DOI:10.12921/cmst.2015.21.04.004

Prediction the Normal Boiling Points of Primary, Secondary and Tertiary Liquid Amines from their Molecular Structure Descriptors

Saadi Saaidpour[1*], **Asrin Bahmani**[2], **Amin Rostami**[2]

[1]*Department of Chemistry, Faculty of Science*
Sanandaj Branch, Islamic Azad University, Sanandaj, Iran

[2]*Department of Chemistry, Faculty of Science*
Kurdistan University, Sanandaj, Iran

[*]*E-mail: sasaaidpour@iausdj.ac.ir*

Received: 23 September 2015; revised: 19 November 2015; accepted: 28 November 2015; published online: 09 December 2015

Abstract: In this article, at first, a quantitative structure–property relationship (QSPR) model for estimation of the normal boiling point of liquid amines is developed. QSPR study based multiple linear regression was applied to predict the boiling points of primary, secondary and tertiary amines. The geometry of all amines was optimized by the semi-empirical method AM1 and used to calculate different types of molecular descriptors. The molecular descriptors of structures were calculated using Molecular Modeling Pro plus software. Stepwise regression was used for selection of relevance descriptors. The linear models developed with Molegro Data Modeller (MDM) allow accurate estimate of the boiling points of amines using molar mass (MM), Hansen dispersion forces (DF), molar refractivity (MR) and hydrogen bonding (HB) (1° and 2° amines) descriptors. The information encoded in the descriptors allows an interpretation of the boiling point studied based on the intermolecular interactions. Multiple linear regression (MLR) was used to develop three linear models for 1° , 2° and 3° amines containing four and three variables with a high precision root mean squares error, 15.92 K, 9.89 K and 15.76 K and a good correlation with the squared correlation coefficient 0.96, 0.98 and 0.96, respectively. The predictive power and robustness of the QSPR models were characterized by the statistical validation and applicability domain (AD).

Key words: liquid amines, boiling points, QSPR, MLR, prediction

I. INTRODUCTION

Fundamentally, an amine is a derivative of ammonia that centers around a single nitrogen atom. Amines are organic compounds and functional groups that contain a basic nitrogen atom with a lone pair. Amines are derivatives of ammonia, in which one or more hydrogen atoms have been replaced by a substituent such as an alkyl or aryl group. Amines are classified as primary, secondary, or tertiary, depending on the number of carbon atoms bonded directly to nitrogen. The substituent groups (R) may be alkyl or aryl. Another group of amines are those in which the nitrogen forms part of a ring (heterocyclic amines) [1].

The amines are used as flotation agents, anticacking agents, corrosion inhibitors, dispersants, emulsifiers and additives, and chemical intermediates.

Besides the amines of which the human body is composed (amino acids), humans have found a range of other uses for amines. Medicines based on amines such as Morphine and Demerol are commonly used as analgesics – medicines that relieve pain. Amines such as Novocaine are commonly used as anesthetics. The amine Ephedra is a common decongestant. Tetramethyl ammonium iodide is used in the disinfection of drinking water. Amines also have many other functions in an array of daily functions. Many amines are used in industries for pest control and tanning of leather. The amine Aniline finds application in the manufacturing of man-made dyes. Some amines, such as methamphetamines and amphetamines, are popular recreational drugs. Similar to ammonia, amines are basic in nature, which means that they have pH above seven. Due to this, they are neutralized by using acids. The process of their neutralization results in the formation of alkylammonium salts, having many industrial applications themselves. Choline, one of these salts, plays a role in the production of some neurotransmitters in the human body that make the brain work properly.

Knowledge of the physical properties of organic compounds is necessary for the design, development and manufacture of products in which they are used. The suitability of a particular compound for a given purpose depends on its physicochemical properties. Physicochemical properties of organic compounds change in a systematic way with changes in the chemical structure, which actually determine the type and magnitude of intra-molecular and intermolecular interactions [2].The boiling-point (BP) of a pure liquid is defined as the temperature at which the vapor pressure of the liquid exactly equals the pressure exerted on it by the atmosphere. When the external pressure is 1 atmosphere, the boiling temperature is called the normal boiling-point. The BP of a compound is an important property, and like the vapor pressure it provides an indication of the attractive forces between the molecules. These intermolecular forces are directly related to the structure of the compound, and therefore the boiling-point may be correlated to the structure. The higher a compound's normal boiling-point, the less volatile that compound is overall, and conversely, the lower a compound's normal boiling-point, the more volatile that compound is overall. Some compounds decompose at higher temperatures before reaching their normal boiling-point [3, 4].

BP is an important property for consideration in certain environmental problems and, further, it is a useful property for testing to develop a QSPR model. For both of these reasons, we have chosen to examine the relationship between BP and a new set of molecular structure descriptors.

Generally, normal boiling points are not difficult to determine: however, when a chemical is unavailable, or hazardous to handle, a reliable procedure for estimating its boiling-point is required. Furthermore, theoretical estimation of boiling points is also important for combinatorial chemistry, when literally millions of new compounds are synthesized and tested. Other physical properties, such as critical temperatures [4], flash points [5], and enthalpies of vaporization [6], can be predicted or estimated from boiling points. With the increased need for reliable data for optimization of industrial processes, it is important to develop reliable QSPR models to estimate normal boiling points for compounds not yet synthesized or whose boiling points are unknown. Several references [7-9] were made to investigations regarding the relationship between the normal boiling-point (NBP) and molecular structure descriptors. Many methods for prediction of BPs have therefore been developed, including many quantitative structure-property relationship (QSPR) studies using multiple linear regression (MLR) [10-14] and neural network (NN) Methods [15-17]. Karelson et al. and Katrizky et al. reviewed the earlier physicochemical, topological, geometrical and constitutional descriptors in QSPR studies, including the methods for boiling-point estimations of diverse organic compounds [18]. Sharma et al. [19] and Kier et al. [20] used topological and electrotopological descriptors for predicting normal boiling points of 34 and 21 amines by MLR method, respectively. Compared with the previous work, the data set used in our investigation is more diverse and the developed model is more general, stable and practicable.

Our goal here is to develop an accurate, simple, fast, and less expensive method for calculation of BP values of 216 amine derivatives. A stepwise multiple linear regression procedure [21] was used for selection of descriptors and modeling. Also, in this work we applied the back propagation neural network (BPNN) [22] and support vector machine regression (SVMR) [23] on this data set, but with no significant difference between results with the MLR method, so we preferred to report the results of the MLR method. The predictive power of the resulting model is demonstrated by testing them on unseen data that were not used during model generation. A physicochemical explanation of the selected descriptors is also given.

II. MATERIALS AND METHODS

II. 1. Data set

The normal boiling points values of 216 liquid amines compounds taken from the literature [24] are presented in the supplementary materials section. The dataset used for this work consists of 90 primary (1°), 30 secondary (2°) and 96 tertiary (3°) amines. The BP data and molecular descriptors were randomly split into training (75%) and test (25%) for each data set. The training set was used to adjust the parameters of the MLR models and the test sets were used to evaluate its prediction ability.

II. 2. Molecular modeling and descriptor generation

All calculations were run on a Dell Inspiron N5010 laptop computer with Intel Core^{TM} i7 processor with Windows 7 operating system. The molecular structures of all compounds were drawn into the HyperChem 8.0 program (Hypercube, Inc., Gainesville, 2011) and pre-optimized using the MM^{+} molecular mechanics method (Polak–Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by more precise optimization with the semi-empirical AM1 method, applying a root mean square gradient limit of 0.01 (Kcal $\cdot$ mol$^{-1}\cdot$ Å^{-1}), as a stopping criterion for optimized structures. The HyperChem output files were transferred into the Molecular Modeling Pro plus (MMP^{+})

(ChemSW Inc., version 6.3.3, Norgwy, 2004) software to calculate six kinds of molecular descriptors: MMP$^+$ software computes six classes of structural descriptors: constitutional; topological; geometrical; electrostatic; quantum chemical and thermodynamic molecular descriptors [25, 26]. Then a total of 72 molecular descriptors were calculated for each compound by the MMP$^+$ on the minimal energy conformations. In order to reduce redundant and non-useful information, constant or near constant values and descriptors found to be highly correlated pair-wise (one of any two descriptors with a correlation greater than 0.75) were excluded in a pre-reduction step; therefore 43 molecular descriptors underwent subsequent variable selection.

II. 3. Stepwise regression for descriptor selection

After the calculation of molecular descriptors, a stepwise regression routine implemented in the Molegro Data Modeller (MDM) software package was used to develop the linear QSPR model using calculated descriptors. The selection of relevant descriptors, which relate the BPs to the molecular structure, is an important step to construct a predictive model. In order to select the subset of descriptors that best explain compounds BP, we have used stepwise regression [27, 28]. The stepwise regression was applied to the input set of 43 molecular descriptors for each chemical of the studied data sets and the related response, in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals. Stepwise regression, included in the MDM software, was used for variables selection (based on the training set). Finally we obtained a four-significant descriptor subset, which keeps most interpretive information for BP. A total of four descriptors were calculated for each compound in each dataset contain molar mass (MM), Hansen dispersion forces (DF), molar refractivity (MR) and hydrogen bonding (HB) (for 1° and 2° amines). The values of selected descriptors are shown in the supplementary materials section.

II. 4. Multiple linear regressions

The datasets used in the QSPR analysis are, as already mentioned, composed of descriptors that should be correlated with the corresponding experimental responses. At this step it is necessary to apply a quantitative method able to find the existing relationship between a limited number of structural descriptors and the modeled response. In MDM, the used method is the MLR approach that can be exemplified by the following formula:

$$\boldsymbol{y_i}{=}\boldsymbol{b_0}{+}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}}\boldsymbol{b_j x_{ij}}{+}\boldsymbol{e_i} \tag{1}$$

where a linear relationship is computed between the studied responses (y_i) and the selected values of the descriptors (x_{ij}); e_i is the random error (also called model residual) and n is the number of descriptors. The intercept (b_0) and the coefficients (b_j) are thus to be estimated. The eq. (1) can be rewritten in a more compact form using the matrix notation:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{y}$ is the responses vector, $\boldsymbol{b}$ the vector of the coefficients, and $\boldsymbol{e}$ the vector of the errors. $\boldsymbol{X}$ is the matrix of the model, where the columns are the descriptors. In this software, to estimate the vector of the coefficients, the **OLS** technique is used:

$$\widehat{\boldsymbol{b}}{=}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y} \tag{3}$$

where $\widehat{\boldsymbol{b}}$ is the vector that estimates the $\boldsymbol{b}$ vector of the coefficients, $\boldsymbol{X}^T$ the transposed $\boldsymbol{X}$ matrix and $^{-1}$ is the inverse matrix operation. The **OLS** minimizes the sum of squares of the difference between the experimental responses and the ones calculated by the model. To work properly, the **OLS** assumes that: (1) a linear relationship exists between the descriptors and the response, (2) the response errors are independent and similarly distributed, (3) the descriptors are not too correlated among them (4) there are more compounds than modeling descriptors (a ratio that should always be higher than 5:1). Once the coefficients of the model are calculated, it is possible to obtain the vector of the $\widehat{\mathrm{y}}$, as in the following formula:

$$\widehat{\boldsymbol{y}}{=}\boldsymbol{X}\widehat{\boldsymbol{b}}{=}\boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}{=}\boldsymbol{Hy} \tag{4}$$

where $\boldsymbol{H}$ is the leverage (or hat) matrix that relates the calculated and the experimental responses. The diagonal elements of the hat matrix $\boldsymbol{h}_{ii}$ are useful to determine the distance of the i object from the centre of the chemical space of the model, [29] thus, for checking the structural applicability domain (AD) of the model. MLR techniques based on least-squares procedures are very often used for estimating the coefficients involved in the model equation [30].

II. 5. Validation of QSPR model

Validation of the developed models is an important aspect of any QSPR study. Once a model is obtained, it is important to determine its reliability and statistical significance. Several procedures are available to assist in this. These can be used to check whether the number of parameters is appropriate for the available data, as well as to provide some estimate of how well the model can predict the property for new molecules. In order to be reliable and predictive, QSPR models should (1) be statistically significant and robust, (2) be validated by making accurate predictions for external data sets not used in the model development, and (3) have a defined domain of application.

Model validation is of crucial importance to QSPR modeling. The training and predictive capability of a QSPR model should be tested through model validation [31-34].

Leave one out (LOO) and leave many out (LMO) cross validation are of the QSPR model internal validation. Predictability of the QSPR model is determined using the LOO-CV and LMO-CV methods. The cross validated explained variance (Q^2_{LOO} or Q^2_{LMO}) is calculated by the following equation:

$$Q^2_{LOO} \text{ or } Q^2_{LMO} = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (yi - \overline{y})^2} \tag{5}$$

where y_i, $\hat{y}_i$ and $\overline{y}$ are the measured, predicted, and averaged (over the entire training set) values of the dependent variable, respectively; the summations cover all the compounds in the training set. The cross validation approach is not sufficient to assess robustness and predictivity. The QSPR model developed using only training set chemicals is then applied to the external validation set chemicals to verify, more reliably, the predictive ability of the model.

The formula for the calculation of Q^2_{ext} is:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{test} (y_i - \overline{y}_{tr})^2} \tag{6}$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted (over the test set) values of the dependent variable, and $\overline{y}_{tr}$ is the averaged value of the property for the training set; the summations cover all the compounds in the validation set. The Q^2 values are good tests for evenly distributed data, but they are not always reliable for unevenly distributed datasets; instead, RMSEs (Root Mean Squared Errors) provide a more reliable indication of the fitness of the model, independently of the applied splitting. Other useful parameter to be considered are the RMSEs calculated on different sets: on training (RMSE$_{tr}$) and prediction (RMSE$_{ext}$). RMSE is calculated as in Equation 7:

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \right]^{0.5} \tag{7}$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted values of the property; n is the number of compounds in each set of data.

Another method for validation of the model is randomization testing or Y-scrambling. Randomization testing is a technique for checking the robustness of a QSPR model and the statistical significance of the estimated predicted power. In this test, the dependent variable vector (BP), Y-vector, is randomly shuffled and a new QSPR model is developed using the original independent variable matrix. The process is repeated several times. It is expected that the resulting QSPR models will generally have low $R^2(R^2Y_{scr})$, low Q^2 (Q^2Y_{scr}) and high RMSE values. If the new models developed from the data set with randomized responses have significantly lower R^2 and Q^2 than the original model, then this is strong evidence that the proposed model is well-founded, and not just the result of chance correlation [35, 36].

II. 6. Chemical domain of model applicability

An important problem of a QSPR model is the applicability domain (AD) [37]. The chemical domain of applicability is a theoretical region in the space defined by the modeled response and the descriptors of the model, for which a given QSPR should make reliable predictions. This region is defined by the nature of the chemicals in the training set, and can be characterized in various ways. The Williams plot of the regression allows a graphical detection of both the outliers for the response and the structurally influential chemicals in a model. The leverage (h) [38, 39] of a compound measures its influence on the model. The leverage of a compound in the original variable space is defined as:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X^T}, \tag{8}$$

where the $\boldsymbol{X}$ is the model matrix derived from the training set descriptor values and the leverage values of training set are diagonal elements of the Hat or Influence matrix $\boldsymbol{H}$(h_i= diag ($\boldsymbol{H}$)). The leverage values are always between 0 and 1. The warning leverage h^* is defined as follows:

$$h^* = 3 \times \frac{\sum_i h_i}{n} = 3 \times \frac{p'}{n} \qquad (i = 1, \ldots, n) \tag{9}$$

where n is the number of training set compounds and p' is the number of model parameters plus one. Observations with standardized residuals greater than (-3; $+3$) range, which lie outside the horizontal reference lines on the plot, are outlier's responses in the QSARINS version 2.2 (standardized residuals $>\pm 3\sigma$, σ is the standard deviation of residuals). Standardized residual (SR$_i$) for each sample is calculated as in Equation 10:

$$\text{SR}_i = \frac{(y_i - \hat{y}_i)}{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}} \tag{10}$$

where y_i and $\hat{y}_i$ are respectively the measured and predicted values of the property; n is the number of compounds in each set of data.

In the standardized residuals plot all values are within the (-3; $+3$) range, which confirms that there are no outliers. Furthermore, there is no clear pattern in the residuals, so nothing seems to be wrong with the model. To visualize the AD of a QSPR model, the plot of standardized residuals versus leverage values (h) (Williams plot) can be used for an immediate and simple graphical detection of both the response outliers and structurally influential chemicals in a model ($h > h^*$). Samples with high leverages have a stronger influence on the model than other samples; they may or may not be outliers, but they are influential. An influential outlier (high residual $+$ high leverage) is the worst case; it can, however, easily be detected using an influence plot. Leverages are useful for the detection of samples which are far from the centre within the space described by the model. If a sample has a very large leverage, it may be different from the rest and can be considered to be an outlier. Large leverage shows a high influence on the model. There are several methods for defining the AD of QSPR models [40], but the most common one is determining the leverage values for each compound [33].

III. RESULTS AND DISCUSSION

Experimental dataset of 90 primary amines, 30 secondary amines and 96 tertiary amines compounds were used to generate QSPR models which involved (4 and 3) descriptors based only on the molecular structure. All descriptors were calculated for the neutral species. The descriptors showed the importance of the effects related to the HB, MM, LF and MR interactions in the liquid media. Positive values in the regression coefficients show that the indicated descriptors contribute positively to the value of BP, whereas negative values indicate that the greater the value of the descriptor, the lower the value of BP.

III. 1. Model Analysis

A number of good models were obtained using the MLR technique. However, the training set was used to develop models (I), (II) and (III) that consisted of 67, 22 and 72 compounds, respectively. The specifications for models are given in Equation 11, 12 and 13. Inspection of the models revealed the superiority of models (I), (II) and (III), owing to better predictive power.

$$\begin{aligned}\mathrm{BP} =&33.37\,(\pm 20.55) + 0.44\,(\pm 0.16)\,\mathrm{MM}\\ &+8.88\,(\pm 1.27)\,\mathrm{DF} + 5.19\,(\pm 1.23)\,\mathrm{HB}\\ &+4.50\,(0.53)\,\mathrm{MR}\\ &n = 67,\ R^2 = 0.95,\ s = 16.56,\\ &F = 290.40,\ Q^2LOO = 0.94,\ \mathrm{model(I)}\end{aligned} \tag{11}$$

$$\begin{aligned}\mathrm{BP} =&0.77(\pm 0.25)\mathrm{MM} + 13.73(\pm 1.44)\mathrm{DF}\\ &+4.65(\pm 0.86)\mathrm{HB} + 1.99(\pm 0.82)\mathrm{MR} + 9.17(2.15)\\ &n = 22,\ R^2 = 0.98,\ s = 11.25,\\ &F = 233.76\ ,\ Q^2LOO = 0.97,\ \mathrm{model(II)}\end{aligned} \tag{12}$$

$$\begin{aligned}\mathrm{BP} =&51.45\,(\pm 15.82)\ +0.29\,(\pm 0.12)\,\mathrm{MM}\\ &+4.14\,(\pm 0.48)\,\mathrm{MR} + 10.76\,(\pm 0.96)\,\mathrm{DF}\\ &n = 72,\ R^2 = 0.9338,\ s = 16.22,\\ &F = 319.80, Q^2LOO = 0.93,\ \mathrm{model(III)}\end{aligned} \tag{13}$$

The squared correlation coefficients, R^2, squared cross validated correlation coefficients, Q^2, Fisher criterion value, F, and standard deviation, s, all give information about the "goodness" of the model. The square of the correlation coefficient (R^2) is indicated the quality of fit of all the data to a straight line is calculated for the checking of training and test set, and is calculated as:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{14}$$

where y_i is the experimental BP of the compound in the sample i, $\hat{y}_i$ represented the predicted BP of the compound in the sample i, $\bar{y}$, is the mean of experimental BP in the train set and n is the total number of samples used in the training set.

The molecular descriptors, experimental BP, predicted BP and residuals values of train and external prediction set by the MLR method are presented in the supplementary materials section. The plots of predicted BP versus experimental BP and the residuals (experimental BP − predicted BP) versus experimental BP value, obtained by the MLR modeling, and the random distribution of residuals about zero mean are shown in Fig. 1, 2 and 3. The stability and validity of model was tested by prediction of the response values for the prediction set. The robustness of the MLR model was also validated with the chance correlation procedure. The dependent variable vector (BP) was randomly shuffled and a new QSPR model was developed using the original independent variable matrix. The new QSPR model is expected to have low R^2 and high RMSE values. Several random shuffles of the y vector were performed and the results are shown in Tab. 1, 2 and 3. The R^2 and RMSE values indicate that the good results for the MLR model are not due to a chance correlation or structural dependency of the training set. The fitting criteria, internal validation criteria and external validation criteria are shown in Tab. 1, 2 and 3.

To visualize the AD of a QSPR model, the plot of standardized residuals (SR) versus leverage values (h) can be used for an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than three standard deviation units, $> 3s$) and structurally influential chemicals in a model ($h > h^*$). The Williams plots for the presented MLR models are shown in Fig. 4, 5 and 6. From these plots, the applicability domain is established inside a squared area within ±3 standard deviations and a leverage threshold ($h^* = 0.224$, 0.6818 and 0.17). For making predictions, predicted BP data must be considered reliable only for those compounds that fall within this AD on which the model was constructed. It can be seen from Fig. 4, 5 and 6 that the majority of compounds in the data sets are inside this area. However, only two compounds (1,2,3-Triaminopropane and Di-(2-hydroxyethyl) amine) in the training set of primary and secondary amine and two compounds (Tri-methylamine and N,N-Di-butyl-aniline) in the training set of tertiary amine slightly exceeds the critical hat value that the developed MLR models have good generalizability and predictivity for the compound with descriptor values significantly far from the centroid of the descriptor space. Also, the N-Methyl-pyrrolidine in the test set of tertiary amine is wrongly predicted (<3s), but with higher leverage values ($h > h^*$).These erroneous predictions could probably be attributed to wrong experimental data rather than to molecular structures.

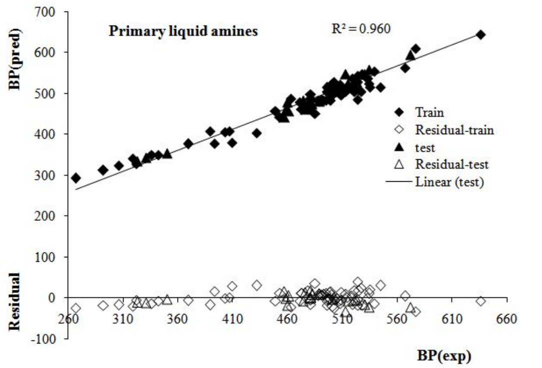

Fig. 1. Plot of predicted BP (K) and residuals versus experimental BP (K) of training and test set of 1° amines

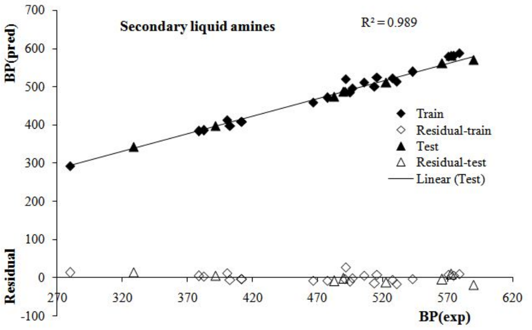

Fig. 2. Plot of predicted BP (K) and residuals versus experimental BP (K) of training and test set of 2° amines

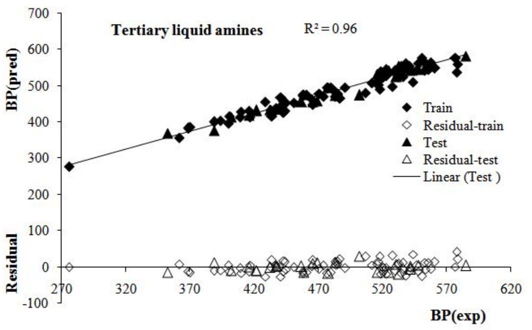

Fig. 3. Plot of predicted BP (K) and residuals versus experimental BP (K) of training and test set of 3° amines

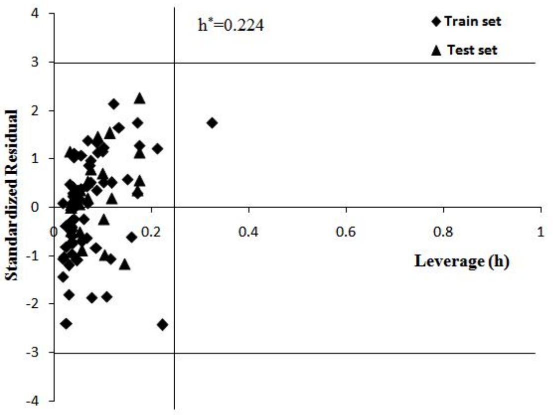

Fig. 4. Williams plot of standardized residual (SR) versus leverage (*h*) for 1° amines

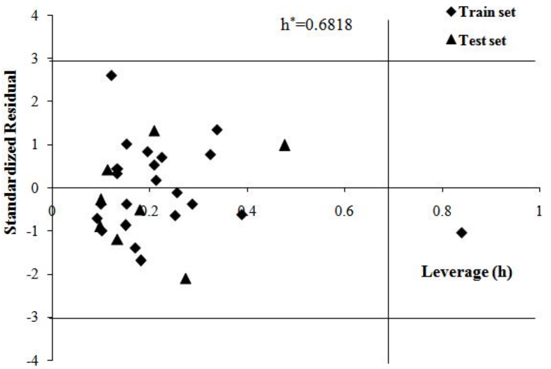

Fig. 5. Williams plot of standardized residual (SR) versus leverage (*h*) for 2° amines

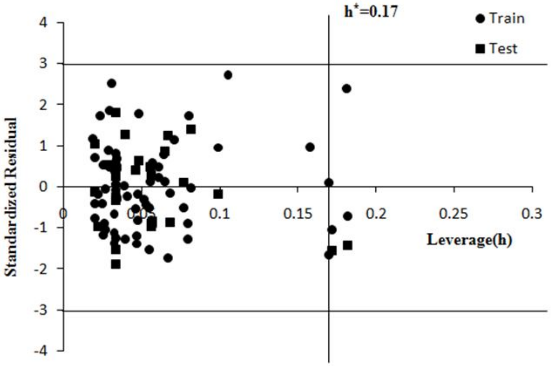

Fig. 6. Williams plot of standardized residual (SR) versus leverage (*h*) for 3° amines

Tab. 1. Fitting, internal validation and external validation criteria for primary amines model

Criteria	Statistical parameters		
Fitting criteria	$R^2 = 0.95$, R^2 adj $= 0.946$	RMSE$_{tr} = 15.92$	$S = 16.55$, $F = 290.40$
Internal validation	Q$^2{}_{LOO} = 0.94$, $Q^2{}_{LMO} = 0.93$	RMSEcv $= 17.85$	$R^2Y_{scr} = 0.062$, $Q^2\ Y_{scr} = 0.10$, RMSE$_{scr} = 69.61$
External validation	$Q^2_{ext} = 0.96$	RMSE$_{ext} = 13.69$	

Tab. 2. Fitting, internal validation and external validation criteria for primary amines model

Criteria	Statistical parameters		
Fitting criteria	$R^2 = 0.95$, R^2 adj $= 0.946$	RMSE$_{tr} = 15.92$	$S = 16.55$, $F = 290.40$
Internal validation	$Q^2{}_{LOO} = 0.94$, $Q^2{}_{LMO} = 0.93$	RMSEcv $= 17.85$	$R^2Y{}_{scr} = 0.062$, $Q^2Y_{scr} = 0.10$, RMSE$_{scr} = 69.61$
External validation	Q$^2{}_{ext} = 0.96$	RMSE$_{ext} = 13.69$	

Tab. 3. Fitting, internal validation and external validation criteria for primary amines model

Criteria	Statistical parameters		
Fitting criteria	$R^2 = 0.95$, R^2 adj $= 0.946$	RMSE$_{tr} = 15.92$	$S = 16.55$, $F = 290.40$
Internal validation	$Q^2{}_{LOO} = 0.94$, $Q^2{}_{LMO} = 0.93$	RMSEcv $= 17.85$	$R^2Y_{scr} = 0.062$, $Q^2Y_{scr} = 0.10$, RMSE$_{scr} = 69.61$
External validation	$Q^2{}_{ext} = 0.96$	RMSE$_{ext} = 13.69$	

III. 2. Interpretation of the selected descriptors

The boiling-point of an organic compound reflects its molecular structure, specifically the type of intermolecular interactions that bind the molecules together in the liquid state. The developed QSPR showed that HB (1° and 2° amine), MM, DF and MR descriptors significantly influence amines normal boiling point. In the MMP+ software has been used from Hansen's approach for calculation cohesion energy [41]. The basic equation governing the assignment of Hansen parameters is that the total cohesion energy, E, must be the sum of the individual energies that make it up.

$$E = E_D + E_P + E_H \tag{15}$$

E_D is dispersion cohesion energy; E_P is polar cohesion energy, and E_H is Hydrogen bonding cohesion energy [41].

The first descriptor is hydrogen bonding (HB). Hydrogen bonding is a molecular interaction and resembles the polar interactions in this respect. The basis of this type of cohesive energy is attraction among molecules because of the hydrogen bonds. The number of hydrogen atoms on the nitrogen available for hydrogen bonding greatly influences the strength of the intermolecular forces. Hydrogen bonding occurs in molecules containing the highly electronegative elements F, O, or N directly bound to hydrogen. Since H has an electronegativity of 2.2 these bonds are not as polarized as purely ionic bonds and possess some covalent character. However, the bond to hydrogen will still be polarized and possess a dipole. The dipole of one molecule can align with the dipole from another molecule, leading to an attractive interaction that we call hydrogen bonding. As you might expect, the strength of the bond increases as the electronegativity of the group bound to hydrogen is increased. So in a sense, HO, and NH are "sticky" - molecules containing these functional groups will tend to have higher boiling points than you would expect based on their molar mass. Hydrogen bonding significantly influences the properties of primary and secondary amines. Hydrogen Bond forming ability is a measure of the tendency of a molecule to form hydrogen bonds. The HB is an intermediate range intermolecular interaction between electron deficient hydrogen and a region of high electron density. Hydrogen bonding is a special type of dipole-dipole interaction between acidic hydrogen ($\delta+$) and a lone pair ($\delta-$). The hydrogen bonding is the powerful intermolecular attraction that results from -N-H…N- hydrogen bonding in amines. So the increase in the boiling point for 1°-amines increase. The results illustrates differences associated with isomeric 1°, 2° and 3°-amines, as well as the influence of chain branching. Since 1° amines have two hydrogens available for hydrogen bonding, we expect them to have higher boiling points than isomeric 2°-amines, which in turn should boil higher than isomeric 3°-amines (no hydrogen bonding). This type of polarity is so strong compared to other Van der Waals interactions. Understandably, hydrogen bonding plays a significant role in physical property. Hydrogen bonding is not a true bond, but a very strong form of dipole-dipole attraction. In this study we have a 1° and 2° amines containing hydrogen bond donor (N–H bonds) and hydrogen bond acceptor (lone pair of nitrogen atom). The hydrogen bonding (HB) is a measure of the tendency of a molecule to form hydrogen bonds. As the hydrogen bond formation increases, BP increases.

Primary and secondary amines can H-bond with themselves, so have relatively high boiling points.

However, because the N-H bond is less polar than the O-H bond, amines have lower boiling points than alcohols. Primary and secondary amines have boiling points similar to aldehydes and ketones. Tertiary amines cannot H-bond with themselves, and so have boiling points near those of ethers and hydrocarbons. The intermolecular hydrogen bonding can dramatically influence physical and chemical properties.

The second descriptor is molar mass (MM). Among the size descriptors, molar mass is the simplest and most commonly used molecular 0D-descriptor, calculated as the sum of the atomic masses of all the atoms in a molecule. It is related to molecular size and is atom-type sensitive. It is defined as $\mathrm{MM} = \sum_{i=1}^{A} m_i$ where m is the atomic mass and i runs over the A atoms of the molecule. By increasing molecular mass of compounds the BP increases. The larger the molar mass, the greater the polarizability of the molecule and hence also the van der Waals attractive forces between near neighbors. Increasing molecular mass leads to increasing the boiling point of amines.

The third descriptor is Hansen dispersion forces (DFs). DF is a measure of dispersion cohesion energy. The most general are the nonpolar interactions. These are derived from atomic forces and have also been called dispersion interactions in the literature. As molecules are built up from atoms, all molecules contain those types of attractive forces. The DF of attraction exists between molecules which have no permanent dipole. The van der Waals force is an attractive force between two atoms or nonpolar molecules, which arises because a fluctuating dipole moment in one molecule induces a dipole moment in the other, and the two dipole moments then interact. With increasing molecular weight, molecular volume and surface area, the van der Waals forces increase. Van der Waals attractive forces exist between all polar and nonpolar molecules. The boiling point increases with increasing dispersion forces.

The fourth descriptor is molar refractivity. The MR is the volume of the substance taken up by each mole of that substance. In SI units, MR is expressed as $\mathrm{m}^3\ \mathrm{mol}^{-1}$. MR is a molecular descriptor of a liquid, which contains both information about molecular volume and polarizability, usually defined by the Lorenz-Lorentz equation [42]:

$$\mathrm{MR} = \frac{n^2-1}{n^2+2}.\frac{\mathrm{MM}}{\rho} = \frac{\varepsilon-1}{\varepsilon+2}.\overline{V} \qquad (16)$$

where MM is the molar mass, ρ the liquid density, and $\overline{V}$ the molar volume, and n the refractive index of the liquid, and its square coincides with the dielectric constant ε. Refractive index measurements yield information about the ability of the molecular electron distribution to be deformed in an electric field or in the presence of other molecules. The polarizability of a molecule is related to the intermolecular forces important in the interaction of molecules. Increasing MR leads to increasing intermolecular forces. However, increasing the intermolecular forces increases the extent of BP of the each compound.

IV. CONCLUSION

In this paper, new QSPR models have been developed for predicting the BP of a diverse set of amines from the molecular structure alone. Stepwise-MLR analysis was followed to develop a model for predicting the BP of amines. The descriptors involved in the correlations reflect the intermolecular interactions. By performing model validation, it can be concluded that the presented model is a valid model and can be effectively used to predict the BP of amines with an accuracy approximating the accuracy of experimental BP determination. The proposed models give reasonable accuracy and are predictive because molecular descriptors can be calculated easily as long as the molecular structure of the concerned compound is known. Therefore, reliable predictions of boiling points in liquid amines can be obtained before they are actually synthesized. This work also demonstrates that molecular descriptors are useful for the structural characterization of amines. Molecular structure is one of the basic concepts of chemistry, since physical, chemical and biological behaviors of molecules are determined by it.

References

[1] J.E. McMurry, *Organic Chemistry* (3rd ed.), Belmont: Wadsworth 1992.

[2] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, NewYork 1976.

[3] W.J. Lyman, W.F. Reehl and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods*, American Chemical Society, Washington DC 1990.

[4] C.H. Fisher, *Boiling Point Gives Critical Temperatures*, Chem. Eng., **96**, 157-158 (1989).

[5] K. Satyanarayana and M.C. Kakati, *Note: Correlation of flash points*, Fire Mater., **15**, 97-100 (1991).

[6] C.E. Rechsteiner, *In handbook of chemical property estimation methods*, McGraw-Hill: NewYork 1982.

[7] J. Ghasemi and S. Saaidpour, *Artificial neural network-based quantitative structural property relationship for predicting boiling points of refrigerants*, QSAR & Comb. Sci., **28**, 1245-1254 (2009).

[8] L.M. Egolf and P.C. Jurs, *Prediction of boiling points of organic heterocyclic compounds using regression and neural network techniques*, J. Chem. Inf. Comp. Sci., **33**, 616-625 (1993).

[9] L.H. Hall and L.B. Kier, *Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information*, J. Chem. Inf. Comput. Sci., **35**, 1039-1045 (1995).

[10] O. Ivanciuc, T. Ivanciuc and A.T. Balaban, *Quantitative structure-property relationship study of normal boiling points for halogen-/ oxygen-/ sulfur-containing organic compounds using the CODESSA program*, Tetrahedron, **54**, 9129-9142 (1998).

[11] D. Plavsic, N. Trinajstic, D. Amic, et al., *Comparison between the structure–boiling point relationships with different descriptors for condensed benzenoids*, New J. Chem., **22**, 1075-1078 (1998).

[12] O. Ivanciuc, T. Ivanciuc and A.T. Balaban, Design *of topological indices. Part 10. Parameters based on electronegativity and covalent radius for the computation of molecular graph descriptors for heteroatom-containing molecules*, J. Chem. Inf. Comput. Sci., **38**, 395-401 (1998).

[13] A.P. Bunz, B. Braun and R. Janowsky, *Application of quantitative structure-performance relationship and neural network models for the prediction of physical properties from molecular structure*, Ind. Eng. Chem. Res., **37**, 3043-3051 (1998).

[14] D.T. Stanton and P.C. Jurs, *Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies*, Anal. Chem., **62**, 2323-2329 (1990).

[15] M.D. Wessel and P.C. Jurs, *Prediction of normal boiling points of hydrocarbons from molecular structure*, J. Chem. Inf. Comput. Sci., **35**, 68-76 (1995).

[16] L.H. Hall and C.T. Story, *Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks*, J. Chem. Inf. Comput. Sci., **36**,1004-1014 (1996).

[17] E.S. Goll and P.C. Jurs, *Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model*, J. Chem. Inf. Comput. Sci., **39**, 974-983 (1999).

[18] A.R. Katritzky, M. Kuanar, S. Slavov, et al., *Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction*, Chem. Rev., **110**, 5714-5789 (2010).

[19] V. Sharma , R. Goswami and A.K. Madan, *A novel highly discriminating topological descriptor for structure–property and structure–activity studies*, J. Chem. Inf. Comput. Sci., **37**, 273-282 (1997).

[20] L.B. Kier and L.H. Hall, *Molecular Structure Description-The Electrotopological State*, Academic, San Diego, CA, USA 1999.

[21] J. Ghasemi and S. Saaidpour, *quantitative structure–property relationship study of n-octanol–water partition coefficients of some of diverse drugs using multiple linear regression*, Anal. Chim. Acta, **604**, 99-106 (2007).

[22] S. Saaidpour, *Prediction of drug lipophilicity using back propagation artificial neural network modeling*, Orient. J. Chem., **30**(2), 793-802 (2014).

[23] Y. Pan, J. Jiang, R. Wang, et al., *A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine*, J. Hazard. Mater., **168**, 962-969 (2009).

[24] R.L. Shriner, C.K.F. Hermann, T.C. Morrill, et al., *The Systematic Identification of Organic Compounds*, eights edition, John Wiley & Sons. Inc. 2004.

[25] L.B. Kier and L.H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, RSP-Wiley, Chichetser, UK 1986.

[26] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany 2000.

[27] R.R. Hocking, *The Analysis and Selection of Variables in Linear Regression*, Biometrics 1976.

[28] S. Saaidpour, *Prediction of the Adsorption Capability onto Activated Carbon of Liquid Aliphatic Alcohols using Molecular Fragments Method*, Iran. J. Math. Chem., **5** (2), 127-142 (2014).

[29] A.C. Atkinson, In Plots, *Transformations and Regression,* Clarendon Press, Oxford 1985.

[30] P. Gemperline, *Practical Guide to Chemometrics*, Taylor & Francis Group, Boca Raton 2006.

[31] A. Golbraikh and A. Tropsha, *Beware of* q^2*!,* J. Mol. Graph. Model., **20**, 269-276 (2002).

[32] P. Gramatica, P. Pilutti and E. Papa, *Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training–Test Sets and Consensus Modeling*, J. Chem. Inf. Comput. Sci., **44**, 1794-1802 (2004).

[33] P. Gramatica, *Principles of QSAR models validation: internal and external*, QSAR & Comb. Sci., **26**, 694-701 (2007).

[34] P. Gramatica, E. Giani and E. Papa, *Statistical external validation and consensus modeling: A QSPR case study for* K_{oc} *prediction*, J. Mol. Graph. Model., **25**, 755-766 (2007).

[35] A. Tropsha, P. Gramatica and V.K. Gombar, *The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models*, QSAR & Comb. Sci., **22**, 69-77 (2003).

[36] L. Eriksson, J. Jaworska, A.P. Worth, et al., *Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs*, Environ. Health Perspect. Aug., **111**, 1351-1375 (2003).

[37] S. Weaver and M.P. Gleeson, *The importance of the domain of applicability in QSAR modeling*, J. Mol. Graph. Model., **26**, 1315-1326 (2008).

[38] A.C. Atkinson, *Plots, Transformations and Regression*, Clarendon Press, Oxford 1985.

[39] M. Shacham, N. Brauner, G.S. Cholakov, et al., *Identifying Applicability Domains for Quantitative Structure Property Relationships*, Elsevier, Amsterdam 2009.

[40] F. Sahigara, K. Mansouri, D. Ballabio, et al., *Comparison of different approaches to define the applicability domain of qsar models*, Molecules, **17**, 4791-4810 (2012).

[41] Ch. M. Hansen, *Hansen Solubility Parameters*, *A User's Handbook*, Second Edition, CRC Press, 2007.

[42] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Vol. I & II, WILEY-VCH, 2009.

Saadi Saaidpour was born in Kermanshah, Iran in 1972. He is an Assistant Professor at the department of chemistry, IAU sanandaj branch, sanandaj, Iran. In 2004, he received his MSc degree in analytical chemistry and in 2008 he received his PhD degree (under the supervision of Prof. Jahan B. Ghasemi) in Analytical chemistry and chemometrics at razi university, Kermanshah, Iran. His research interests concern chemometrics methods, computational chemistry and QSAR & QSPR studies.

Asrin Bahmani is a Student for PhD, in Organic Chemistry, Department of Chemistry, University of Kurdistan, Sanandaj, Iran. In 2005, she received her BSc degree in chemistry and in 2009 she received her MSc degree in organic chemistry at University of Kurdistan, Sanandaj-Iran. Her research interests concern computational chemistry, docking and QSAR/ QSPR studies.

Amin Rostami was born in Kurdistan, Iran in 1976. He obtained his PhD (2006) under the supervision of Prof. A. Khazaei from Bu-Ali Sina University. Currently, he is working as associate of professor at university of Kurdistan. His research interests are in the area of catalysis, green chemistry and medicinal chemistry. He has over 65 publications.

DOI:10.12921/cmst.2015.21.04.004