

Prediction of $\Delta H_{f(gas)}^{\circ}$, $\Delta H_{f(liquid)}^{\circ}$ of Amines Derivatives Using chemometrics (A Quantitative Structure–Property Relationship Study)

MORTEZA REZAEI AND ESMAT MOHAMMADINASAB*

Department of Chemistry, Arak Branch, Islamic Azad University, Arak, Iran

ARTICLE INFO

Article History:

Received 7 January 2017

Accepted 9 June 2018

Published online 30 July 2019

Academic Editor: Ali Reza Ashrafi

Keywords:

Amines derivatives

Standard enthalpy of formation

Molecular descriptors

Chemometrics

ABSTRACT

In this study, multiple linear regression method that is based on property-structure model has been used to predict the standard enthalpies of formation for the gas and liquid phases of the 33 different types of amines. It was indicated that among studied topological and geometric descriptors to predict the $\Delta H_{f(liquid)}^{\circ}$, descriptors as PSA, H, MaxZL and V have more importance than the other descriptors. Also, the results of experiments on studied amines were compared with the results of multiple linear regression calculations and it was observed that such descriptors as MaxZL, MaxPA, DE, J and WW are the best descriptors for predicting the values of $\Delta H_{f(gas)}^{\circ}$ of this class of amines.

© 2019 University of Kashan Press. All rights reserved

1 INTRODUCTION

The progress of the chemometrics is stoutly related to the extension of computers in chemistry [1–2]. Representative applications of chemometric methods are the improvement of quantitative structure activity relationships or the valuation of analytical–chemical data. Those methods in chemistry are the obedience that uses mathematical and statistical methods to obtain appropriate information on chemical material.

The most significant consequence of using statistical and analytic methods and chemometrics for modeling and performing statistical computations is that it

*Corresponding Author (Email: e-mohammadinasab@iau-arak.ac.ir).

DOI: 10.22052/ijmc.2018.72357.1264

reduces time, costs and environmental pollution and allows for predicting most of the chemicals' properties with high precision and without the need to laboratory equipment. The experimental methods for determination of properties of chemical substances are not always available. Also, chemical hazards are caused by chemical substances causing significant damage to the environment. Different types of chemical substances have been associated with unsafe health effects. Common chemical hazards include: skin inflammation, deform burns, eye damage or blindness caused by corrosive chemical products. A chemical hazard is a type professional hazard caused by contact to chemicals in the place of work. Exposure to chemicals in the workplace can reason acute or long-term unfavorable health effects. These hazards can cause physical and/or health dangers.

Modeling using chemometrics involves three elements namely experimental data extracted from reliable resources, appropriate descriptors for chemical structure and an effective computational method in which the model is formula. Some modeling techniques of chemometrics enabled chemists to express the relationships between molecules structure and their properties which could predict continuous variables such as boiling point, enthalpy of formation, Gibbs free energy, etc. or rank discrete variables such as bitter/sweet, poisonous/non-poisonous using structural parameters. The chemists use various labels and formulas to convey information about chemicals and their structures. There are numerous methods for determining the numerical value assigned to structures. The experimental molecular descriptors derive from symbolic representation of a molecule and can be classified in terms of various types of molecular representation [3]. They are derived from different theories such as quantum chemistry, organic chemistry and physical chemistry and are analyzed using statistical and chemometric methods. Chemometrics has numerous applications in different branches of chemistry. One of the significant applications of chemometrics has been the studies which associate chemical structures and properties in a data-set of compound [4-6].

Quantitative structure activity/property relationships (QSAR/QSPR) studies, as one of the most important areas in chemometrics, play a fundamental role in predicting the biological activity/property of new compounds. The most significant step in building QSAR models is the appropriate representation of the structural and physicochemical features of chemical entities [7].

The structure-property relationship among numerous computational methods plays the most important role in predicting chemicals' properties. Chemical structure changes reflect changes in the physicochemical properties. One of the applications of this relationship is to predict various physicochemical properties of organic compounds.

2. MATERIAL AND METHODS

2.1 MATERIAL

Amines are considered as one of the most abundant nitrogen-containing organic compounds which are present in definite plants and a restricted chemical intermediary, histamine that arises in most animal tissues. They are used for different applications from industrial manufacturing to medicine.

A amine solution is polluted with various material such as hydrocarbon and non-hydrocarbon, organic acids, cations, anions or even oxygen. Some amines are toxic and carcinogenic. Experimental determination of properties of amines is both time and material consuming; it can only be done if the compounds are available. Modeling and computational methods using the software conveniences are used that abridged mentioned difficulties in determining the physico-chemical properties of amines [8].

2.2 COMPUTATIONAL METHOD

The chemometric methods include Multivariate Calibration, Multivariate Linear Regression, Partial Least Square, Multivariate Nonlinear Regression, K nearest neighbours and etc. Calibration levels is usually included experimental design, model selection, parameter estimation and prediction of indefinite. In a multivariate calibration there is not the possibility of correcting annoying without entering additional information. However, in multivariate Calibration methods it is possible to separate the useful information from inappropriate in a set. In nonlinear regression observational statistics consist of independent variables and their associated observed dependent variables are modeled by a function which is a nonlinear combination of the model parameters. The *k*-nearest neighbors algorithm is used in classification or regression machine learning responsibilities. In kNN classification an input is classified by a majority choose of its neighbours whereas regression involves organizing a relationship between input points and the rest of the data. The kNN is a very simple non-parametric method of marking unknown inputs.

In this research is studied multivariate linear regression method for thermodynamics properties prediction of this class of amines. The advantage of multiple linear regressions to simple linear regression is that it has two or more independent variables. In multiple regression method, it is not necessary that there be a linear relationship between dependent variable *y* and independent variable *x* but the dependent variable *y* is a linear combination of coefficients (parameters)

[9–11]. The linear relationship between a dependent variable and independent variables is calculated by Equation 2.1:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (2.1)$$

where y_i is modeled dependent variable; p is the number of predictors (independent variables) β_i is regression coefficients, x_{ip} is i amount of the j independent variable; and ε_i is the observed error amount for i .

In present research, standard enthalpies of formation for the gas and liquid phases and two-dimensional and geometric descriptors of studied amines are considered as dependent and independent variables in the regression model, respectively. Recent studies have showed that molecular indices of molecules determine the physical-chemical properties of them, so it's especially important to know that how change in the molecular structures changes the physical-chemical properties [12–14].

2.3 MOLECULAR DESCRIPTORS

The two-dimensional (2D) indices and Geometric descriptors in this study due to are molecular descriptors which can be related to thermodynamics functions such as the standard enthalpies of formation for the gas and liquid phases of 33 different types of amines. It is worth mentioning, the standard enthalpy of formation of a chemical compound which was used as modeled variable is defined as the change of enthalpy during the formation of 1 mole of the compound from its constituent elements under standard conditions at the specified temperature 25°C and $P = 1$. It should be noted that thermodynamics functions of organic chemical are susceptible to molecular geometry and structural indices, for this reason in this study mentioned geometric and topological descriptors as independent variables were employed.

The studied 2D and geometric descriptors in this research are calculated using related formulas. To analyze the relationship between thermodynamics properties in contrast with molecular descriptors, the research data were collected as follows:

First, standard enthalpies of formation for the gas and liquid phases; $\Delta H_{f(\text{gas})}^\circ$ and $\Delta H_{f(\text{liquid})}^\circ$ related to the 33 different types of the amines were obtained from National Institute of Standard and Technology chemistry web book and were listed in Table 1.

Second, the two and three-dimensional descriptors of the amines were calculated [15–25] and utilized as independent variables in the regression model. As it is clear, Two-dimensional descriptors derive from 2D representation of a molecule, atoms connection. A molecular graph or chemical graph is a

representation of the structural formula of a chemical compound that bond type and the atom type are not necessarily specified. The numerical value of two-dimensional descriptors is calculated using the matrixes which represent the molecular diagram. These diagrams can be sensitive to one or several structural characteristics of a molecule such as size, shape, symmetry, branched, or ring shaped molecules [26]. The geometric descriptors or spatial descriptors are used to depict a three-dimensional representation of molecule as a geographical solid structure. The geometric descriptors not only represent the nature and type of atomic bonds but also the spatial structure of atoms in a molecule. These descriptors involve spatial descriptors and Morse molecular mass distribution function [27].

Table 1. The used molecular descriptors, explanation and references.

No	Indices	Explanation	References
1	Wiener	$W = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij}$	15-16
2	Hyper wiener	$WW(G) = \sum d(U, V)^2 + \sum d(U, V) / 2$	17
3	Wiener Polarity	$W_p(G) = \{u, v\} d_G(u, v) = 3, u, v \in V \}$	18
4	Platt	$F = \sum_{i=1}^M D_e$	19
5	Randic	$\chi = \sum \left(\frac{1}{d_i d_j} \right)^{\frac{1}{2}}$	20
6	Balaban	$J = \frac{m}{\mu + 1} \sum_{i=1}^n \sum_{j=1}^n [(D_i)(D_j)]^{0/5}$	21-22
7	Harary	$H = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (D_{ij})^{-2}$	23
8	Sz	$Sz_v(G) = \sum_e n_u(e G) \cdot n_v(e G)$	24
9	Geometry indices	Geometric indices are taken from Chemicalize Web Book	25

In this study, 2D descriptors including Platt, Projection Surface Area, Balaban, Randić, Harary, Szeged, Wiener, Wiener-Polarity, Hyper-Wiener and geometric descriptors including Dreiding-Energy, Volume, Minimal projection Area, Minimal Z Length, Maximal projection Area and Maximal Z Length using the literatures [25] were used for the estimation of thermodynamic properties of amines.

Third, the relationships between $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ of 33 different types of amines with 2D, and Geometric descriptors were investigated by SPSS software, and multiple linear regression method was used to find out the most appropriate molecular descriptors for predicting this class of thermodynamics function of amines.

3. RESULTS AND DISCUSSION

The standard enthalpies of formation of liquid and gas phases ($\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$) of 33 different types of the amines were listed in Table 2.

After calculating the descriptors using the mentioned resources, we investigated the relationship between experimental values of the standard enthalpies of formation for the gas and liquid phases, namely $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ with above descriptors using Excel software.

Table 2. The literature values of the $\Delta H_{f(\text{gas})}^{\circ} / \text{kJmol}^{-1}$ and $\Delta H_{f(\text{liquid})}^{\circ} / \text{kJmol}^{-1}$ of training set.

No	Name	$\Delta H_{f(\text{Liq})}^{\circ}$	$\Delta H_{f(\text{gas})}^{\circ}$	No	Name	$\Delta H_{f(\text{Liq})}^{\circ}$	$\Delta H_{f(\text{gas})}^{\circ}$
1	Aminoethane	-84.5	-57.7	20	4-methylpyridine	59.2	103.8
2	3-amino-1-propene	10.0	-	21	Aminocyclohexane	-141.5	-98.7
3	1-aminopropane	-101.5	-69.9	22	1-aminohexane	-239.0	-193.9
4	2-aminopropane	-112.3	-83.7	23	di-n-Propylamine	-154.6	-116.5
5	Methylethylamine	-45.73	-	24	triethylamine	169.0	-
6	Trimethylamine	-45.73	-23.7	25	2-methylaniline	-4.7	53.2
7	Pyrrolidine	-41.0	-3.4	26	3-methylaniline	42.3	61.1
8	1-aminobutane	-127.7	-95.0	27	4-methylaniline	19.0	41.8
9	1-amino-2-methylpropane	-132.5	-98.6	28	N-Methylaniline	33.4	83.9
10	2-Aminobutane	-	-106.0	29	Quinuclidine	-	-4.3
11	2-methyl-2-aminopropane	-150.6	-120.0	30	1-aminoheptane	-266.0	-216
12	Diethylamine	-131.0	-99.80	31	2,4,6-trimethylpyridine	31.0-	-
13	Pyridine	99.96	140.2	32	di-n-butylamine	-212.3	-171.1
14	n-methylpyrrole	62.38	103.1	33	tri-n-propylamine	-206.9	-160.8
15	Piperidine	-86.44	-47.15	34	Carbazole	-	205.0
16	Methylbutylamine	-141.2	-108.3	35	2-aminobiphenyl	-	184.4
17	Aniline	31.3	87.03	36	Diphenylamine	132.0	202
18	2-methylpyridine	-68.99	-26.5	37	n,n-diphenylamino methane	120.5	-

The multiple linear regression method was used to explore the relationship between standard enthalpy value of formations and molecular descriptors using SPSS software version 16 with multiple linear regression method and backward procedure with an estimated error of 0.05. Whether or not the regression model explains a statistically significant percentage of data was found through the *ANOVA Table* of output based on the MLR model in terms of the relationship between properties and structural indices.

In this method, 15 molecular descriptors were considered as independent variables and the values of $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ were considered as dependent variables. As it can be seen in Table 3 and 3, the 12th and the 5th models indicate a highly acceptable validity to maintain the structure-property relationship for predicting $\Delta H_{f(\text{liquid})}^{\circ}$ and $\Delta H_{f(\text{gas})}^{\circ}$ among the proposed models, respectively.

Table 3. The efficient statistical parameters for prediction of $\Delta H_{f(\text{liquid})}^{\circ}/\text{kJmol}^{-1}$ in 12 models.

No	Predictors	R	R ²	R ² _{Adj}	STD	F	MS
1	PSA, Platt, X, J, W, H, WW, Wp, Sz, DE, V, MinPA, MinZL, MaxPA, MaxZL	0.972	0.945	0.896	34.474	19.307	22946.43
2	PSA, Platt, X, J, W, H, WW, Sz, DE, V, MinPA, MinZL, MaxPA, MaxZL	0.972	0.944	0.901	33.521	21.878	24583.90
3	PSA, Platt, X, J, W, H, WW, Sz, DE, V, MinPA, MaxPA, MaxZL	0.972	0.944	0.906	32.658	24.819	26471.95
4	PSA, Platt, X, J, W, H, WW, Sz, DE, V, MinPA, MaxZL	0.972	0.944	0.910	31.951	28.078	28665.22
5	PSA, Platt, X, J, H, WW, Sz, DE, V, MinPA, MaxZL	0.971	0.944	0.914	31.278	31.952	31259.62
6	PSA, X, J, H, WW, Sz, DE, V, MinPA, MaxZL	0.971	0.943	0.917	30.683	36.505	34368.841
7	PSA, X, J, H, WW, Sz, V, MinPA, MaxZL	0.971	0.942	0.920	30.223	41.768	38154.538
8	PSA, X, J, H, WW, V, MinPA, MaxZL	0.970	0.941	0.922	29.808	48.264	42884.507
9	PSA, X, J, H, WW, V, MaxZL	0.970	0.940	0.924	29.463	56.397	48956.987
10	PSA, J, H, WW, V, MaxZL	0.968	0.937	0.922	29.823	63.951	56879.331
11	PSA, H, WW, V, MaxZL	0.967	0.935	0.923	29.606	77.746	68146.928
12	PSA, H, MaxZL, V	0.964	0.928	0.918	30.511	90.859	84583.718

Table 4. The efficient statistical parameters for prediction of $\Delta H_{f(\text{gas})}^{\circ}/\text{kJmol}^{-1}$ in 5 models.

No	Predictors	R	R ²	R ² _{Adj}	STD	F	MS
1	MaxPA, PSA, MinPA, MinZL, DE, J, WW, Wp, MaxZL	0.967	0.936	0.911	39.011	37.214	56635.512
2	MaxZL, MaxPA, DE, Wp, PSA, MinPA, WW, , J	0.967	0.936	0.914	38.211	43.633	63709.935
3	MinPA, MaxPA, DE, J, WW, PSA, MaxZL,	0.967	0.935	0.917	37.622	51.407	72762.430
4	MaxZL, MaxPA, PSA, DE, J, WW	0.967	0.934	0.919	37.147	61.460	84807.622
5	MaxZL MaxPA, DE, J, WW	0.964	0.930	0.917	37.646	71.472	101291.556

4. STATISTICAL COEFFICIENTS

Various statistical tests and coefficients can be used for validation of applied models. In the following, the most appropriate and important ones are represented.

4. 1. **R:** Correlation coefficient indicates the correlation between two variables. Statistically, the higher correlation between variables X and Y, the more accurate the prediction will be. The R=0.964 in 12th and 5th models in Tables 3 and 4, illustrate a strong correlation between $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ as dependent variables and significant molecular descriptors as independent variables using the MLR method respectively.

4. 2. **R²:** The values of R²=0.930 and 0.928 in 12th and 5th models in Tables 3 and 4, illustrate that 93% and 92.8% of $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ can be explained by the best models, respectively.

4. 3. **Sig:** level of significance (Sig) shows the maximum amount of errors in decision level in terms of testing the hypothesis. Thus, the appropriate level of significance (Sig. <0.05) indicates that the association between the dependent variables $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ and the their predictor variables is statistically significant.

4. 4. **F Statistic:** The lower the significance level is, the higher the Fischer statistic will be. The higher values of Fisher number equal to 71.472 and 90.859 in 12th and 5th models according to the Tables 3 and 4, respectively, affirms the fact that there is sufficient data and the molecular indices in the best models have been selected appropriately.

4. 5. R^2_{Adjust} : refers to adjusted determination coefficient. It should be noted that the percentage of adjusted determination coefficient does not represent the influence of all the independent variables, but it only illustrates the real influence of applied independent variables on the dependent variable. Thus, the equation $R^2_{\text{Adjust}}=0.917$ for $\Delta H^{\circ}_{f(\text{gas})}$ can be used to explain the $\Delta H^{\circ}_{f(\text{gas})}$. Variations in terms of the values of independent variables MaxZL, MaxPA, DE, J and WW, and $R^2_{\text{Adjust}}=0.918$ for $\Delta H^{\circ}_{f(\text{liquid})}$ indicates that variations of $\Delta H^{\circ}_{f(\text{liquid})}$ can be explained in terms of the values of PSA, H, MaxZL and V.

4. 6. **STD**: Standard error of estimate is used to indicate the spreadness of values in a distribution. It is a standard method for determining the normal or the abnormal values. The indices mentioned specify that the selected model possess the necessary quality for predicting the thermodynamics functions of amines.

Finally, the comparison of models and the values of statistical coefficients showed that the best models for predicting $\Delta H^{\circ}_{f(\text{gas})}$, $\Delta H^{\circ}_{f(\text{liquid})}$ of amines using the MLR method are as follows:

$$\Delta H^{\circ}_{f(\text{gas})} = -4.725 \text{ MaxPA} - 75.145 \text{ MaxZL} - 76.773 \text{ J} + 7.307 \text{ DE} + 0.132 \text{ WW} + 585.256 \quad (4.1)$$

$$\Delta H^{\circ}_{f(\text{liquid})} = -2.143 \text{ PSA} + 21.900 \text{ H} - 4.923 \text{ V} - 25.839 \text{ MaxZL} + 357.763 \quad (4.2)$$

Equation 4.1 with multiple linear regression model shows that the $\Delta H^{\circ}_{f(\text{gas})}$ of studied amines can be determined using the obtained coefficients for molecular descriptors of Maximum Projection Area, Maximum Z Length, Wiener-Polarity, Dreiding-Energy, Balaban. Also Equation 4.2 shows that the $\Delta H^{\circ}_{f(\text{liquid})}$ of mentioned molecules can be determined using the obtained coefficients for molecular descriptors of Projection Surface Area, Harary, Maximum Z Length and Volume.

Table 4 shows the predicted and residual (literature-calculated) of the standard enthalpies of formation for gas and liquid phases of 33 different types of amines.

In Figures 2 and 3, the values of R^2 show the linear correlation between experimental and predicting values resulted from the model that is related to the dependent variable. The larger values of R^2 indicate a stronger correlation.

R^2 refers to the determination coefficient and represents the square of the multiple correlation coefficients. The obtained value for $R^2 = 0.9298$ in Figure 1. emphasizes the fact that 92.98% of change in $\Delta H^{\circ}_{f(\text{gas})}$ has been explained with 5th model in Table 4. Also the value of R^2 equal to 0.9285 in Figure 2 indicates that 92.85% of change in $\Delta H^{\circ}_{f(\text{liquid})}$ has been explained with 12th model in Table 3.

These results show that there is the high linear correlation between the experimental and standard enthalpies of formation obtained using Equations 4.1 and 4.2. The values of determination coefficient of above figures confirm that change in the dependent variable with a good approximate can be determined using the values of independent variables.

Table 5. The predicted and residual $\Delta H_{f(\text{gas})}^{\circ}/\text{kJmol}^{-1}$ and $\Delta H_{f(\text{liquid})}^{\circ}/\text{kJmol}^{-1}$ of the training set.

No	$\Delta H_{f(\text{gas})}^{\circ}$ pred	$\Delta H_{f(\text{gas})}^{\circ}$ [Residual]	$\Delta H_{f(\text{liquid})}^{\circ}$ Pred	$\Delta H_{f(\text{liquid})}^{\circ}$ Residual	No	$\Delta H_{f(\text{gas})}^{\circ}$ pred	$\Delta H_{f(\text{gas})}^{\circ}$ [Residual]	$\Delta H_{f(\text{liquid})}^{\circ}$ Pred	$\Delta H_{f(\text{liquid})}^{\circ}$ Residual
1	-30.636	-27.063	-59.91	-24.58	20	93.4281	10.37186	45.20	13.99
2	-	-	-54.81	64.80	21	-47.917	-50.7829	-113.21	-28.29
3	-70.431	0.53069	-103.74	2.24	22	-120.43	-73.4701	-200.67	-38.32
4	-74.786	-8.9137	-94.01	-18.29	23	-151.29	34.79249	-187.93	33.33
5	-	-	-76.25	30.53	24	-	-	-157.10	-11.89
6	11.1656	-34.8656	-34.39	-11.33	25	26.7642	26.43577	7.32	-12.04
7	-36.082	32.68201	-51.00	10.00	26	35.4288	25.67119	8.08	34.22
8	-102.9	7.900822	-143.11	15.41	27	36.3095	5.49046	6.16	12.83
9	-130.8	32.18318	-145.40	12.90	28	44.9229	38.97707	25.97	7.43
10	-111.63	5.632586	-	-	29	29.0198	-33.3198	-	-
11	-179.01	59.01339	-148.49	-2.10	30	-200.28	-15.7218	-252.09	-13.90
12	-96.797	-3.00289	-114.25	-16.74	31	-	-	0.34	-31.34
13	115.065	25.13491	78.17	21.79	32	-185.46	14.35947	-242.41	30.11
14	113.651	-10.5511	58.85	3.52	33	-114.2	-46.5986	-211.32	4.43
15	-38.674	-8.47554	-64.05	-22.79	34	221.935	-16.9347	-	-
16	-105.01	-3.28895	-145.37	4.17	35	101.844	82.56617	-	-
17	96.7445	-9.71449	34.80	-3.50	36	183.179	18.82063	125.81	6.19
18	48.296	-74.796	32.21	-101.20	37	-	-	110.08	10.41
19	87.3428	16.25723	43.49	17.60	38	346.321	-19.3205	-	-

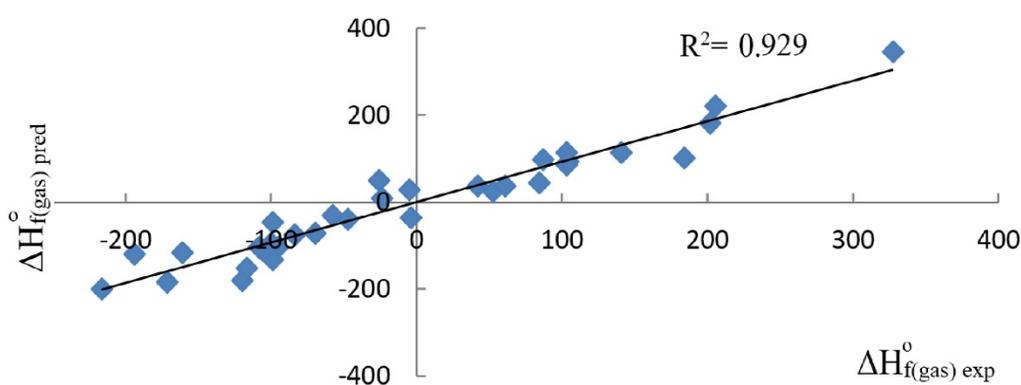


Figure 1. The diagram of the experimental versus predicted $\Delta H_{f(\text{gas})}^{\circ}/\text{kJmol}^{-1}$.

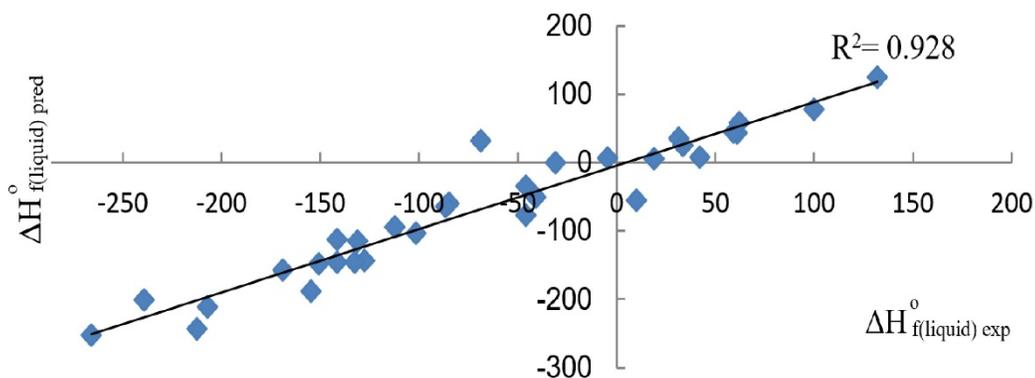


Figure 2. The diagram of the experimental versus predicted $\Delta H_{f(\text{liquid})}^{\circ} / \text{kJmol}^{-1}$.

To analyze the equal distribution of errors, curves indicating the relationship between experimental results and residuals were sketched. It was demonstrated that the errors around the axis X have an equal distribution indicating that the proposed method is an appropriate method for predicting the standard enthalpies of formation of amines. The comparison of experimental and predicted standard enthalpies of formation using the MLR model indicates that the Equations 3.1 and 3.2 might be used successfully to predict the $\Delta H_{f(\text{gas})}^{\circ}$, $\Delta H_{f(\text{liquid})}^{\circ}$ of studied organic molecules, respectively (see Figures 3 and 4).

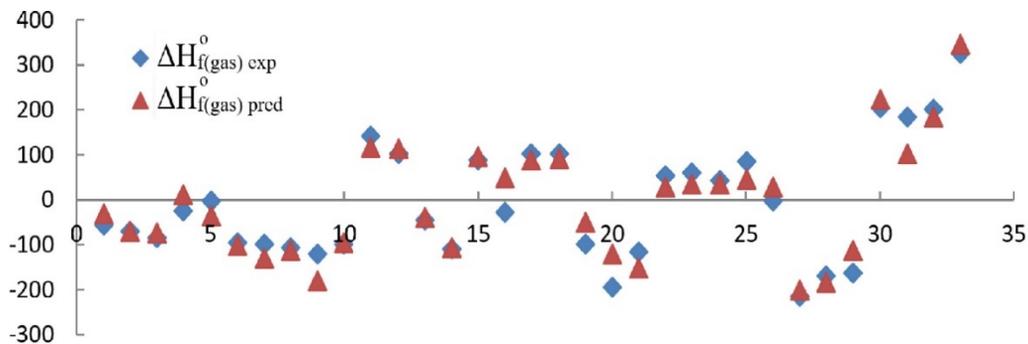


Figure 3. Comparison of the experimental and predicted $\Delta H_{f(\text{gas})}^{\circ} / \text{kJmol}^{-1}$.

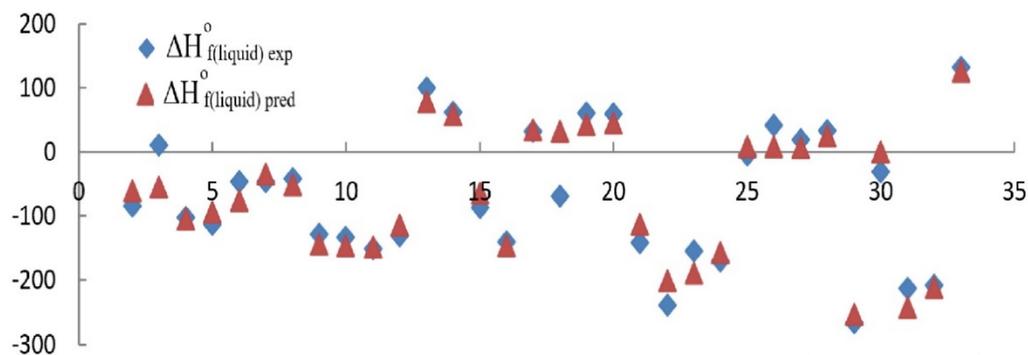


Figure 4. Comparison of the experimental and predicted $\Delta H_{f(\text{liquid})}^{\circ} / \text{kJmol}^{-1}$.

For evaluating the uniform distribution form of residuals, the scatter curves of the relation between residual values with the standard enthalpies of formation have been plotted. It was also shown that the residuals around x-axis have almost a uniform distribution which proves the suitability of the selected pattern for experimental values of the standard enthalpies of formation of studied amines (Figures 5 and 6).

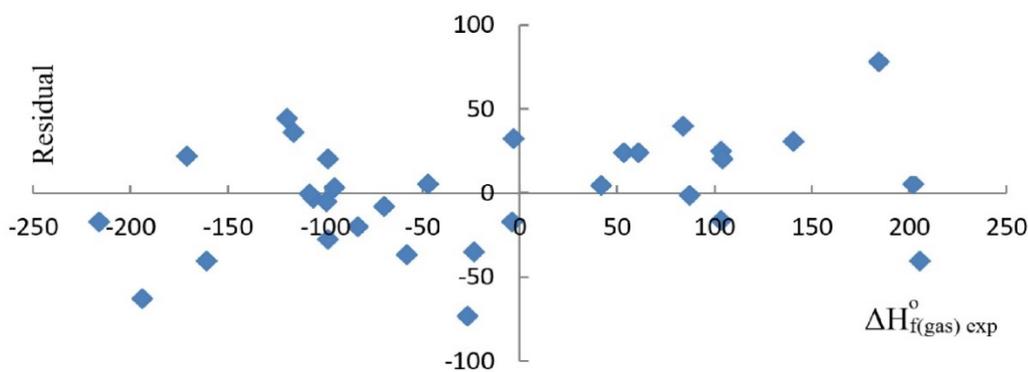


Figure 5. The scatter curves of the residuals against experimental $\Delta H_{f(\text{gas})}^{\circ} / \text{kJmol}^{-1}$ values.

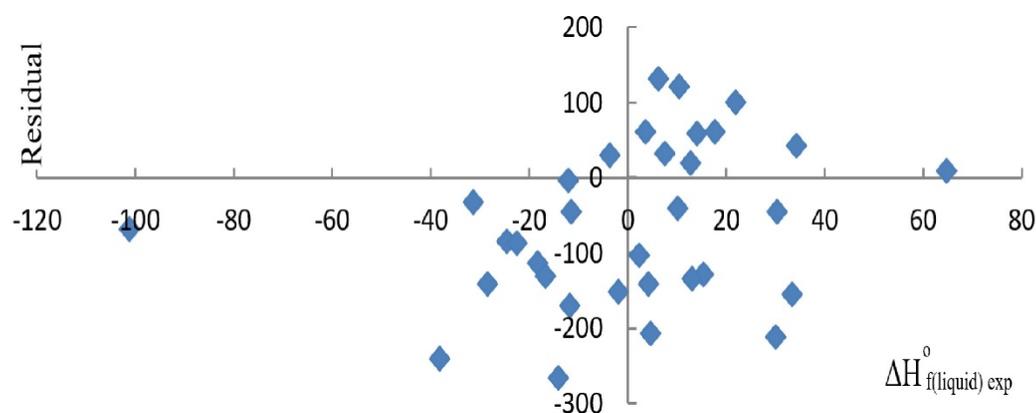


Figure 6. The scatter curve of the residuals against experimental $\Delta H_{f(\text{liquid})}^{\circ} / \text{kJmol}^{-1}$ values.

5. VALIDATION METHODS

5. 1. VIF TEST

A variance inflation factor (VIF) quantifies how much the variance is inflated. The VIF may be calculated for each predictor by doing a linear regression of that predictor on all the other predictors. The computation formula of VIF is: $1/(1-R^2)$.

The permissible values of VIF are between 1-10, confirming that there is no multicollinearity. The analysis of VIF values for all the descriptors indicate that the best descriptors for predicting $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ values are: PSA, H, V and MaxZL, and also J, DE, WW, MaxPA and MaxZL on the 5th and 12th models, respectively (See Table 6). As it can be seen in the Table 8, the higher VIF is, the lower the tolerance will be.

Table 6. The collinearity statistics.

	Descriptors	Tolerance	VIF	Beta
5th Model for predicting $\Delta H_{f(\text{gas})}^{\circ}/\text{kJmol}^{-1}$	MaxZL	0.741	1.349	-0.493
	MaxPA	0.282	3.544	-0.435
	DE	0.198	5.057	0.769
	J	0.622	1.607	-0.250
	WW	0.209	4.790	0.362
12th Model for predicting $\Delta H_{f(\text{liquid})}^{\circ}/\text{kJmol}^{-1}$	PSA	0.873	1.146	-0.172
	H	0.256	3.901	1.592
	V	0.217	4.599	-1.450
	MaxZL	0.740	1.351	-0.199

5. 2. β

The values of standard coefficients β of the best descriptors which were used for predicting $\Delta H_{f(\text{gas})}^{\circ}$ and $\Delta H_{f(\text{liquid})}^{\circ}$ in the Equations 4.1 and 4.2 were listed in Table 6. The standard correlation coefficient β value is a measure of how strongly each predictor variable influences the dependent variable. According to the data in the Table 6, and the DE variable which is used for predicting $\Delta H_{f(\text{gas})}^{\circ}$, the value of β equals to 0.769 indicate that compared to MaxZL, MaxPA, J, WW and PSA predictor variables, the DE index has the strongest influence on $\Delta H_{f(\text{gas})}^{\circ}$. Also, the value of $\beta=1.592$ reveals that the descriptor H has the highest influence on dependent variable $\Delta H_{f(\text{liquid})}^{\circ}$ than the PSA, V and MaxZL predictors.

5. 3. LEAVE-ONE-OUT (LOO) CROSS-VALIDATION

To determine the LOO cross validation, at first, a molecule from training set of 33 primary amines was removed and then QSPR was examined on the 32 remaining molecules. Considering the amount of RSS and TSS, the amount of Q^2 was calculated based on Formula (5.1) and this cycle was repeated resulting in elimination of 25% of amines which were being studied. This resulted in examination and calculation of Q^2 for the remaining amines.

$$Q^2 = 1 - \frac{\sum (Y_{\text{exp}(\text{train})} - Y_{\text{pred}(\text{train})})^2}{\sum (Y_{\text{exp}(\text{train})} - \bar{Y}_{\text{Training}})^2} = 1 - \frac{RSS}{TSS} \quad (5.1)$$

where, RSS is the residual sum of squares and TSS represents the total sum of square. This formula is the most widely used measure of the ability of a QSPR model to reproduce the data in the training set. We have computed the values of Q^2 (Equation 5.1) using %25 of training set randomly. The values of Q^2 are defined as positive and less than one. (see Table 7).

Table 7. The Q^2 LOO values of $\Delta H_{f(\text{gas})}^\circ/\text{kJmol}^{-1}$ and $\Delta H_{f(\text{liquid})}^\circ/\text{kJmol}^{-1}$.

No	$Q^2_{\text{LOO}}/\Delta H_{f(\text{liquid})}^\circ$	$Q^2_{\text{LOO}}/\Delta H_{f(\text{gas})}^\circ$	No	$Q^2_{\text{LOO}}/\Delta H_{f(\text{liquid})}^\circ$	$Q^2_{\text{LOO}}/\Delta H_{f(\text{gas})}^\circ$
1	0.9281	0.9297	5	0.9301	0.9423
2	0.9284	0.9298	6	0.9582	0.9522
3	0.9287	0.9318	7	0.9622	0.9553
4	0.9300	0.9320	8	0.9626	0.9559

The small difference between mean Q^2 values of $\Delta H_{f(\text{gas})}^\circ$ and $\Delta H_{f(\text{liquid})}^\circ$ are equal to 0.9410 and 0.9411 respectively, and the R^2 values of them verify the precision and accuracy of the model for predicting the mentioned thermodynamic functions.

5. 4. APPLICABILITY DOMAIN

Applicability domain (AD) determines the degree of generalization of an obtained predictive model. The AD is an important tool for reliable application of a QSPR model. This study reports that the AD can be easily used for identification of the X-outliers for training set amines and detection of the test amines residing outside the AD through the use of molecular descriptors. The leverage approach along with the Williams Plot was used to determine the applicability domain in predictive QSAR models. To create the Williams Plot, the leverage h_i for each compound in training set, in which QSPR model was used to predict its property, was calculated according to the equation (5.2):

$$h_i = x^T (X^T X)^{-1} x \quad (5.2)$$

where, x represents the descriptor vector of the considered compound and X represents the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) was determined as:

$$h^* = 3(p+1)/n \quad (5.3)$$

where n is the number of training compounds and p is the number of predictor variables. According to the equation (5.3), the values of h^* in 12th and 5th models (in which, $n=33$ and $p=4$ and 5) will be equal to 0.454 and 0.545, respectively.

In this research, in 12th and 5th models, the descriptor vector x includes the PSA, H, MaxZL, V and MaxZL MaxPA, DE, J, WW descriptors and X is PSA, H, MaxZL, V and MaxZL, MaxPA, DE, J, WW descriptors matrix related to training set of amines, respectively. (The leverage values are shown in Table 8).

Then, the defined applicability domain (AD) was visualized through the use of a Williams plot, the plot of the standardized residuals versus the leverage values (h). A compound with $h_i > h^*$ seriously influences the regression performance and may be excluded from the applicability domain, see Figures 7 and 8. The results indicated that among training set of 33 amines, there is just one and two outlier for prediction of $\Delta H_{f(\text{liquid})}^\circ$ and $\Delta H_{f(\text{gas})}^\circ$, respectively.

Table 8. The leverage values of the training set.

No.	$h/\Delta H_{f(\text{liquid})}^\circ$	$h/\Delta H_{f(\text{gas})}^\circ$	No.	$h/\Delta H_{f(\text{liquid})}^\circ$	$h/\Delta H_{f(\text{gas})}^\circ$
1	0.1283	0.1918	18	0.0564	0.0599
2	0.0674	0.0779	19	0.0623	0.0760
3	0.0703	0.0644	20	0.0518	0.0610
4	0.0576	0.0965	21	0.1183	0.0699
5	0.0859	0.1677	22	0.0692	0.0481
6	0.1723	0.0389	23	0.1335	0.0525
7	0.1564	0.0592	24	0.1001	0.0470
8	0.0527	0.0457	25	0.1090	0.0397
9	0.0670	0.2521	26	0.1056	0.6654
10	0.1057	0.0348	27	0.0505	0.1683
11	0.0484	0.1486	28	0.1753	0.2217
12	0.1320	0.1605	29	0.0324	0.4169
13	0.1388	0.0261	30	0.1777	0.2626
14	0.0275	0.0267	31	0.3195	0.1386
15	0.0432	0.1031	32	0.2707	0.1867
16	0.1387	0.0306	33	0.6401	0.9050
17	0.0333	0.0541			

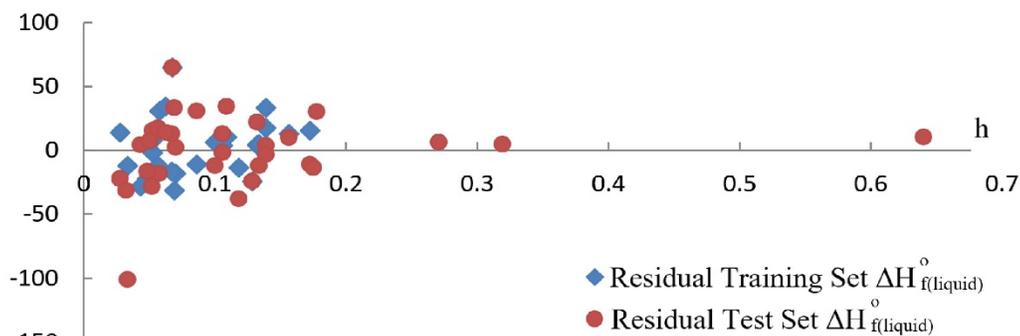


Figure 7. The Williams plot of residual $\Delta H_{f(\text{liquid})}^\circ / \text{kJmol}^{-1}$ versus leverage.

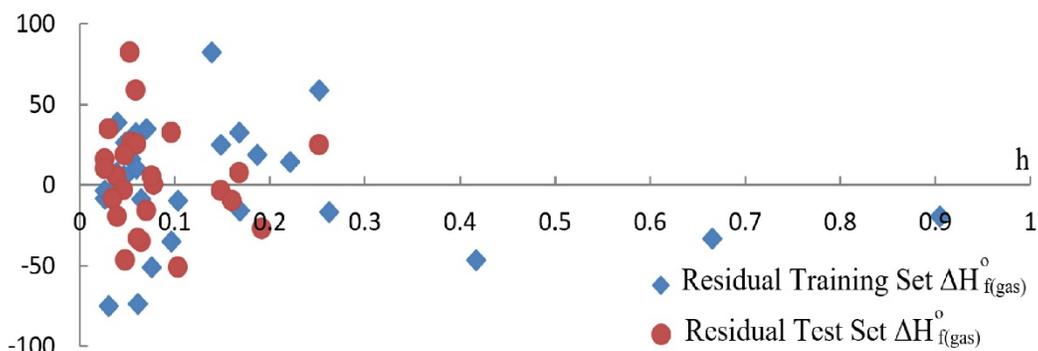


Figure 8. The Williams plot of residual $\Delta H^{\circ}_{f(\text{gas})}/\text{kJmol}^{-1}$ versus leverage.

5. 5. RESIDUALS SKEWNESS

The normality of residuals was analyzed for obtaining more confidence in the accuracy of applied method. The skewness is a function indicating the symmetry or asymmetry of the distribution function. In general, if the skewness is placed at an interval between $[2, -2]$, the data follow a normal distribution. The observed values 1.082, and 0.429 for residuals skewness of $\Delta H^{\circ}_{f(\text{liquid})}$ and $\Delta H^{\circ}_{f(\text{gas})}$ indicate their normality, respectively.

5. 6. DW

In a regression analysis, the Durbin–Watson statistic is used to detect the presence of autocorrelation in the residuals. The most optimum variation interval of Durbin-Watson coefficient is between 0-4. The values DW equal to 2.349 and 1.919 of 12th and 5th models in Tables 3 and 4 are acceptable values indicating the lower correlation between the errors and the independence of the residuals respectively.

5. 7. CORRELATION

The predictive power, firmness and reliability of a QSPR model depends on the used molecular descriptors. Correlation measures the strength and direction of a linear relationship. As it's seen in Tables 9 and 10, the values of correlations between descriptors of final models indicate that descriptors in selective models had been chosen correctly.

Table 9. Correlations between descriptors of 12th model of the Table 2.

Correlations	PSA	H	V	MaxZL
PSA	1			
H	-0.334	1		
V	-0.343	0.645	1	
MaxZL	-0.075	0.199	0.418	1

Table 10. Correlations between descriptors of 5th model of the Table 3.

Correlations	J	MaxZL	DE	WW	MaxPA
J	1				
MaxZL	0.137	1			
DE	-0.403	0.242	1		
WW	-0.307	0.382	0.651	1	
MaxPA	-0.074	0.214	0.615	0.769	1

Therefore, various statistical coefficients, tables and diagrams analyzed in the multiple linear regression method possess the necessary and sufficient validity for predicting the thermodynamic functions of amines.

6. CONCLUSIONS

In this study, a model was presented for predicting the enthalpies formation of amines in two gas and liquid phases. This model was based on topological and geometrical descriptors and the important rule of these molecular descriptors were specified for predicting the studied thermodynamic functions in this class of organic molecules. According to the result, it can be seen that there is good correlation between the $\Delta H_{f(\text{gas})}^{\circ}$ of studied amines in this report with molecular descriptors of MaxPA, MaxZL, Wp, DE and J. Also it can be identified that the $\Delta H_{f(\text{liquid})}^{\circ}$ of mentioned molecules can be determined using the obtained coefficients for molecular descriptors of PSA, H, MaxZL and V.

REFERENCES

1. M. A. Sharaf, D. L. Illman and B. R. Kowalski, *Chemometrics, Chemical Analysis Series*, Wiley, New York, 82, 1986.
2. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufmann, *Chemometrics—a Textbook*, Elsevier, Amsterdam, 1988.
3. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry*, Wiley-Vch Verlag GmbH, 2008.
4. H. Martens and T. Naes, *Multivariate Calibration*, New York, Wiley, 1989.
5. W. G. Hunter, *Statistics and Chemistry and the Linear Calibration Problem*, In B. R. Kowalski, *Chemometrics: Mathematics and Statistics in Chemistry*, Boston: Riedel, 1984.
6. J. F. Mac Gregor and T. Kourti, Statistical control of multivariate processes, *Control Engineering Practice* **3** (3) (1995) 403–414.

7. S. Ahmadi and E. Habibpour, Application of GA-MLR for QSAR modeling of the arylthioindole class of tubulin polymerization inhibitors as anticancer agents, *Anti-Cancer Agents Med. Chem.* **17** (4) (2017) 552–565.
8. M. B. Smith and J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure*, Wiley-Interscience, New York, 2007.
9. G. W. Snedecor and G. W. Cochran, *Statistical Methods*, Oxford and IBH, New Delhi, 2010.
10. I. Gutman and O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer-Verlag, Berlin, 1986.
11. M. Randić and S. C. Basak, Multiple regression analysis with optimal molecular descriptors, *SAR QSAR Environ.* **11** (1) (2000) 1–23.
12. J. Sangster, Octanol-water partition coefficients of simple organic compounds, *J. Phys. Chem. Ref. Data* **18** (3) (1989).
13. E. Mohammadinasab, Determination of critical properties of alkanes derivatives using multiple linear regression, *Iranian J. Math. Chem.* **8** (2) (2017) 199–220.
14. M. Goodarzi and M. Mohammadinasab, Theoretical investigation of relationship between quantum chemical descriptors, Topological indices, energy and electric moments of zig-zag polyhex carbon nanotubes TUHC6 [2p,q] with various circumference [2p] and fixed lengths, *Fullerenes Nanotubes Carbon Nanostructures* **21** (2013) 102–112.
15. H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69** (1947) 17–20.
16. D. J. Klein, Z. Mihalic, D. Plavsic and N. Trinajstić, Molecular topological index: a relation with the Wiener index, *J. Chem. Inf. Computer. Sci.* **32** (4) (1992) 304–305.
17. G. Cash, S. Klavžar and M. Petkovšek, Three methods for calculation of the hyper-Wiener index of molecular graphs, *J. Chem. Inf. Comput. Sci.* **42** (2002) 571–576.
18. H. Wang, On the extremal Wiener-polarity index of Hückel graphs, *Comput. Math. Methods Med.* **2016**, Article ID 3873597, 6 pages.
19. J. Baskar Babujee, Topological indices and new graph structures, *Appl. Math. Sci.* **6** (108) (2012) 5383–5401.
20. M. Randic, Characterization of molecular branching, *J. Am. Chem.* **97** (23) (1975) 6609–6615.
21. A. T. Balaban and T. S. Balaban, New vertex invariant and topological indices of chemical graphs based on information on distance, *Math. Chem.* **8** (1991) 383–397.

22. A. T. Balaban, Topological index based on topological distances in molecular graph, *Pure Appl. Chem.* **55** (1983) 199–206.
23. K. Xu and K. Ch. Das, On Harary index of graphs, *Discrete Appl. Math.* **159** (2011) 1631–1640.
24. P. V. Khadikar, N. V. Deshpande, P. P. Kale, A. Dobrynin, I. Gutman and G. Domotor, The Szeged index and an analogy with the Wiener index, *J. Chem. Inf. Comput. Sci.* **35** (1995) 547–550.
25. Web search engine developed by ChemAxon, *software available at <http://www.chemicalize.org>*.
26. J. Devillers and A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science, Netherlands, 1999.
27. I. Gutman and B. Furtula (eds), *Novel Molecular Structure Descriptors-Theory and Applications*, University of Kragujevac and Faculty of Science, Kragujevac, 2010.