Journal of Molecular Graphics and Modelling 44 (2013) 113–119

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Prediction of boiling points of organic compounds by QSPR tools

Dai Yi-min*, Zhu Zhi-ping, Cao Zhong, Zhang Yue-fei, Zeng Ju-lan, Li Xun

School of Chemistry and Biological Engineering, Hunan Provincial Key Laboratory of Materials Protection for Electric Power and Transportation, Changsha University of Science and Technology, Changsha 410004, China

ARTICLE INFO

Article history:
Received 8 January 2013
Accepted 24 April 2013
Available online 4 May 2013

Keywords:
Normal boiling point
Organic compound
Equilibrium electro-negativity
Molecular descriptor
Quantitative structure–property relationship (QSPR)

ABSTRACT

The novel electro-negativity topological descriptors of Y_C, W_C were derived from molecular structure by equilibrium electro-negativity of atom and relative bond length of molecule. The quantitative structure–property relationships (QSPR) between descriptors of Y_C, W_C as well as path number parameter P_3 and the normal boiling points of 80 alkanes, 65 unsaturated hydrocarbons and 70 alcohols were obtained separately. The high-quality prediction models were evidenced by coefficient of determination (R^2), the standard error (S), average absolute errors (AAE) and predictive parameters (Q^2_{ext}, R^2_{CV}, R^2_m). According to the regression equations, the influences of the length of carbon backbone, the size, the degree of branching of a molecule and the role of functional groups on the normal boiling point were analyzed. Comparison results with reference models demonstrated that novel topological descriptors based on the equilibrium electro-negativity of atom and the relative bond length were useful molecular descriptors for predicting the normal boiling points of organic compounds.

1. Introduction

The normal boiling point temperature (BP) can be defined as the temperature at which the vapour pressure of a pure liquid reaches 760 mmHg. The normal boiling point is usually used in the estimation of many other properties such as critical temperatures, flash points and enthalpies of vaporization, etc. [1–3]. Accordingly, boiling point is one of the most important thermal properties used to characterize and identify compounds. It depends on the intermolecular interactions in the liquid and on the differences in the molecular internal partition function between the gas and liquid phase [4]. Therefore, the normal boiling point can be directly related to the chemical structure of the molecule. The direct measurement of normal boiling point of organic compounds is extremely laborious, but experimental values of many compounds are unavailable in literatures, and their measurement is costly and time-consuming as it requires pure compounds. Also high molecular weight compounds decompose prior to reaching their normal boiling points and necessitate measurements under reduced pressure and later corrections to ambient pressure leading to errors [5]. In view of these shortcomings, it is essential to develop reliable methods for estimating normal boiling points of the compounds is essential. The classical approaches to estimate boiling points of compounds are based on their structures, including function methods, group contribution methods and quantitative structure–property relationships [6–12]. Earlier function methods for the estimation of boiling points employed physical parameters such as parachor and molar refractivity. But the function methods may be complex and require additional chemical properties and simplified assumptions to complete the calculation [13,14]. Efforts were also made to estimate boiling points by group contribution methods based on the assumption that cohesion forces in liquids are predominantly short-range and proceed from the division of a molecule into predefined structural groups, each of which adds a constant increment to the value of the property [1]. However, the group contribution methods do not take into account interactions between different groups or the role of the group position inside the structure [15]. Even more sophisticated methods are not comprehensive enough to cover multiple substitutions of functional groups [16]. Accordingly, the function methods and the group contribution methods are not adequately accurate. Quantitative structure–property relationships (QSPR) are important complementary tools in computational chemistry to represent explain and, most importantly, predict a variety of physicochemical, industrial and environmental properties [17–21]. With the increased need for reliable data for optimization of industrial processes, it is important to develop reliable QSPR models for the estimation of normal boiling points for compounds not yet synthesized or whose boiling points are unknown. In recent years, several methods on QSPR were developed for the correlation and prediction of boiling points [22–25]. Katritzky et al. focused on predicting the boiling point for a huge set of 612 organic compounds containing C, N, O, S, F, Cl, Br and I atoms based on molecular descriptors calculated solely from structure using CODESSA PRO. The general

* Corresponding author. Tel.: +86 13786140837; fax: +86 7315258733.
E-mail address: yimindai@sohu.com (Y.-m. Dai).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.04.007

performance of the final eight parameters model is good and the results show that sulfur-containing compounds are quite difficult to be predicted, and lead to a lower model performance [22]. Srinivasan et al. estimated normal boiling points of trialkyl phosphates using retention indices by gas chromatography. The results show that measurement of retention index is a reliable, simple and fast method for estimating normal boiling points of trialkyl phosphates [23]. Goll and Jurs developed a suite of methods for the prediction of boiling points based on molecular descriptors such as the Wiener index and surface charge areas [24]. However most of them could not reveal the real connection among atoms, and are not suitable for heteroatom-containing and multiple bond organic compounds.

The aim of this work was to establish new QSPR models for predicting the boiling points of 80 alkanes, 65 unsaturated hydrocarbons and 70 alcohols and to find the structural factors which affect the boiling points of studied compounds. Novel topological descriptors Y_C, W_C based on equilibrium electro-negativity and relative bond length together with path number P_3 were used to develop multiple linear regression (MLR) models for estimating and predicting the boiling points of alkanes, unsaturated hydrocarbons, and alcohols. The stability and predictive power of these models were validated using leave-one-out cross-validation and external test.

2. Materials and methods

2.1. Experimental data set

Experimental data sets of the normal boiling points of 215 organic compounds are obtained from literature [26]. The data sets include alkanes, alkenes, alkynes, dienes, aromatics, alcohols.

In supplementary material tables all the chemicals in the experimental data sets, molecular descriptor values, experimental and predicted boiling points are listed.

2.2. Molecular descriptors

Electro-negativity is one of the most important properties used to characterize atoms, which represents the ability of atoms to obtain or lose electrons. The larger electro-negativity of an atom has, the stronger the ability of the atom attracts electrons. Based on the Pauling electro-negativity, the group electro-negativity χ_G can be calculated by the step-wise addition method, which can be expressed as follows [27–30]:

$$\chi_0 = \frac{1}{n_{1l}}\sum_{l=1}^{n_{1l}}\chi_{1l} \quad \text{the equilibrium of the first level}$$

$$\chi_{1l} = \frac{1}{n_{2l}}\sum_{l=1}^{n_{2l}}\chi_{2l} \quad \text{the equilibrium of the second level}$$

$$\vdots \qquad\qquad \vdots$$

$$\chi_{(k-1)l} = \frac{1}{n_{kl}}\sum_{l=1}^{n_{kl}}\chi_{kl} \quad \text{the equilibrium of the} k\text{th level}$$

$$\vdots \qquad\qquad \vdots$$

Then group electro-negativity χ_G is defined as:

$$\chi_G = \left\{\frac{1}{n_{1l}}\sum_{l=1}^{n_{1l}}\left[\frac{1}{n_{2l}}\sum_{l=1}^{n_{2l}}\cdots\left(\frac{1}{n_{il}}\sum_{l=1}^{n_{il}}\chi_{il}\right)\cdots\right]\right\} \quad (1)$$

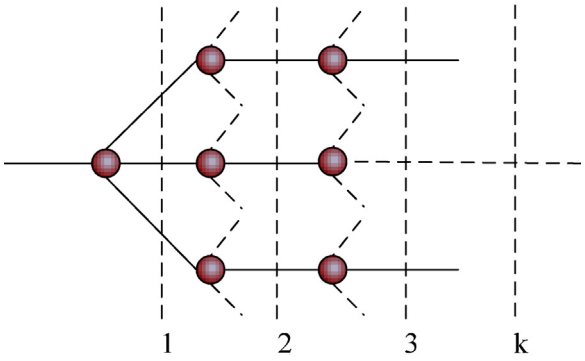

Fig. 1. Plot of group structure tree.

where $n_{1l}, n_{2l}\ldots n_{kl}$ are the sum of atom or group directly attached to the ground atom, which is the left atom next to the dotted line labeled 1, 2, 3,..., k... of each level in Fig. 1, and $\sum_{l=1}^{n_{1l}}\chi_{1l}, \sum_{l=1}^{n_{2l}}\chi_{2l}, \ldots, \sum_{l=1}^{n_{2l}}\chi_{2l}$ are the sum of electro-negativity of atom or group directly attached to the ground atom. We think that electro-negativity randomly changes in the formation of a molecule. As long as a molecule is formed, the electro-negativity of atom in the molecule is fixed, that is to say, the electro-negativity is in the state of equilibrium, which is called the equilibrium electro-negativity of atom [28]. The definition of equilibrium electro-negativity for atom i is defined as following:

$$\chi_i = \frac{\chi_{iA} + \sum \chi_G}{1 + \sum l} \quad (2)$$

where χ_{iA} is the Pauling electro-negativity for atom i, χ_G is the electro-negativity of group directly attached to atom i calculated by Eq. (1), and l is the group number directly attached to atom i. Equilibrium electro-negativity χ_i can efficiently characterize the electro-negativity of each atom in a molecule, and the equilibrium electro-negativity of atom can closely reflect atomic the chemical environment. Therefore, the equilibrium electro-negativity can effectively reflect the chemical information at the atom and group levels [29,30]. Accordingly, the equilibrium electro-negativity matrix $\boldsymbol{X}$ is defined to reflect every atomic chemical environmental change of a molecule. It is worth mentioning that χ_i can make up for the absence of hydrogen atom in a carbon skeleton graph because the equilibrium electro-negativity of hydrogen atoms is fully taken into account.

In addition, the relative bond length of two adjacent vertices is used to distinguish saturated, unsaturated bond and heteroatom compounds. Here, L_{ij} is the shortest distance between vertices i and j, and is calculated by summing the bond length between two adjacent vertices in the shortest path. If employ the C—C bond length $L_{C-C} = 0.154$ nm is taken as 1, then the relative bond length between vertices i and j is calculated: $d_{ij} = \Sigma L_{ij}/L_{C-C}$, for example, C=O relative bond length is $0.122/0.154 = 0.7922$ [31].

The equilibrium electro-negativity of atom χ_i and the relative bond length d_{ij} between two adjacent vertices are used to elucidate the properties and interaction of vertices in a molecule. The distance matrix $\boldsymbol{D}$ of n atoms in a molecule, a symmetric matrix, can be expressed as: $\boldsymbol{D} = [d_{ij}]_{n\times n} = [\Sigma L_{ij}/L_{C-C}]_{n\times n}$, where d_{ij} is the relative bond length of the shortest path between the vertice i and vertice j. In addition, addition matrix $\boldsymbol{S}$, vertex matrix $\boldsymbol{R}$ and electro-negativity matrix $\boldsymbol{X}$ are defined in order to distinguish the length of carbon backbone, the size, the degree of branching and the charge distribution of a molecule from the atomic species, respectively. According to corresponding definition matrixes $\boldsymbol{D}$, $\boldsymbol{S}$, $\boldsymbol{R}$ and $\boldsymbol{X}$ are

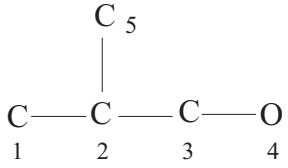

Fig. 2. The carbon backbone graph of 2-methyl-1-propanol.

expressed as follows:

$$\boldsymbol{D} = \begin{bmatrix} 0 & L_{12}/L_{C-C} & \cdots & L_{1(n-1)}/L_{C-C} & L_{1n}/L_{C-C} \\ L_{21}/L_{C-C} & 0 & \cdots & L_{2(n-1)}/L_{C-C} & L_{2n}/L_{C-C} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ L_{i1}/L_{C-C} & \cdots & 0 & \cdots & L_{in}/L_{C-C} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ L_{n1}/L_{C-C} & L_{n2}/L_{C-C} & \cdots & L_{n(n-1)}/L_{C-C} & 0 \end{bmatrix}$$

$$\boldsymbol{S}_i = \sum_{j=1}^{n} d_{ij} \qquad \boldsymbol{S} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

Taking into them account, two novel electro-negativity topological descriptors Y_C and W_C are respectively defined as:

$$Y_C = \log \sum_{j}^{N} \left(s_j \cdot r_j \cdot \chi_j\right) \tag{3}$$

$$W_C = \frac{1}{P_n^2} \sum_{j}^{n} s_j \tag{4}$$

Path number was initially put forward by Wiener, which could effectively reflect chemical information of the degree of branching and shape of a molecule [10]. Here we use P_3 which is defined as the distance between any two vertices is equal to 3. For example, the molecular carbon skeleton graph and revised distance matrix of 2-methyl-1-propanol are expressed as follows (Fig. 2):

The distance matrix $\boldsymbol{D}$, addition matrix $\boldsymbol{S}$, vertices matrix $\boldsymbol{R}$ and the equilibrium electro-negativity matrix $\boldsymbol{X}$ are respectively expressed as following:

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1.0000 & 2.0000 & 2.9286 & 2.0000 \\ 1.0000 & 0 & 1.0000 & 1.9286 & 1.0000 \\ 2.0000 & 1.0000 & 0 & 0.9286 & 2.0000 \\ 2.9286 & 1..9286 & 0.9286 & 0 & 2.9286 \\ 2.0000 & 1.0000 & 2.0000 & 2.9286 & 0 \end{bmatrix}$$

$$\boldsymbol{S} = \begin{bmatrix} 7.9286 \\ 4.9286 \\ 5.9286 \\ 8.7144 \\ 7.9286 \end{bmatrix} \qquad \boldsymbol{R} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 1 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 2.3040 \\ 2.3535 \\ 2.4203 \\ 2.6534 \\ 2.3040 \end{bmatrix}$$

Consequently, topological indices Y_C and W_C for 2-methyl-1-propanol according to the above definition are calculated as:

$$Y_C = \log(7.9286 \times 1 \times 2.3040 + 4.9286 \times 3 \times 2.3535 + 5.9286 \times 2 \times 2.4203 + 8.7144 \times 1 \times 2.6534 + 7.9286 \times 1 \times 2.3040) = 2.0904$$

$$W_C = \frac{7.9286 + 4.9286 + 5.9286 + 8.7144 + 7.9286}{5 \times 4} = 1.7714$$

$$P_3 = 2$$

2.3. Model validation

Validation of QSPR model is a very important aspect. It is acknowledged that the three aspects of goodness-of-fit, stability, and predictive power are all very important for QSPR models [32]. The quality of goodness-of-fit of the models is quantified by the coefficient of determination (R^2), the standard error (S), the Fisher statistic value (F) and average absolute error (AAE). The coefficient of determination is reported as a measure of the total variance of the response explained by the regression models. The standard error indicates dispersion degree of random error. The larger R^2 and F, the smaller S, and the model will have more fitting ability. However, good fitting ability does not stand for good robustness and predictive ability, thus internal validation is considered to be necessary for model validation [33,34]. The most popular validation criterion to explore the robustness of a predictive model is through the analysis of each one of individual objects that configure the final equation. This procedure is known as leave-one-out (LOO) cross-validation (CV). The leave-one-out cross-validations were performed in training test. The R^2_{CV} describes the stability of a regression model obtained by focusing on sensitivity of the model to the elimination of any single data point, which is defined as the following Eq. (5) [35]:

$$R^2_{CV} = 1 - \frac{\sum_{i=1}^{\text{training}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{training}} (y_i - \bar{y}_{tr})^2} \tag{5}$$

where, $y_i, \hat{y}_i$ and, $\bar{y}_{tr}$ are the experimental, predicted, and the mean boiling point values of the training set, respectively. It is worth mentioning that the models have excellent fitting ability and stability, yet, it cannot guarantee the true predictive ability of the models. Therefore, to assess such predictive power the use of an external validation is essential. Roy et al. pointed out that the external validation was an indispensable validation method used to determine the true predictive ability of the QSPR model [35–37]. And the predictive ability of a model on external validation set can be expressed by Q^2_{ext} by the following Eq. (6) [35]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{\text{prediction}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{prediction}} (y_i - \bar{y}_{tr})^2} \tag{6}$$

where, y_i and $\hat{y}_i$ are the experimental and predicted boiling point values of the validation set, respectively, and $\bar{y}_{tr}$ is the mean boiling point values of the training set.

Thus, an additional parameter R^2_m, which penalizes a model for large differences between observed and predicted values of the prediction set compounds, was also calculated for model external prediction [35–38]. The expression of R^2_m is defined by the following Eq. (7):

$$R^2_m = R^2 \left(1 - \sqrt{R^2 - R_0^2}\right) \tag{7}$$

R^2 and R_0^2 are determination coefficients of linear regression equation between the observed and predicted set compounds with and without intercept, respectively.

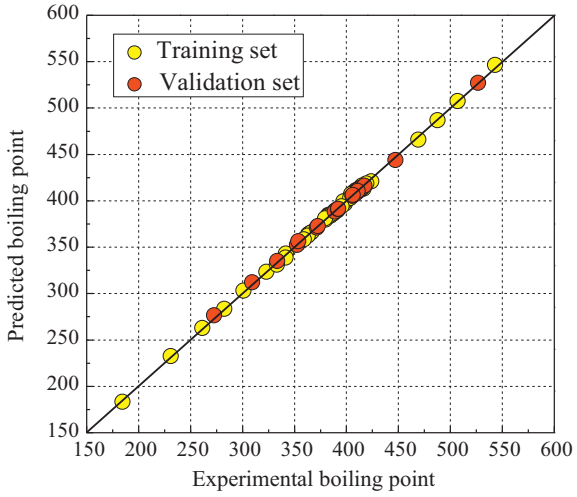

Fig. 3. Plot of predicted versus experimental values of alkanes boiling points.

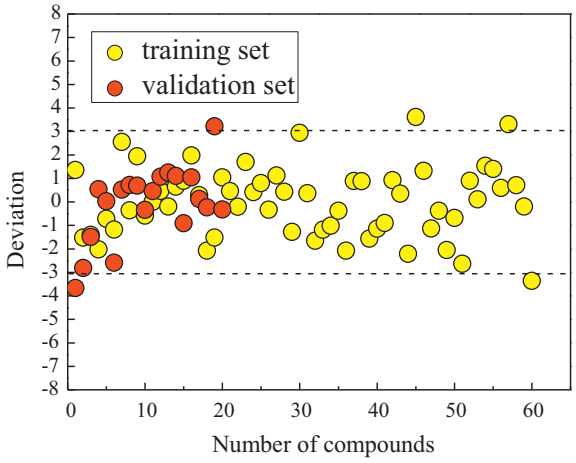

Fig. 4. Residuals plot of predicted versus experimental values of alkanes boiling points.

3. Results and discussion

3.1. With alkanes

In order to build a QSPR model, the dataset was randomly divided into two subsets, namely, the training set and the validation set. For alkanes series, 60 alkanes were selected as the training set in the modeling, the remaining 20 chemicals were used for external validation set. Making use of the multiple linear regression method, the following linear model was obtained, in which the molecular descriptors Y_C, W_C and P_3 were used as independent variables:

$$\begin{aligned} &BP = (68.6864 \pm 1.2129)\,Y_C + (23.9021 \pm 0.5879)\,W_C + (4.9672 \pm 0.1043)\,P_3 + (107.2374 \pm 1.6815) \\ &n = 60, \quad R^2 = 0.9993, \quad Q^2_{ext} = 0.9995, \quad S = 1.5215, \quad F = 26,800.92, \quad R^2_{CV} = 0.9991 \end{aligned} \tag{8}$$

The statistical results indicate that the model coefficient of determination, the standard error and the Fisher statistic value are 0.9993, 1.5215 and 26,800.92, respectively. And the average absolute error is only 0.98 K between the predicted and experimental boiling points for 60 alkanes compounds. The coefficient of determination (R^2) and predictive parameter (R^2_{CV}) of the model appear good. The result suggests our quantitative structure property relationship model is satisfactory. The scatter plot of the predicted against the experimental boiling points is plotted in Fig. 3. Fig. 3 indicates a good correlation between the experimental and predicted values. The residuals of predicted of boiling points against the experimental values are plotted in Fig. 4. As the predicted residuals are distributed on both sides of the zero line, the residuals exceed seldom the standard deviation of $\pm 2S$. Therefore, one may conclude that there is not systematic error in the development of MLR model. The external predictive power is confirmed by a high Q^2_{ext} value (0.9995) that reveals model applicability also to predict the boiling points of unknown alkanes belonging to its chemical domain. This result is even more relevant considering that the model was strongly externally validated on a number of chemicals equivalent to that included in the training set.

In addition, R^2_m value helps to identify the best model when different models show different patterns in internal and external predictivity. In QSPR models, the recommended criterion is that R^2_m value should be >0.5 [37]. For the present QSPR study, R^2_m for both training and external validation set are 0.9852, 0.9822, respectively, all the models pass the criterion for R^2_m statistic.

3.2. With unsaturated hydrocarbons

The datasets of boiling points of unsaturated hydrocarbons containing alkene, alkyne, diene and aromatic hydrocarbon were taken from the literature [26]. Since there were many molecules for calibrating the model, the whole set was split randomly into two parts, namely, the training set containing 50 compounds to develop the model and the external validation set of remained 15 compounds to evaluate the model performance, according to previous method. For the complete training set of 50 compounds the following QSPR model was obtained:

$$\begin{aligned} &BP = (61.0548 \pm 4.0296)\,Y_C + (24.5478 \pm 2.9283)\,W_C + (5.7240 \pm 0.7979)\,P_3 + (122.4451 \pm 5.3434) \\ &n = 50, \quad R^2 = 0.9910, \quad Q^2_{ext} = 0.9856, \quad S = 6.4564, \quad F = 1687.25, \quad R^2_{CV} = 0.9875. \end{aligned} \tag{9}$$

The statistical results show that molecular descriptors Y_C, W_C and P_3 have good correlation with boiling points in terms of the coefficient of determination, the standard error and the Fisher statistic value. The analysis of plots has shown to be quite useful to detect anomalies or confirm the quality of a model. Fig. 5 shows that the calculated versus experimental boiling points obtained with Eq. (9) follows a straight line. And the average absolute error is only 1.62 K. Fig. 6 shows the dispersion as a function of the predicted property. Horizontal lines in this figure indicate the standard deviation limits of $\pm 2S$. The residuals exceed seldom the standard deviation of $\pm 2S$ from Fig. 6. Accordingly, from Figs. 5 and 6 and the statistical results of Eq. (9), it can be concluded that the 3-variable-model is excellent.

The model cross-validated R^2_{CV} values ($R^2_{CV} = 9875$) are very close to the corresponding R^2 value ($R^2 = 0.9910$). Clearly, the cross-validation demonstrates the final model to be statistically significant. The external predictive power is confirmed by a high Q^2_{ext} value ($Q^2_{ext} = 0.9856$) that reveals model applicability also to predict the boiling points of unknown series compounds.

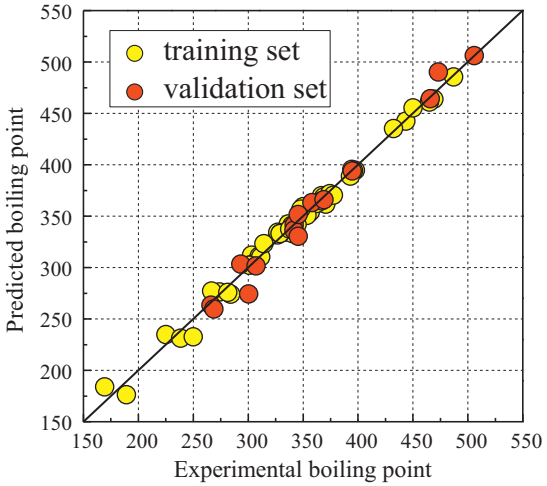

Fig. 5. Plot of predicted versus experimental values of unsaturated hydrocarbons boiling points.

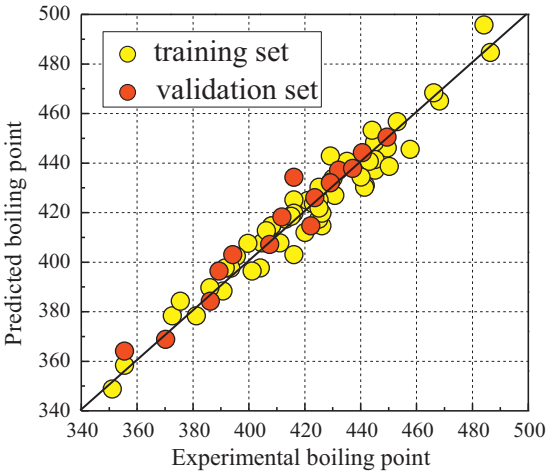

Fig. 7. Plot of predicted versus experimental values of alcohol boiling points.

3.3. With alcohols

Here a general data set of 70 alcohols was considered. These data sets were randomly divided into two subsets, one containing 55 alcohols was used as the training set in the modeling, and the other containing 15 alcohols was used for external validation set. Using MLR method, a linear relationship model between boiling point values of alcohols and the three optimal descriptors was developed as follows:

$$\begin{aligned} &BP = (1.6085 \pm 11.7109)Y_C + (49.8169 \pm 5.6986)W_C + (2.0728 \pm 0.8397)P_3 + (282.5385 \pm 14.8904) \\ &n = 55, R^2 = 0.9485, Q^2_{ext} = 0.9520, S = 6.8623, F = 296.44, R^2_{CV} = 0.9373 \end{aligned} \quad (10)$$

For our QSPR model, the cross-validated correlation coefficient $R^2_{CV} = 0.9373$, as compared to the coefficient of determination $R^2 = 0.9458$, indicates the high stability of the regression equation. The scatter plot of the boiling point predicted using the model versus experimental boiling point is presented in Fig. 7. Fig. 7 shows the regression line of the above proposed model with very satisfactory performances.

In order to investigate the error distribution we present the plot of the residuals of predicted boiling points against the experimental values as shown in Fig. 8. As can be seen from Fig. 8 the residuals show random dispersion. Therefore, it is in agreement with the general multiple linear theory. And parameter $R^2_m = 0.8586$, $Q^2_{ext} = 0.9520$ it can be concluded that the QSPR model based on this new topological index has good predictive stability and reliability for predicting the boiling point of alcohols.

3.4. Model interpretation

The descriptors involved in Eqs. (8)–(10) are electro-negativity topological descriptors of Y_C, W_C and path number parameter P_3. It can be observed that the descriptor of Y_C play an important role in determining boiling points of saturated and unsaturated hydrocarbons. The Y_C index calculated from the hydrogen-suppressed graph of the molecule, encoded information about the length of carbon backbone, the size and the degree of branching of a molecule. As the size of index Y_C grew bigger, volume of a molecule increase drastically, intermolecular interactions become stronger accordingly, the boiling point value is higher. The second descriptor W_C used is connected with functional groups information of compounds. The electrostatic descriptor W_C can effectively offer information about binding and formation energies, the charge distribution of a molecule, dipole moment and molecular orbital energy level. According to the electro-negativity equilibrium theory, the charge on the bond is proportional to the molecular

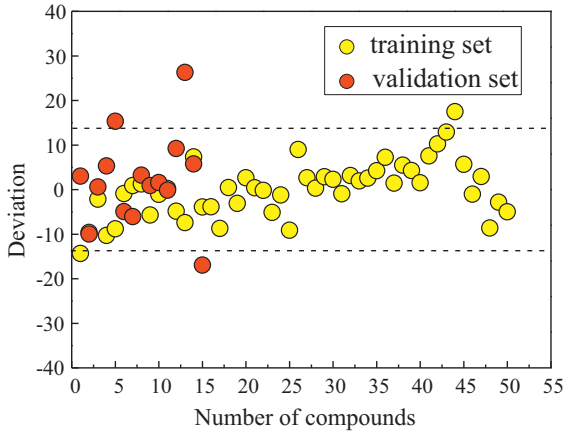

Fig. 6. Residuals plot of predicted versus experimental values of unsaturated hydrocarbons boiling points.

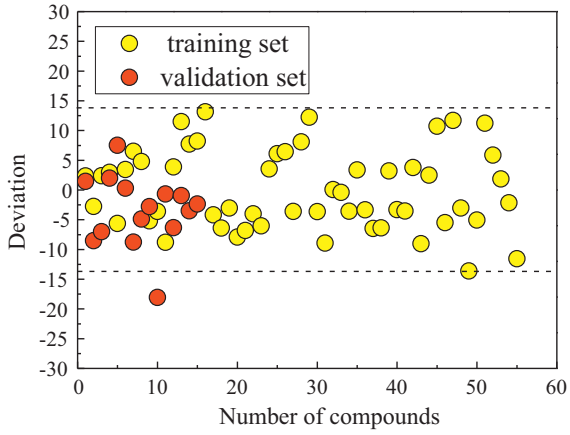

Fig. 8. Residuals plot of predicted versus experimental values of alcohol boiling points.

Table 1
Comparative statistical performances of different developed models.

Type of compound	N	Descriptors	Method[a]	R^2	R^2_{CV}	S	Reference
Alkanes	94	Tls(W, P)	MLR			0.97	Wiener [10]
Alkanes	74	Tls(5)	MLR	0.999		1.86	Needham [42]
Alkanes (C_2–C_7)	72	Tls(LOVI's)	RA	0.994		3.9	Balaban et al. [43]
Furans, thiophenestetrahydrofurans	209	Tls, electronic, geometrical	RA	0.969		11.2	Stanton [44]
Alkanes, alcohols	245	Atom-type E	MLR			8	Hall et al. [45]
Alcohols	58	Weighted path related Tls	MRA	0.978		3.64	Randić et al. [46]
Diverse organic compounds	298	Constitutional, topological, geometrical and CPSA(8)	MLR	0.954	0.953	16.15	Katritzky et al. [47]
Hydrocarbons,	143	Tls	MLR	0.9821	0.9816	7.2549	Zhou et al. [41]
Alkanes	80	Molecualr descriptor	MLR	0.9993	0.9991	1.5215	This work
Unsaturated	65			0.9910	0.9875	6.4564	
Hydrocarbons		(Y_C, W_C, P_3)					
Alcohols	70			0.9458	0.9373	6.8623	

[a] MLR: multiple linear regression; RA: regression analysis; PCA: principal analysis; QR: quadratic regression; Tls: topological indices.

electro-negativity. It induces electron redistribution in neighbor molecules and thereby establishes dipole: inductive dipole attraction [39]. The larger the intermolecular inductive dipole attraction is, the higher the boiling point value is. Position factor of group could distinguish most isomers. Nevertheless, the position of the structural group in the molecule may strongly influence the properties of the compounds. When an atom with a lone pair is able to form H-bond, it gets a partial positive charge [40]. The descriptor W_C of alcohols belongs to the class of the charged partial surface area descriptors which combine shape, electronic information and the hydrogen-bonding ability to characterize the polar interactions between molecules. Path number parameter P_3 can effectively elucidate chemical information of degree of branching and shape of a molecule. Accordingly, alkanes are nonpolar compounds only having dispersion forces between molecules. The dispersion forces change positively with increase of carbon number. However, the increasing rate decreases gradually. Similarly, Y_C and W_C increase with carbon number but the changing rate gets lower, which show the same regularity as the situation of the dispersion. For alcohols, the interaction forces between molecules include hydrogen-bonding forces, orientational forces and induction forces, which make the interaction forces between molecules more complicated [41]. The results reveal that the normal boiling points of alcohols are predominantly decided by molecular descriptor of W_C. Descriptor Y_C characterizing molecular size and degree of branching has minor contribution to normal boiling point for alcohols. Comparison of the results indicates that the alkanes and unsaturated hydrocarbons have better predictive power than the alcohols of boiling point using electro-negativity topological descriptors of Y_C, W_C and path number parameter P_3.

All the molecular descriptors involved in the regression models, which have explicit physical meanings, may account for the structural features responsible for the boiling point of organic compounds.

3.5. Comparison with published models

The models in this paper are now compared with those previously reported by other groups [41–47]. The statistical qualities of the different published models and current models are listed in Table 1. Comparative comments can be made, although it is impossible to make a perfect comparison of the published models, as different data sets and different algorithms were used for model building and validation. It is interesting to note that the descriptors selected in different models, mainly in those obtained from training sets similar in dimension and typology, have comparable structural and physical meanings.

Aside from simple correlations of boiling points with the carbon number or molecular weight for homologous series of compounds, Wiener was first to correlate boiling points based topological descriptors of the Wiener index (W) and Wiener polarity index (P) [10]. Based on these indices, he predicted the boiling points of alkanes with an average error of 1 °C. Boiling points of organic compounds depend on individual atoms or groups, polarizability of the molecule, length of carbon backbone, shape and branching of the molecule and their contribution to the structural environment in the molecule. For example most of the conventional topological indices such as Balaban index, Randić and Hall molecular connectivity indices are found unsuccessful for compounds with multiple bonds and heteroatoms, since these indices do not take into account the contribution of each of the individual atom types or groups [42–47]. Katritzky et al. combined a QSPR approach with molecular descriptors extracted from AM1 semi-empirical calculations to fit the boiling points of a training set of 298 organic compounds, including saturated and unsaturated hydrocarbons, amino, ester, hydroxyl and halogenated compounds. The two descriptors linear equation showed R^2 of 0.954 and squared cross-validation coefficient R^2_{CV} of 0.953 with the root mean square error of 16.15 K. This returned an overall R^2 value of 0.973 and an average prediction error of 2.3% with the use of four parameters [47].

It should be stressed the new topological descriptors Y_C, W_C proposed in this work is composed of two parts. The first part is the molecular distance matrix by which the structure of molecules could be described objectively and quantitatively. The second part is the extended matrix from which the structure and the composition of molecules could be further identified by the equilibrium electro-negativity and the relative bond length. Our models, the structure and composition of molecules could be determined accurately and completely. From the above discussion, it could be demonstrated that our method with novel topological indices could result in significant improvements both in accuracy and in stability for predicting boiling points of organic compounds. Another, molecular descriptors involved in the regression models which have explicit physical meaning. It is combination of the equilibrium electro-negativity and the relative bond length based on the distance matrix that leads to the good application of electro-negativity topological descriptors of Y_C, W_C and path number parameter P_3 to study on normal boiling points of organic compounds.

4. Conclusions

Novel electro-negativity topological descriptors of Y_C, W_C were proposed to be characterized the structure and the composition of heteroatom-containing and multiple bond organic compounds by the equilibrium electro-negativity and the relative bond length of molecules. The QSPR models were developed for 80 alkanes, 65 unsaturated hydrocarbons and 70 alcohols using electro-negativity topological descriptors of Y_C, W_C and path number parameter P_3.

The final models were validated to be statistically reliable and predictive using the leave-one-out cross validation and an external test set. Comparison results with reference models demonstrate that this new method is very efficient and provides satisfactory results in significant improvements, both in accuracy and stability for predicting the normal boiling points of organic compounds.

Acknowledgements

The authors wish to express their grateful thanks to National Natural Science Foundation of China (21275022, 21003014), the Foundation of Hunan Province Science and Technology Department (2012GK3058); the Foundation of Hunan Provincial Key Laboratory of Materials Protection for Electric Power and Transportation (2010CL01); the Doctoral Foundation of Changsha University of Science and Technology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2013.04.007.

References

[1] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, Chemical Reviews 110 (2010) 5714–5789.
[2] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative structure–property relationship modeling of diverse materials properties, Chemical Reviews 112 (2012) 2889–2919.
[3] W.J. Lyman, W.F. Reechl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC, 1990.
[4] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. Gonzalez, A new search algorithm for QSPR/QSAR theories: normal boiling points of some organic molecules, Chemical Physics Letters 412 (2005) 376–380.
[5] H.D. Li, H. Higashi, K. Tamura, Estimation of boiling and melting points of light, heavy and complex hydrocarbons by means of a modified group vector space method, Fluid Phase Equilibria 239 (2006) 213–222.
[6] M.R. Riazi, T.A. Al-Sahhaf, Physical properties of heavy petroleum fractions and crude oils, Fluid Phase Equilibria 117 (1996) 217–224.
[7] P.Y. Chan, C.M. Tong, M.C. Durrant, Estimation of boiling points using density functional theory with polarized continuum model solvent corrections, Journal of Molecular Graphics and Modelling 30 (2011) 120–128.
[8] J.A. Palatinus, C.M. Sams, C.M. Beeston, F.A. Carroll, A.B. Argenton, F.H. Quina, Kinney revisited: an improved group contribution method for the prediction of boiling points of acyclic alkanes, Industrial and Engineering Chemistry Research 45 (2006) 6860–6863.
[9] Q. Wang, Q.Z. Jia, P.S. Ma, Prediction of the acentric factor of organic compounds with the positional distribution contribution method, Journal of Chemical and Engineering Data 57 (2012) 169–189.
[10] H. Wiener, Structural determination of paraffin boiling points, Journal of the American Chemical Society 69 (1947) 17–20.
[11] S.H. Kumar, A comparative QSPR study of alkanes with the help of computational chemistry, Bulletin of the Korean Chemical Society 29 (2008) 67–76.
[12] D. Sola, D.A. Ferri, M. Banchero, L. Manna, S. Sicardi, QSPR prediction of N-boiling point and critical properties of organic compounds and comparison with a group-contribution method, Fluid Phase Equilibria 263 (2008) 33–42.
[13] C.E. Rechsteiner, in: W.J. Lyman, W.F. Reehl, D.H. Rosenblatt (Eds.), Handbook of Chemical Property Estimation Methods, McGraw-Hill, New York, 1982.
[14] A.L. Horvath, Molecular Design: Chemical Structure Generation from the Properties of Pure Organic Compounds, Amserdam, Elseiver, 1992.
[15] K.G. Joback, R. Reid, Estimation of pure-component properties from group-contributions, Chemical Engineering Communications 57 (1987) 233–243.
[16] X. Wen, Y. Qiang, Group vector space method for estimating melting and boiling points of organic compounds, Industrial & Engineering Chemistry Research 41 (2002) 5534–5537.
[17] R.D. Cramer III, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins, Journal of the American Chemical Society 110 (1988) 5959–5967.
[18] Y.M. Alvarez-Ginarte, Y. Marrero-Ponce, J.A. Ruiz-García, L.A. Montero-Cabrera, J.M. Garcia-de la Vega, P. Noheda-Marin, R. Crespo-Otero, F. Torrens-Zaragoza, R. García-Domenech, Applying pattern recognition methods plus quantum and physico-chemical molecular descriptors to analyze the anabolic activity of structurally diverse steroids, Journal of Computational Chemistry 29 (2008) 317–333.
[19] G. Gece, The use of quantum chemical methods in corrosion inhibitor studies, Corrosion Science 50 (2008) 2981–2992.
[20] A.N. Gorban, G.S. Yablonsky, Extended detailed balance for systems with irreversible reactions, Chemical Engineering Science 66 (2011) 5388–5399.
[21] S. Ajmani, A. Agrawal, S.A. Kulkarni, A comprehensive structure–activity analysis of protein kinase B-alpha (Akt1) inhibitors, Journal of Molecular Graphics and Modelling 28 (2010) 683–694.
[22] A.R. Katritzky, V.S. Lobanov, M. Karelson, Normal boiling points for organic compounds: correlation and prediction by a quantitative structure–property relaionship, Journal of Chemical Information and Computer Science 38 (1998) 28–41.
[23] K. Panneerselvam, M.P. Antony, T.G. Srinivasan, P.R. Vasudeva Rao, Estimation of normal boiling points of trialkyl phosphates using retention indices by gas chromatography, Thermochimical Acta 511 (2010) 107–111.
[24] E.S. Goll, P.C. Jurs, Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model, Journal of Chemical Information and Computer Science 39 (1999) 974–983.
[25] W.Q. Liu, C.Z. Cao, Quantitative structure–property relationship of normal boiling point of aliphatic oxygen-containing organic compounds, CIESC Journal 63 (2012) 3739–3746.
[26] D.R. Lide, CRC Handbook of Chemistry and Physics, 83rd ed., CRC Press, Boca Raton, FL, 2002–2003.
[27] C.M. Nie, Group electro-negativity, Journal of Wuhan University (Nat. Sci. Ed.) 46 (2000) 176–180.
[28] C.M. Nie, Y.M. Dai, S.N. Wen, Z.H. Li, C.Y. Zhou, G.W. Peng, Topological homologous regularity for additive property of alkanes, Acta Chimica Sinica 63 (2005) 1449–1455.
[29] Y.M. Dai, X. Li, Z. Cao, D.W. Yang, K.L. Hang, Modeling flash point scale of hydrocarbon by novel topological electro-negativity indices, CIESC Journal 60 (2009) 2420–2425.
[30] C. Zhou, X. Chu, C. Nie, Predicting thermodynamic properties with a novel semiempirical topological descriptor and path numbers, Journal of Physical Chemistry B 111 (2007) 10174–10179.
[31] Y.M. Dai, Atomic equilibrium electro-negativity and its application research in molecular design and molecular modeling, Central South University, Changsha, 2012 (PhD Dissertation).
[32] J.J. Shi, L.P. Chen, W.H. Chen, N. Shi, H. Yang, W. Xu, Prediction of the thermal conductivity of organic compounds using heuristic and support vector machine methods, Acta Physico-Chimica Sinica 28 (2012) 2790–2796.
[33] V. Consonni, D. Ballabio, R. Todeschini, Evaluation of model predictive ability by external validation techniques, Journal of Chemometrics 24 (2010) 194–201.
[34] A. Tropsha, P. Gramatica, V. Gpmbar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR, QSAR & Combinatorial Science 22 (2003) 69–77.
[35] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Further exploring r^2_m metrics for validation of QSPR models, Chemometrics and Intelligent Laboratory Systems 107 (2011) 194–205.
[36] K. Roy, I. Mitra, S. Kar, Lattice enumeration for inverse molecular design using the signature descriptor, Journal of Chemical Information and Modeling 52 (2012) 1787–1797.
[37] P.P. Roy, K. Roy, On some aspects of variable selection for partial least squares regression models, QSAR & Combinatorial Science 27 (2008) 302–313.
[38] P.P. Roy, S. Kovarich, P. Gramatica, QSAR model reproducibility and applicability: a case study of rate constants of hydroxyl radical reaction models applied to polybrominated diphenyl ethers and (benzo-) triazoles, Journal of Computational Chemistry 32 (2011) 2386–2396.
[39] Y.X. Wu, C.Z. Cao, H. Yuan, Estimation of the ionization potential for polyhalogenated hydrocarbons by weakest bound potential method, Journal of Physical Organic Chemistry 25 (2012) 110–117.
[40] M. Bortolotti, M. Brugnara, C.D. Volpe, D. Maniglio, S. Siboni, Molecular connectivity methods for the characterization of surface energetics of liquids and polymers, Journal of Colloid and Interface Science 296 (2006) 292–308.
[41] C.Y. Zhou, C.M. Nie, S. Li, Z.H. Li, A novel semi-empirical topological descriptor Nt and the application to study on QSPR/QSAR, Journal of Computational Chemistry 28 (2007) 2413–2423.
[42] D.E. Needham, I.C. Wei, P.G. Seybold, Molecular modeling of the physical properties of alkanes, Journal of the American Chemical Society 110 (1988) 4186–4194.
[43] A.T. Balaban, D. Ciubotariu, M. Medeleanu, Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors, Journal of Chemical Information and Computer Science 31 (1991) 517–523.
[44] D.T. Stanton, Development of a quantitative structure–property relationship model for estimating normal boiling points of small multifunctional organic molecules, Journal of Chemical Information and Computer Science 40 (2000) 81–90.
[45] L.H. Hall, L.B. Kier, Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information, Journal of Chemical Information and Computer Sciences 35 (1995) 1039–1045.
[46] M. Randić, A.T. Balaban, S. Basak, On structural interpretation of several distance related topological indices, Journal of Chemical Information and Computer Science 41 (2001) 593–601.
[47] A.R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson, Correlation of boiling points with molecular structure. 1. A training set of 298 diverse organic and a test set of 9 simple inorganics, Journal of Physical Chemistry 100 (1996) 10400–10407.