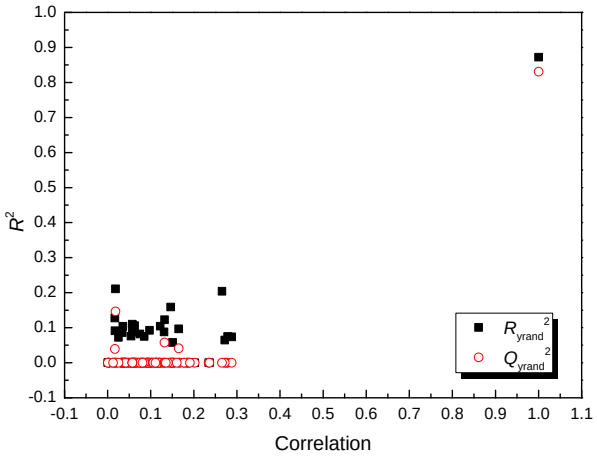


Fig. 1 – R^2 and Q^2 *vs.* correlation coefficient between the original and permuted response data.

Generally, the larger the magnitude of the *F* ratio, the better the model predicts the property values in the training set. The large *F* ratio of 60.39 points out that Eq. (9) does an excellent job of predicting lg (MIE). An adjusted R^2 value of 0.858 is achieved, indicating good agreement between the correlation and the variation in the data. In order to determine the robustness and neglect of chance correlation, the model was validated by Y-randomization tests. The obtained R_{yrand}^2 and Q_{yrand}^2 *vs.* the correlation coefficient between the original and permuted response data are shown in Figure 1. The quality for all randomized models is significantly poorer when compared to the original model. The intercepts for R_{yrand}^2 and Q_{yrand}^2 were regressed to be -0.036 and -0.052, respectively, which are within the limit values suggested in the literature, i.e., below 0.3 for R_{yrand}^2 and 0.05 for Q_{yrand}^2.[38] Therefore, there is no risk of chance correlation in the developed model. The involved descriptors were evaluated by the statistical parameters shown in Table 2. The high absolute *t*-values of the descriptors involved in the MLR model indicate that the regression coefficients of

these descriptors are remarkably higher than the standard error. The *t*-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (*i.e.*, descriptors' interactions). Descriptors with *t*-probability values below 0.05 (95% confidence) are usually considered being statistically significant in a particular model, which means that their influence on the response variable is not merely by chance.[39] The *t*-probability values of these descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values (less than five) indicate that these descriptors are weakly correlated with each other. Therefore, the proposed model can be regarded as an optimal regression equation.

Table 2

Characteristics of the descriptors selected in the optimal MLR model

Descriptor	Descriptor type	Coefficient	Error	*t*-value	*t*-probability	VIF
Constant		-1.068	0.155	-6.881	0.000	
P1s	WHIM descriptors	-0.294	0.092	-3.181	0.002	1.018
HIC	GETAWAY descriptors	0.264	0.037	7.217	0.000	1.595
H0u	GETAWAY descriptors	0.335	0.059	5.708	0.000	1.689
HATS3m	GETAWAY descriptors	2.377	0.461	5.161	0.000	1.104
nRNH2	Functional group counts	0.638	0.121	5.283	0.000	1.015
nOxiranes	Functional group counts	-0.466	0.123	-3.788	0.000	1.055

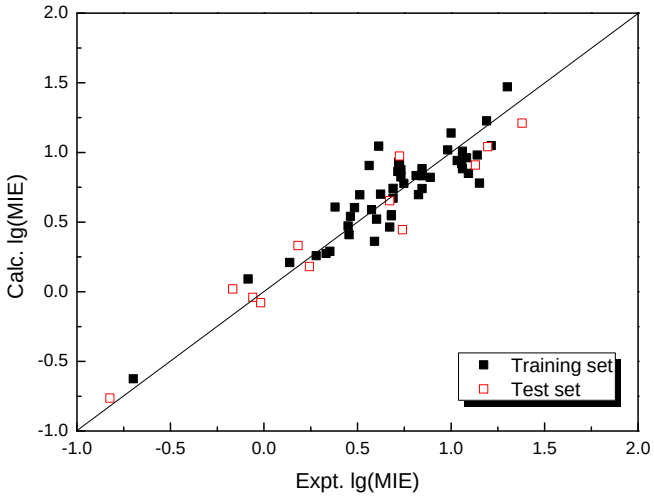


Fig. 2 – Calculated *vs.* experimental values for the MLR model.

The calculated lg (MIE) values from the MLR model for the training and test sets are given in Table 1 and Figure 2. The errors are distributed on both sides of the zero point, thus one may conclude that there is no systematic error in the model development.

The AAE for the training and test set is 0.124 and 0.137, respectively, giving rise to an AAE of 0.126 for the entire dataset.

The following statistical parameters were obtained for the test set, which obviously satisfy the generally accepted condition and thus demonstrate the predictive power of the present model:

$$Q^2_{F1} = 0.946 > 0.5, \quad r^2 = 0.940 > 0.6$$

$(r^2 - r_0^2)/r^2 = (0.940 - 0.980)/0.940 < 0.1$ or

$(r^2 - r'^2_0)/r^2 = (0.940 - 0.990)/0.940 < 0.1$

$0.85 \le k = 1.082 \le 1.15$ or $0.85 \le k' = 0.887 \le 1.15$

$$\overline{r_{\mathrm{m}}^2} = 0.909 > 0.5\,,\quad \Delta r_m^2 = 0.034 < 0.2$$

The results of the present MLR model are comparable in quality to those obtained by Wang *et al.*[22] in the same property, further demonstrating the reliability of the MLR model developed in our study.

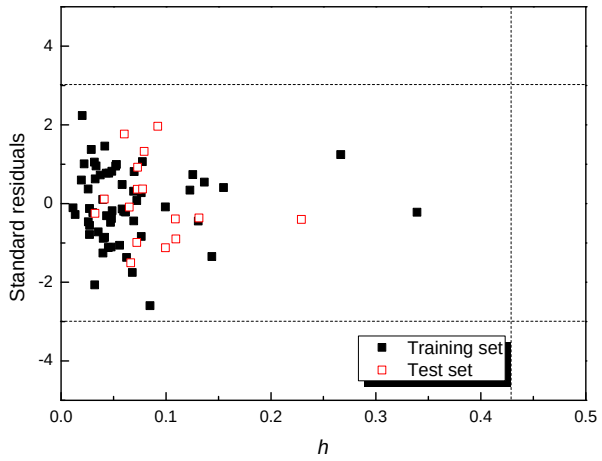


Fig. 3 – Williams plots of the MLR model.

One validated QSPR model cannot be expected to reliably predict the studied property for the entire universe of compounds. The predictions for only those samples that fall in the defined AD can be considered reliable. The AD visualized by the plot of standardized residuals *versus* leverage values (the Williams plot) is shown in Figure 3. The h^* equals 0.4375 (vertical dotted line in Figure 3). The X outliers can be assigned to those samples with special features poorly expressed in the training set; while the Y outliers can be related to the experimental errors. From the Williams plot of the MLR model, it can be seen that neither of the compounds for the training and test sets is identified as an outlier. All the samples are located within the model AD and are predicted accurately, which further indicates the reliability of the proposed model. Thus, this model could be utilized to screen the existing databases.

The relative contributions (C_i) of the six descriptors to the MLR model were calculated based on a previously reported procedure40. The significance of the descriptors to the MLR model decreases in the following order: HIC (19.7%) > H0u (17.4%) > HATS3m (17.0%) > nRNH2 (15.5%) ≈ nOxiranes (15.5%) > P1s (14.8%).

The importance of the GETAWAY descriptors on the MIE values is apparent, since the GATEWAY descriptors explain 54.1% of the contributions (19.7% of HIC, 17.4% of H0u and 17.0% of HATS3m).The first important descriptor is HIC, which is defined as

$$HIC = -\sum_{i=1}^{nAT} \frac{h_{ii}}{D} \cdot \log_2 \frac{h_{jj}}{D} \qquad (10)$$

where nAT is the number of molecule atoms; h_{ii} and h_{jj} are the leverages of the two considered atoms; D = 1, 2 or 3 (1 for linear, 2 for planar and 3 for non-planar molecules). HIC mainly catches the information related to molecular complexity and the presence of multiple bonds. The positive correlation coefficient for HIC indicates that one chemical with larger value for this descriptor would have higher MIE value. H0u is calculated as

$$H0u = \sum_{i=1}^{nAT-1} \sum\nolimits_{j>1} h_{ij} \cdot w_i \cdot \delta(0; d_{ij}; h_{ij}) \qquad (11)$$

where w_i is the unit weight (u); d_{ij} is the topological distance between atoms i and j; $\delta(0;\ d_{ij};\ h_{ij})$ is a Dirac-delta function (δ = 1 if d_{ij} = 0 and h_{ij} >0, zero otherwise). HATS3m is calculated as

$$HATS3m = \sum_{i=1}^{nAT-1} \sum\nolimits_{j>1} (m_i \cdot h_{ii}) \cdot (m_j \cdot h_{jj}) \cdot \delta(3; d_{ij}) \qquad (12)$$

where m is the atomic mass; $\delta(3;\ d_{ij})$ is a Dirac-delta function (δ = 1 if d_{ij} = 3, zero otherwise). H0u and HATS3m, as based on spatial autocorrelation, encode information on structural fragments and are suitable for describing differences in congeneric series of molecules. The correlation coefficients of H0u and HATS3m are positive, suggesting that these descriptors would increase the MIE values.

The nRH2 descriptor is the number of primary amines (aliphatic). The positive correlation coefficient of nRH2 indicates that the chemicals containing more primary amines (aliphatic) have higher MIE values. The nOxiranes descriptor is the number of Oxiranes. The negative sign of this descriptor indicates that the increase in the number of the Oxiranes groups in the molecules would give rise to the decrease in the MIE values. The presence of P1s in the model reflects the influence of the electrotopological state indices on the MIE values.

2. Results of the ANN model

ANN has been widely used as a powerful nonlinear tool for pattern classification and building predictive models in QSPR studies. The most attractive advantage of ANN is its inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function. In the present study, the quasi-Newton BFGS algorithm[32] was applied to develop the ANN models. The advantages of the BFGS algorithm are that specifying rate or momentum is not necessary and the training is much more rapid.[41] The six descriptors obtained from the MLR model were used as inputs to the network. The number of hidden neurons is a key parameter determining the ANN performances. The usual rule of thumb is that the weights and biases should be smaller than the samples so that the model obtained by the network is stationary.[42] Consequently, a 6-2-1 network architecture was achieved after the trial and error process, with *s* of 0.149 for the training set.

The predictive results from the ANN model for the entire dataset are presented in Table 1 and Figure 4. The proposed ANN model is predictive as it satisfies the conditions for the test set:

$$Q^2_{F1} = 0.952 > 0.5 \quad , \quad r^2 = 0.943 > 0.6$$

$$(r^2 - r_0^2)/r^2 = (0.943 - 0.900)/0.943 < 0.1 \text{ or}$$

$$(r^2 - r'^2_0)/r^2 = (0.943 - 0.997)/0.943 < 0.1$$

$$0.85 \le k = 1.044 \le 1.15 \text{ or } 0.85 \le k' = 0.920 \le 1.15$$

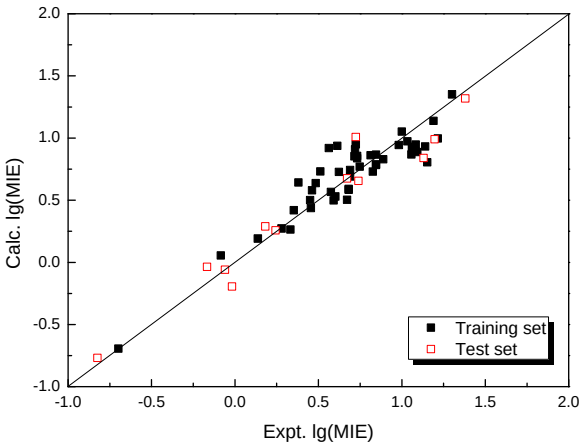


Fig. 4 – Calculated *vs*. experimental values for the ANN model.

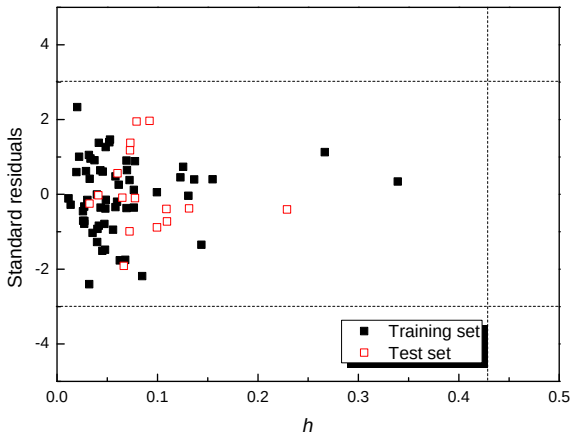


Fig. 5 – Williams plots of the ANN model.

$$\overline{r^2_m} = 0.942 > 0.5 \, , \quad \Delta r^2_m = 0.032 < 0.2$$

The AAE for the training and test set is 0.115 and 0.118, respectively, resulting in an AAE of 0.116 for the entire dataset. The Williams plot for the ANN model is shown in Figure 5. There is no outlier compound for the training and test sets, suggesting that the predictions from the ANN model are reliable.

CONCLUSION

In this work, QSPR models for the prediction of the minimum ignition energy were successfully developed by the MLR and ANN methods according to all OECD principles. The proposed

models have satisfactory predictive ability and robustness for the prediction of the minimum ignition energy, which were verified by internal validation (cross-validations and Y-randomization tests) and external validation through test set. Moreover, the applicability domain of the developed models was assessed and visualized by the Williams plot. The developed QSPR models can be used to predict the minimum ignition energy for those chemicals within the applicability domain. This investigation extends the research method to predict the minimum ignition energy.

Acknowledgements: This work was supported by the Natural Science Foundation of Hubei Province (No. 2012FFA098), the Scientific Innovation Team Project of the Education Department of Hubei Province (No. T201507), and Hubei Collaborative Innovation Center for Key Technologies in Textiles.

REFERENCES

1. N. Baati, "Models for Thermal Stability and Explosive Properties of Chemicals from Molecular Structure", *Ph.D. Thesis*, École Polytechnique Fédérale De Lausanne, 2016.
2. S. P. M. Bane, J. L. Ziegler, P. A. Boettcher, S. A. Coronel and J. E. Shepherd, *J. Loss Prev. Process Ind.*, **2013**, *26*, 290-294.
3. J. Devillers and A. T. Balaban (Eds.), "Topological Indices and Related Descriptors in QSAR and QSPR", Gordon and Breach, The Netherlands, 1999.
4. M. Karelson, "Molecular Descriptors in QSAR/QSPR", Wiley-Interscience, New York, 2000.
5. X. J. Yao, Y. W. Wang, X. Y. Zhang, R. S. Zhang, M. C. Liu, Z. D. Hu and B. T. Fan, *Chemom. Intell. Lab. Syst.*, **2002**, *62*, 217-225.
6. J. Xu, B. Chen, Q. Zhang and B. Guo, *Polymer*, **2004**, *45*, 8651-8659.
7. J. Xu, B. Guo, B. Chen and Q. Zhang, *J. Mol. Model.*, **2005**, *12*, 65-75.
8. J. Xu, L. Liu, W. Xu, S. Zhao and D. Zuo, *J. Mol. Graph. Model.*, **2007**, *26*, 352-359.
9. J. Xu, H. Liang, B. Chen, W. Xu, X. Shen and H. Liu, *Chemom. Intell. Lab. Syst.*, **2008,** *92*, 152-156.
10. J. Xu, L. Wang, L. Wang, X. Shen and W. Xu, *J. Comput. Chem.*, **2010**, *32*, 3241-3252.
11. J. Xu, H. Zhang, L. Wang, W. Ye, W. Xu and Z. Li, *Fluid Phase Equil*., **2010**, *291*.
12. G. Liang, J. Xu and L. Liu, *Fluid Phase Equil*., **2013**, *353*, 15-21.
13. X. Wang, Y. Sun, L. Wu, S. Gu, R. Liu, L. Liu, X. Liu and J. Xu, *Chemom. Intell. Lab. Syst.*, **2014**, *134*, 1-9.
14. Y. Yuan, Y. Sun, D. Wang, R. Liu, S. Gu, G. Liang and J. Xu, *Fluid Phase Equil.*, **2015**, *391*, 31-38.
15. D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu and J. Xu, *Chemom. Intell. Lab. Syst*., **2015**, *143*, 7-15.
16. X. Yu, B. Yi, Z. Xie, X. Wang and F. Liu, *Chemom. Intell. Lab. Syst.*, **200**7, *87*, 247-251.
17. A. Guendouzi and S. M. Mekelleche, *Chem. Phys. Lipids*, **2012**, *165*, 1-6.
18. A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, O. Igglessi-Markopoulou and G. Kollias, *Mol. Divers.*, **2010**, *14*, 225-235.
19. G. Melagraki and A. Afantitis, *Chemom. Intell. Lab. Syst.*, **2013,** *123*, 9-14.
20. O. Deeb, P. V. Khadikar and M. Goodarzi, *J. Iran. Chem. Soc.*, **2011**, *8*, 176-192.
21. "The principles for establishing the status of development and validation of (quantitative) structure activity relationships [(Q)SARs]", http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf, 2004.
22. B. Wang, L. Zhou, K. Xu and Q. Wang, *Ind. Eng. Chem. Res.*, **2017**, *56*, 47-51.
23. "EC, Regulation N(1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorization and restriction of chemicals (REACH)." EC, Brussels, 2006.
24. H. F. Calcote, C. A. Gregory, C. M. Barnett and R. B. Gilmer, *Ind. Eng. Chem.*, **1952**, *44*, 2656-2662.
25. C. T. Klein, D. Polheim, H. Viernstein and P. Wolschann, *Pharm. Res.*, **2000**, *17*, 358-365.
26. N. L. Allinger, Y. H. Yuh and J. H. Lii, *J. Am. Chem. Soc.*, **1989**, *111*, 8551-8566.
27. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, **1985**, *107*, 3902-3909.
28. R. Todeschini, V. Consonni, A. Mauri and M. Pavan, TALETE srl, Milan, 2006.

29. H. Liu and P. Gramatica, *Bioorgan. Med. Chem.*, **2007**, *15*, 5251-5261.
30. R. D. Snee, *Technometrics*, **1977**, *19*, 415-428.
31. G. R. Famini, C. A. Penski and L. Y. Wilson, *J. Phys. Org. Chem.*, **1992**, *5*, 395-408.
32. M. D. Wessel and P. C. Jurs, *Anal. Chem.*, **1994**, *66*, 2480-2487.
33. A. Golbraikh and A. Tropsha, *J. Mol. Graph. Model.*, **2002**, *20*, 269-276.
34. P. P. Roy and K. Roy, *QSAR Comb. Sci.*, **2008**, *27*, 302-313.
35. P. K. Ojha, I. Mitra, R. N. Das and K. Roy, *Chemom. Intell. Lab. Syst*., **2011**, *107*, 194-205.
36. A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR Comb. Sci.*, **2003**, *22*, 69-77.
37. A. Atkinson, "Plots, transformations, and regression", Clarendon Press, Oxford, UK, 1985.
38. R. Kiralj and M. M. C. Ferreira, *J. Braz. Chem. Soc*., **2009**, *20*, 770-787.
39. L. F. Ramsey and W. D. Schafer, "The Statistical Sleuth", Wadsworth Publishing Company, USA, 1997.
40. F. Zheng, E. Bayram, S. P. Sumithran, J. T. Ayers, C.-G. Zhan, J. D. Schmitt, L. P. Dwoskin and P. A. Crooks, *Bioorg. Med. Chem*., **2006,** *14*, 3017-3037.
41. L. Xu, J. W. Ball, S. L. Dixon and P. C. Jurs, *Environ. Sci. Chem.*, **1994**, *13*, 941-851.
42. Y.-H. Qi, Q.-Y. Zhang and L. Xu, *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 1471-1475.