Journal of Physical and Theoretical Chemistry
of Islamic Azad University of Iran, 16 (3, 4) 117-131: Fall 2019 & Winter 2020
(J. Phys. Theor. Chem. IAU Iran)
ISSN 1735-2126

QSPR models to predict thermodynamic properties of some mono and polycyclic aromatic hydrocarbons (PAHs) using GA-MLR

Fatemeh Dialamehpour and Fatemeh Shafiei*

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

Received February 2020; Accepted July 2020

ABSTRACT

Quantitative Structure-Property Relationship (QSPR) models for modeling and predicting thermodynamic properties such as the enthalpy of vaporization at standard condition (ΔH˚vap kJ mol-1) and normal temperature of boiling points (T˚bp K) of 57 mono and Polycyclic Aromatic Hydrocarbons (PAHs) have been investigated. The PAHs were randomly separated into 2 groups: training and test sets. A set of molecular descriptors was calculated for selected compounds using the Dragon software. The Genetic Algorithm (GA) method and backward stepwise regression were used to select the suitable descriptors. Multiple Linear Regression (MLR) technique was used to obtain a linear relationship between descriptors and chemical properties. The predictive ability of the GA-MLR models was implemented using squared cross-validation and external validation methods. The aforementioned results and discussion lead us to conclude that the training set models established by GA-MLR method have good correlation of thermodynamic properties, which means QSPR models could be efficiently used for estimating and predicting of the above mentioned properties of the mono and PAHs.

Keywords: polycyclic aromatic hydrocarbons (PAHs); quantitative structure-property relationships (QSPR); normal temperature of boiling points; first Zagreb index

INTRODUCTION

Polycyclic aromatics hydrocarbons (PAHs) are a class of chemicals that occur naturally in coal, crude oil, and gasoline. They also are produced when coal, oil, gas, wood, garbage, and tobacco are burned. Cigarette smoke contains many PAHs. The major source of PAHs is the incomplete combustion of organic material such as coal, oil and wood [1, 2].

Some PAHs are well known as carcinogens, mutagens, and teratogens and therefore pose a serious threat to the health and the well-being of humans. The most significant health effect to be expected from inhalation exposure to PAHs is an excess risk of lung cancer [3-5].

Quantitative structure-property relationships (QSPR) models are

*Corresponding author: f-shafiei@iau-arak.ac.ir
shafa38@yahoo.com

mathematical equations that relate properties such as the various physical and physicochemical properties of compounds to a wide range of molecular descriptors [6-8].

Molecular descriptors are of outstanding importance in the research fields of QSPR/QSAR, where they are the independent chemical information used to predict the properties/activities of compounds [9].

It is well known that a single molecular descriptor is unable to carry all the information of the molecular structure and thus sets of suitable and relevant descriptors for a particular response must be selected [10,11].

Relationship between vapor pressure and molecular descriptor of PAHs has been investigated [12].

QSPR model has been investigation for predicting the vapor pressure of typical PAHs such as benzo(a)pyrene with the lowest vapor pressure and naphthalene with the highest vapor pressure using molecular weight descriptor [13,14].

The multiple linear regression (MLR), artificial neural network (ANN), and support vector machine (SVM) were applied to study the relationship between adsorption coefficients and physico-chemical properties of 39 aromatic compounds [15].

Quantitative structure-activity relationship (QSAR) models have been used to determine activity of PAHs using information indices [16].

QSAR method has been applied to predict mutagenicity of 48 nitrated polycyclic aromatic hydrocarbons (nitro-PAHs) [17].

Several activities of PAHs, such as carcinogenesis, mutagenicity, phototoxicity, and biocatalytic oxidation, have been studied using QSAR analyses and molecular descriptors [18].

2D and 3D QSAR models have been used to study relationship between biological activities (antidepressants and antipsychotics) and chemical structures of PAHs [19].

A QSPR study to predict polarizability of 40 PAHs and fullerenes using molecular descriptor has been researched [20].

QSAR model to correlate the photolysis half-lives of PAHs with their quantum chemical descriptors by partial least squares (PLS) method has been developed [21].

MLR method has been used to construct QSPR model for the prediction of boiling point of 61 PAHs [22].

Physicochemical and thermodynamic properties of organic pollutant play an important key role to understand their behavior in environment. However, the information behind the property-behavior phenomena of chemical compounds is less found in the literature. Therefore, computational methods had to be applied for process optimization. In the present study the applicability of the QSPR models based on molecular descriptors derived from molecular structures have been developed for the prediction of thermodynamic properties of 57 mono and PAHs such as the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) and normal temperature of boiling points (T°_{bp} K). For this purpose genetic algorithm - multiple linear regressions (GA-MLR) were used to select the suitable descriptors for construct QSPR models.

MATERIALS AND MATHEMATICAL METHODS

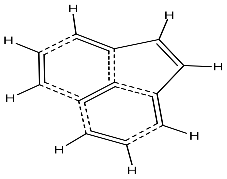
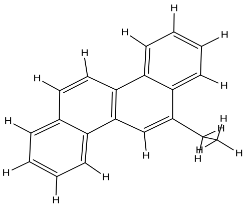
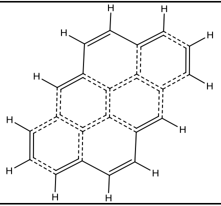
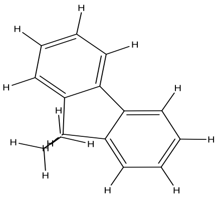
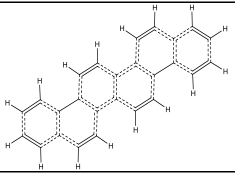
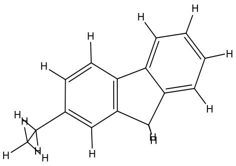
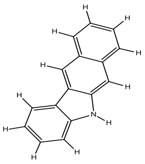
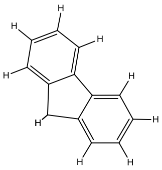
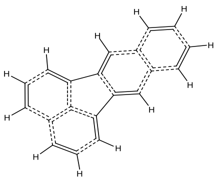
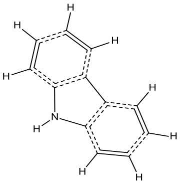
Mono and PAHs are used in the manufacture of cellulose esters, fibers, plastics, lacquers, drugs, disinfectants, cosmetics, dyestuffs, anti-icers, corrosion inhibitors, *etc* [23]. The name and chemical structure of the mono and PAHs discussed in this study are shown in Table **1.** Thermodynamic properties such as the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) and normal temperature of boiling points (T°_{bp} K) of 57 mono and Polycyclic Aromatic Hydrocarbons (PAHs) are taken from national institute of standards and technology (NIST) chemistry and chemspider web book, respectively. These properties are listed in Table **2**. The 57 datasets were randomly divided into 2 groups: training and test sets consisting of 47, 10 data point, respectively.

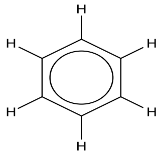
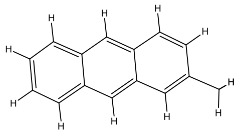
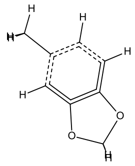
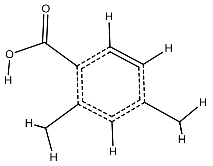
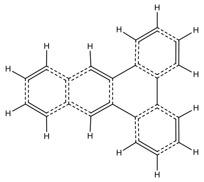
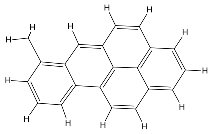
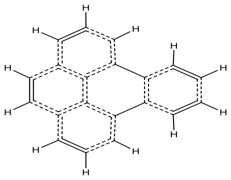
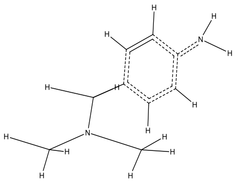
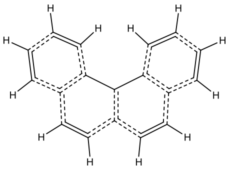
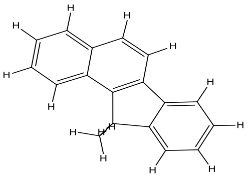
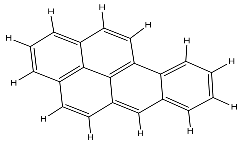
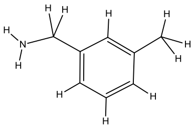
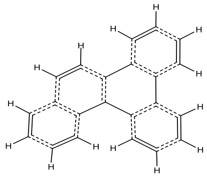
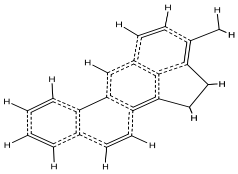
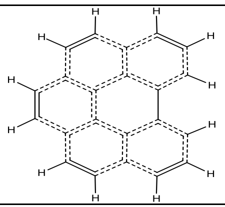
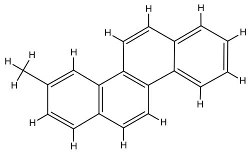
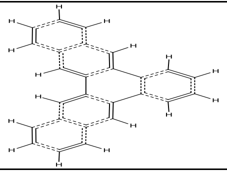
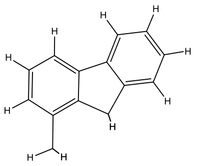
The chemical structures of molecules were drawn by Gauss View 05 program and then they were optimized with Gaussian09 using Hartree–Fock (HF) level of theory and 6-311G* basis set method.

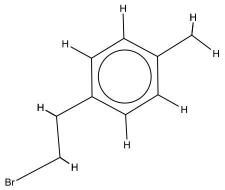
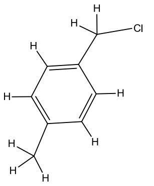
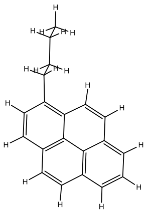
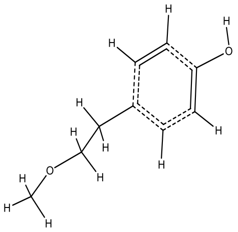
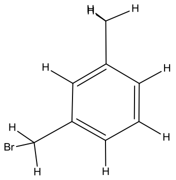
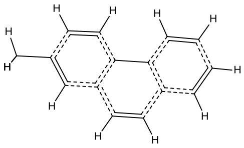
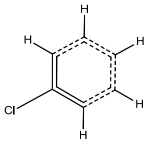
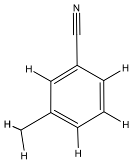
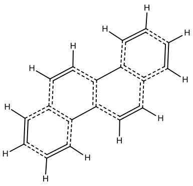
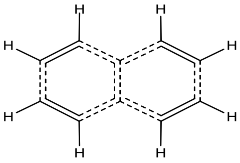
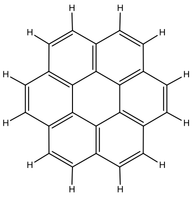
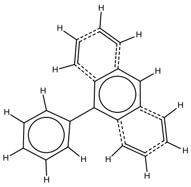
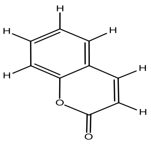
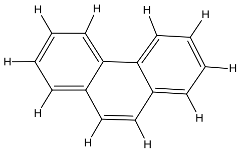
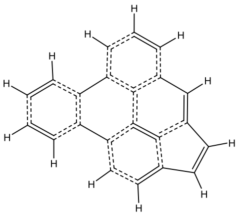
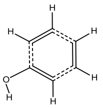
A set of descriptors was calculated for selected compound using the Talete srl, Dragon for Windows Version 5.4- 2006 package. A lot of descriptors include different categories like topological, Getaway, 3D-MoRSE, constitutional, and molecular properties which have been used [24].

The genetic algorithm (GA) is written in MATLAB (version 2010a) environment and backward stepwise regressions have been used to decrease the number of descriptors. The software package SPSS 21.0 for Windows is used to implement multilinear regression [25].

Table 1. Iupac Name and chemical structure of 57 mono and PAHs used in present study

NO	Iupac Name	Structure	NO	Iupac Name	Structure
1	Acenaphthylene		30	6-Ethylchrysene	
2	Anthanthrene		31	9-Ethylfluorene	
3	Benzo(C)picene		32	2-Ethyl-9H-fluorene	
4	5H-Benzo[b]carbazole		33	Fluorene	
5	Benzo[k]fluoranthene		34	9H-carbazole	

6	Benzene		35	2-Methylanthracene	
7	5-Methyl-1,3-benzodioxole		36	2,4-Dimethylbenzoic acid	
8	Benzo[b] triphenylene		37	7-Metylbenzo[A] pyrene	
9	Benzo[e] pyrene		38	4-[(Dimethylamino) methyl]aniline	
10	Benzo[c] phenanthrene		39	11-Methylbenzo(a) fluorene	
11	Benzo[a] pyrene		40	3-Methylbenzylamine	
12	Benzo[G] chrysene		41	3-Methylcholanthrene	
13	Benzo[ghi] perylene		42	3-Methylchrysene	
14	Benzo[h] pentaphene		43	1- Methylfluorene	

15	1-(2-Bromoethyl)-4-methylbenzene		45	(4-Methylphenyl) methyl chloride	
16	1-Butylpyrene		46	4-(2-Methoxyethyl) phenol	
17	Alphabromomxylene			2-Methylphenanthrene	
18	Chlorobenzene		47	m-Tolunitrile	
19	Chrysene		48	Naphthalene	
20	Coronene		49	9-Phenylanthracene	
21	Coumarin		50	Phenanthrene	
22	Cyclopentabenzo(E) Pyrene		51	Phenol	

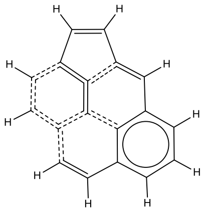
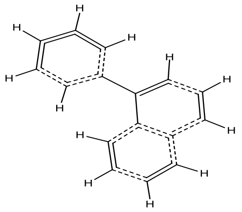
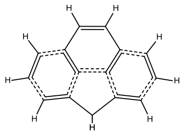
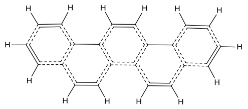
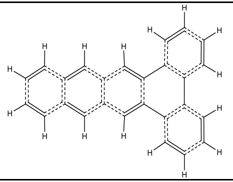
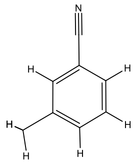
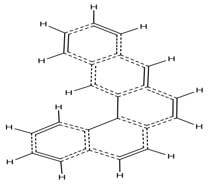
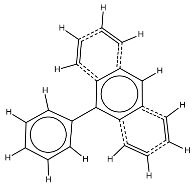
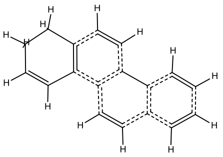
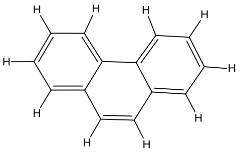
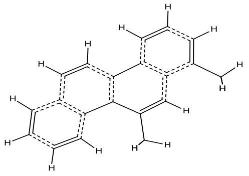
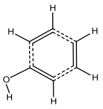
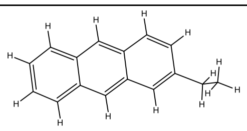
23	Cyclopenta[cd] pyrene		52	1-Phenylnaphthalene	
24	4H-Cyclopenta[def]phenanthrene		53	Picene	
25	Dibenzo[a,c] naphthacene		54	2-(Propan-2- yl) aniline	
26	Dibenzo[b,g] phenanthrene		55	Pyrene	
27	1,2-Dihydrochrysene		56	Triphenylene	
28	1,11-Dimethylchrysene		57	9-Vinylanthracene	
29	2-Ethylanthracene				

Table 2. The observed, predicted and residuals values for training and test sets of mono and PAHs using Equations 8,9 (*Compounds selected for test set in external validation procedure)

NO	Observed (T°_{bp} K)	Predicted	Residual	Observed (ΔH°_{vap} kJ mol^{-1})	Predicted	Residual
1	572.05	566.574	5.476	51.70	51.99	-0.29
2	770.25	774.249	-3.999	73.60	76.61	-3.01
3	*877.25	*869.383	*7.867	86.50	90.71	-4.21
4	729.15	721.283	4.523	*68.90	*65.07	* 3.83
5	753.15	748.627	8.792	71.60	73.31489	-1.71
6	*351.95	*359.916	*-7.966	*30.70	*29.63	* 1.07
7	*745.85	*753.816	*-1.8	54.90	49.62	5.28
8	791.15	792.95	1.479	76.10	72.70	3.39

NO	Observed (T°_{bp} K)	Predicted	Residual	Observed (ΔH°_{vap} kJ mol^{-1})	Predicted	Residual
9	740.65	744.357	-3.707	70.20	69.62	0.57
10	709.85	707.612	2.238	66.70	63.64	3.05
11	768.15	765.912	2.379	73.40	69.90	3.50
12	*797.85	*795.471	*-2.776	76.90	76.24	0.66
13	774.15	776.926	-2.709	74.10	73.39	0.71
14	877.25	880.026	1.199	86.50	81.94	4.56
15	499.15	495.416	3.734	*44.40	*44.81	* -0.41
16	693.65	691.023	2.627	64.80	63.59	1.21
17	484.65	481.654	2.996	43.00	41.43	1.56
18	*405.15	*400.163	* 4.987	*35.20	*37.05	* -1.85
19	721.15	717.782	3.368	67.90	69.27	-1.37
20	798.15	804.149	-5.999	77.00	80.76	-3.76
21	570.15	576.548	-6.398	*53.80	*53.69	* 0.11
22	800.15	801.248	-1.098	77.2	73.49	3.71
23	711.45	712.356	-0.906	66.8	68.93	-2.13
24	626.15	630.576	-4.42599	57.4	55.78	1.62
25	877.25	876.051	1.199	*86.5	*84.83	* 1.67
26	797.85	795.701	2.149	76.90	80.87	-3.97
27	692.95	690.624	2.326	64.70	64.51	0.19
28	733.35	731.607	1.743	69.30	67.56	1.74
29	637.15	635.426	1.724	58.60	60.92	-2.32
30	730.35	728.196	2.154	69.00	64.63	4.37
31	594.85	591.464	3.386	54.10	52.20	1.90
32	*598.15	*595.881	*2.269	54.20	55.02	-0.82
33	568.15	573.685	-5.535	51.20	49.89	1.31
34	628.15	624.837	3.313	57.60	53.26	4.33
35	626.65	625.265	1.385	*57.50	*56.60	* 0.90
36	541.95	536.233	5.717	53.60	47.90	5.69
37	752.55	758.58	-6.03	71.50	69.15	2.35
38	504.95	506.134	-1.184	46.80	48.27	-1.47
39	673.45	673.245	0.205	*62.60	*65.96	* -3.36
40	473.35	470.571	2.779	43.60	48.47	-4.87
41	779.55	775.676	3.874	74.70	68.52	6.18
42	722.55	724.375	-1.825	68.10	68.52	-0.42
43	471.75	472.911	-1.161	41.70	50.40	-8.70
44	471.75	470.14	1.61	41.70	46.87	-5.17
45	513.05	518.757	-5.707	*49.6	*47.90	* 1.70
46	611.95	612.797	-0.847	55.90	60.48	-4.58
47	486.95	490.29	-3.34	45.00	43.53	1.47
48	494.65	496.153	-1.503	43.90	50.34	-6.44
49	690.15	693.082	-2.932	64.60	60.77	3.83
50	610.55	609.353	1.197	55.80	57.14	-1.34
51	*454.95	*458.479	*-3.529	43.50	45.35	-1.854
52	609.55	611.652	-2.102	55.70	59.40	-3.70
53	792.15	790.571	1.579	76.20	78.50	-2.30
54	498.75	500.859	-2.109	46.20	44.84	1.36
55	677.15	674.982	2.168	63.00	61.53	1.47
56	*698.15	*696.869	*1.281	*65.30	*68.96	* -3.66
57	*650.15	649.164	*0.986	60.00	59.61	0.38

RESULTS

Statistical coefficients

In order to build and test models, a data set of 57 compounds was randomly separated into a training set of 47 compounds, which was used to build the model and a test set of 10 compounds, which was used to evaluate the built model. The obtained models were evaluated by statistical parameters, such as squared multiple correlation coefficient (R^2) adjusted correlation coefficient (R^2adj), Fisher ratio (F), Root Mean Square Error (RMSE), Durbin-Watson statistic (D) and significance (Sig).

The squared multiple correlation coefficient (R^2) [26] is defined by the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{RSS}{TSS} \qquad (1)$$

where TSS is Total Sum of Squares; RSS: Residual Sum of Squares; y_i is the observed property, $\hat{y}_i$ is the property predicted by the model, and $\bar{y}_i$ is the average property.

The R^2 value increases when the number of variables in the model increases, while the adjusted R^2 value increases only if the new variables improve the model more than expected by chance. Therefore, the adjusted R^2, which is defined below, was also used [27]:

$$R^2_{adj} = R^2 - \frac{p(1-R^2)}{(n-p-1)} \qquad (2)$$

where p is the total number of regressors in the model, n is the sample size, and R^2 is the correlation coefficient.

The RMSE [28] for the training or prediction sets was calculated as follows:

$$RMSE = \sqrt{\frac{\sum(y_{obs} - y_{pred})^2}{n}} \qquad (3)$$

In Equation (3), y_{pred} and y_{obs} indicate predicted and observed property values respectively.

A linear regression equation significancy is drawn from Fisher's coefficient (F) [27]. It obtains by the following equation:

$$F = \frac{n-p-1}{p} \times \frac{ESS}{RSS} \qquad (4)$$

In Equation (4), n is number of molecules; p is number of explanatory variables.

QSPR models

The GA–MLR analysis led to the derivation of 4 models for the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}), with 7-10 descriptors (Table **3**). The statistical parameters of the models are almost the same; so, the model 4, which has the lowest number of descriptors, has been chosen. This model includes the seven descriptors namely: R7v, RDF080u, EEig13r, SP20, IC0, GGI5 and Mor08p. With the selected descriptors, we have built the linear model using the training set data, and obtained the following equation:

$$\Delta H^{\circ}_{vap} = -19.244 + 439.881\ (R7v) - 0.349\ (RDF080u) + 6.550\ (EEig13r) + 1.293\ (SP20) + 13.913\ (IC0) + 7.185\ (GGI5) - 6.370\ (Mor08p) \qquad (5)$$

N=47, R=0.991, R^2=0.981, R^2_{adj} =0.975, F=147.592, DW=1.688, Sig=0.000, RMSE= 1.992

The linear models for the normal temperature of boiling points (T°_{bp} K) contain 8 -10 descriptors. Table 4 shows the regression parameters and statistical parameters models for the thermal energy of 47 mono and PAHs. The best linear model for T°_{bp} includes eight descriptors which is: Jhetp, EEig15x, E2m, EEig13r, nCb, ZM1V, EEig13x and$ATS2e_u$. The model is presented below:

$$T^{\circ}_{bp} = 202.122 + 53.092\ (Jhetp) + 0.744\ (ZM1V) + 128.541\ (EEig15x) - 91.724\ (EEig13x) + 34.289\ (EEig13r) + 217.075\ (ATS2e_u) - 456.520\ (E2m) - 15.897\ (nCb) \quad (6)$$

N=47, R=0.985, R^2=0.971, R^2adj =0.964, F=152.577, DW=1.797, Sig=0.000, RMSE= 4.781

DISCUSSION

In this paper, we have carried out a QSPR analysis to derive a quantitative relationship between chemical structure of 57 mono and PAHs and their thermodynamic properties. In this step, to find the best model for predicting the mentioned properties, we will use the following sections.

Collinearity and Multicollinearity

It can be seen that the correlation coefficient of each of the models near to 1, but in regression analysis collinearity and multicollinearity should be checked. Collinearity and multicollinearity occur when two or more than two independent variables (molecular descriptors) are inter-correlated.

Table 3. Statistical parameters of the models calculated with the SPSS software for ΔH°_{vap} ($kJ.mol^{-1}$)

Model	Independent Variable	R	R^2	R^2_{adj}	RMSE	F	Sig
1	Mor12p, R7v, RDF080u, EEig13r, SP20, IC0, Mor14p, GGI5, Mor08p, ITH,	0.992	0.984	0.977	1.910	128.781	0.000
2	Mor12p, R7v, RDF080u, EEig13r, SP20, IC0, Mor14p, GGI5, Mor08p	0.992	0.983	0.976	1.935	134.329	0.000
3	Mor12p, R7v, RDF080u, EEig13r, SP20, IC0, GGI5, Mor08p,	0.991	0.982	0.975	1.953	141.875	0.000
4	**R7v** (R autocorrelation of lag 7 / weighted by van der Waals volume), **RDF080u** (Radial Distribution Function - 080 / unweighted), **EEig13r** (Edge adjacency indices), **SP20** (shape profile no. 20), **IC0** (Information Content index (neighborhood symmetry of 0-order)), **GGI5** (topological charge index of order 5), **Mor08p** (signal 08 / weighted by polarizability)	0.991	0.981	0.975	1.992	147.592	0.000

Table 4. Statistical parameters of the models calculated with the SPSS software for T°_{bp} (K)

Model	Independent Variable	R	R^2	R^2_{adj}	RMSE	F	Sig
1	Jhetp, EEig15x, E2m, BELp6, EEig13r, nCar, nCb, ZM1V, EEig13x, $ATS2e_s$	0.987	0.974	0.966	4.249	128.641	0.000
2	Jhetp, EEig15x, E2m, BELp6, EEig13r, nCb, ZM1V, EEig13x, $ATS2e_t$	0.986	0.972	0.964	4.696	136.830	0.000
3	**Jhetp** (2D matrix-based descriptors Barysz matrix weighted by polarizability (Dz(p)), **EEig15x** (Edge adjacency indices), **E2m (**2nd component accessibility directional WHIM index / weighted by mass), **EEig13r** (Edge adjacency indices), nCb (number of substituted benzene C (sp2), **ZM1V** (first Zagreb index by valence vertex degrees), **EEig13x** (Edge adjacency indices), **ATS2eu (**Broto-Moreau autocorrelation of lag 2 (log function)	0.985	0.971	0.964	4.781	152.577	0.000

Good regression model should not exist in a correlation between the independent variables or should not have happened multicollinearity.

To study the correlation between the molecular descriptors in the models 5, 6, we used SPSS program to obtain the variance inflation factor (VIF), Pearson coefficient correlation (PCC) and collinearity statistics in ANOVA table.

If the VIF value lies between 1 and 10, there is no multicollinearity; if VIF<1 or >10, there is multicollinearity and a recheck is necessary [29-31]. The VIF is calculated as follows:

$$\mathrm{VIF} = \frac{1}{1 - \mathrm{R}^2} \qquad (7)$$

From Table **5**, we can infer that, the multicollinearity has existed, because the Pearson correlation between IC0and GGI5descriptors are bigger than0.5, therefore there is a linearity between these descriptors. After removing GGI5 and then in the next step IC0 and RDF080u descriptors we corrected Equation (5) as follows:

$$\Delta H^{\circ}_{vap} = 41.944 + 678.635\,(R7v) + 1.784\,(SP20) - 10.041\,(Mor08p) + 9.327(Mor12p) \qquad (8)$$

N = 47, R =0.961, R^2 = 0.923, R^2_{adj} = 0.914, F = 98.618, DW = 1.688, Sig = 0.000, RMSE = 2.691

The suitable linear model for QSPR study of the thermal energy (Equation 6) includes eight molecular descriptors. The results of the correlation between these descriptors are listed in Table **6**. Based on these results, there are high correlations between EEig13x and EEig13r descriptors that indicate possible collinearity problems. After removing EEig13r descriptor, and the next steps ATS2e$_u$, nCb, EEig13r and E2m from this model, we corrected Equation **(6)** as follows:

$$T^{\circ}_{bp} = 324.787 + 1.778\,(ZM1V) \qquad (9)$$

N = 47, R = 0.966, R^2 = 0.932, R^2_{adj} = 0.929, F = 573.406, DW = 1.839, Sig = 0.000, RMSE = 3.086

Table 5. Correlation between the molecular descriptors (Eq.(5))

	Pearson Correlation for ΔH°_{vap}							Collinearity Statistical			Corrected model
Descriptor	**Mor12p**	**R7v**	**RDF080u**	**SP20**	**IC0**	**GGI5**	**Mor08p**	**VIF (1)**	**VIF (2)**	**VIF (3)**	**VIF (4)**
Mor12p	1.000							5.546	4.408	3.553	1.177
R7v	0.367	1.000						2.920	2.653	2.310	1.459
RDF080u	-0.254	-0.265	1.000					3.935	3.337	-----	-----
SP20	0.377	0.273	-0.156	1.000				3.159	3.015	1.742	1.554
IC0	0.411	0.104	0.079	-0.062	1.000			5.487	4.290	3.076	-----
GGI5	0.046	0.064	-0.335	0.197	-0.683	1.000		7.680	-----	-----	-----
Mor08p	-0.453	-0.408	-0.120	-0.405	-0.060	-0.218	1.000	9.066	8.048	3.553	1.667

Table 6. Correlation between the molecular descriptors (Eq.(6))

	Pearson Correlation for T°_{bp}								Collinearity Statistical			Corrected model
Descriptor	**Jhetp**	**EEig15x**	**E2m**	**EEig13r**	**nCb**	**ZM1V**	**EEig13x**	**ATS2e$_u$**	**VIF**	**VIF**	**VIF**	**VIF**
Jhetp	1.000								5.134	3.951	2.720	---
EEig15x	-0.075	1.000							44.555	5.967	---	---
E2m	-0.321	-0.089	1.000						111.174	3.447	2.802	---
EEig13r	0.464	-0.266	-0.097	1.000					76.795	---	---	---
nCb	-0.510	-0.321	0.182	-0.235	1.000				10.328	23.264	---	---
ZM1V	0.366	-0.379	-0.198	0.018	-0.042	1.000			73.891	42.102	5.080	1.000
EEig13x	-0.479	0.065	0.171	-0.949	0.241	-0.091	1.000		3.718	2.123	---	---
ATS2e$_u$	0.466	0.325	-0.288	0.596	-0.670	-0.448	-0.59	1.000	24.711	46.971	---	---

Validation

The success of any QSAR/ QSPR models depends on the accuracy of the input data, selection of appropriate descriptors, statistical tools and validation of the developed model. In this section, for verification, the validity of the regression models and the predictive ability and statistical significance of the QSPR models, squared cross-validation coefficient for leave-one-out (Q^2_{LOO}) and external validation through test set were used [32,34]. The Q^2_{LOO} value (Eq. 10) computed from 20 % of randomly chosen data was found to be positive and smaller than one.

$$Q^2 = 1 - \frac{\sum(Y_i - \hat{Y}_{i|i})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{PRESS}{TSS} \quad Q^2 \leq 1 \quad (10)$$

In Equation (10), the notation i|i indicates that the quantity is predicted by a model estimated when the i-th sample was left out from the training set.

The Q^2_{LOO} values of the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) and the normal temperature of boiling points (T°_{bp} K) of the mono and PAHs were calculated 0.929 and 0.959 respectively. Another method for judgment of reliability of predictions of models has been checked by 10 compounds as external validation test set.

The external prediction accuracy of the mentioned models was examined using R^2, R^2_{adj}, RMSE, DW, F and Sig values. These statistical results for training and test sets of studied properties are listed in Table **7**.

Figs (1, 2) show the linear correlation between the observed values versus predicted values of ΔH°_{vap} and T°_{bp} were obtained using Equations (8,9).

Regular Residuals

The residual is the difference between the observed (experimental) value of the dependent variable (y) and the predicted (calculated) value (ŷ). The residual of the GA-MLR calculated values of ΔH°_{vap} and T°_{bp} show a relatively random pattern (see Figs. 3, 4). This relatively random pattern shows that a linear model provides a decent fit to the data.

Table 7. Statistical parameters of models for training and test sets based on Equations 8,9

Data set	property	N	R	R^2	R^2_{adj}	RMSE	DW	F	sig
training	ΔH°_{vap}	47	0.961	0.923	0.914	2.691	1.688	98.618	0.000
test	ΔH°_{vap}	10	0.990	0.980	0.974	2.190	1.802	168.207	0.000
training	T°_{bp}	47	0.966	0.932	0.929	3.086	1.839	573.406	0.000
test	T°_{bp}	10	0.973	0.947	0.912	2.885	1.791	686.321	0.010

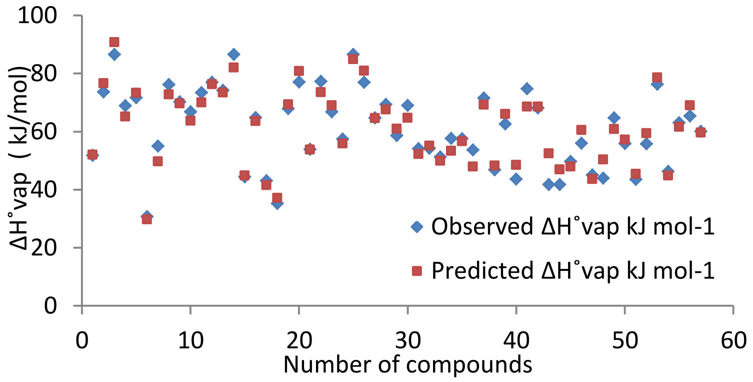


Fig. 1. Comparison between predicted and observed values of the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) of the mono and PAHs by the GA-MLR method.

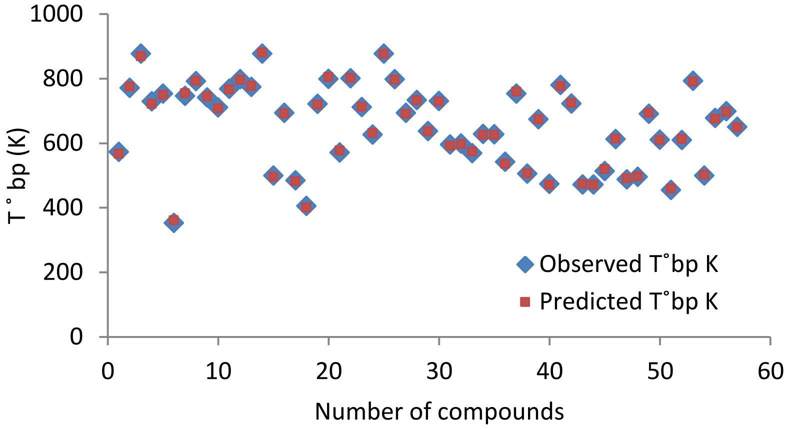


Fig. 2. Comparison between predicted and observed values of the normal temperature of boiling points (T°_{bp} K) of the mono and PAHs by the GA-MLR method.

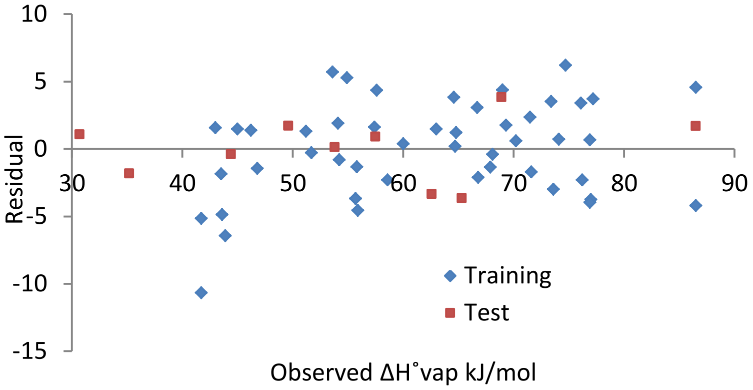


Fig. 3. Plot of residuals against the observed values of the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) of the mono and PAHs for training and test sets.

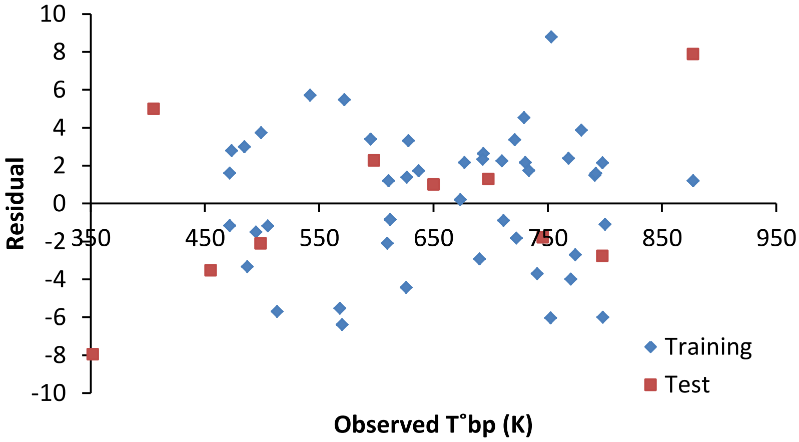


Fig. 4. Plot of residuals against the observed values of the normal temperature of boiling points (T°_{bp} K) of the mono and PAHs for training and test sets.

Interpretation of the best descriptors

The obtained results and discussion lead us to conclude that the four descriptors(Mor08p, Mor12p, SP20 and R7v) have been classified into topological index, 3D-MoRSE and GETAWAY descriptors can be used successfully for modeling and predicting the enthalpy of vaporization at standard condition of the studied compounds (see Table 8).

Topological indices are designed by transforming a molecular graph into a number. Topological indices possess the remarkable ability of being able to correlate and predict a very wide spectrum of properties for a vast range of molecular species [35].

GETAWAY descriptor calculated from the leverage matrix obtained by the centered atomic coordinates [36].This descriptor could be used for satisfactory prediction of the thermal energy.

3D-MoRSE descriptor incorporates the information about the whole molecule structure, and it is a very flexible 3D structure encoding framework for chemoinformatics and QSAR/QSPR purposes [37, 38].

As can be seen, only the one descriptor (ZM1V) in topological block is useful to predict the normal temperature of boiling points (T°_{bp} K) of the mono and PAHs than the other descriptors (Table 8).

The first Zagreb index (ZM1) is the sum of the square vertex degrees of all the non-hydrogen atoms. First Zagreb index by valence vertex degrees (ZM1V) is obtained in the same way as the ZM1 index by substituting the simple vertex degree by the valence vertex degree [38].

CONCLUSIONS

QSPR studies are mathematical correlations between molecular property and molecular descriptors. In this investigation, QSPR models have been developed to predict the normal temperature of boiling points (T°_{bp} K) and the enthalpy of vaporization at standard condition (ΔH°_{vap} kJ mol^{-1}) of 54 mono and PAHs. Multiple linear models were connected for modeling and predicting properties which are used in present study. Molecular descriptors calculated by the DRAGON software. The suitable descriptors were selected with the aid of the genetic algorithm (GA) technique and multiple linear regression (MLR) method. To assess the vigor and prescient capacity of the built models, leave-one-out cross-validation, internal and external validation methods were implemented.

Our results suggest that combining the four descriptors (Mor08p, Mor12p, SP20 and R7v) can be used for satisfactory prediction of ΔH°_{vap} of mono and PAHs. These descriptors are classified as Topological, 3D-MoRSE descriptors, and GETAWAY descriptors.

Table 8. Molecular descriptors used for ΔH°_{vap} and T°_{bp}

Property	Symbol	Description	Block
T°_{bp}	ZM1V	first Zagreb index by valence vertex degrees	Topological indices
ΔH°_{vap}	Mor08p	signal 8 / weighted by polarizability	3D-MoRSE descriptors
	Mor12p	signal 12 / weighted by polarizability	3D-MoRSE descriptors
	SP20	shape profile no. 20	Randic molecular profiles
	R7v	R autocorrelation of lag 7 / weighted by van der Waals volume	GETAWAY descriptors

The results and discussion lead us to conclude that the models established by GA-MLR method have good correlation of thermodynamic properties, which means QSPR models could be efficiently used for predicting of the above mentioned properties of the mono and PAHs.

The QSPR model involving one descriptor (ZM1V) provides a useful tool in predicting the normal temperature of boiling points (T°bp K) of mono and PAHs. This descriptor is classified as topological index.

REFERENCES

[1] P. Vollhardt and S. Neil, Organic Chemistry: Structure and Function; 5th. New York: W.H. Freeman and Company, 2007.

[2] M. Pogorzelec and K. Piekarska, Sci Total Environ. 631 (2018) 1431.

[3] Z. Wang, C. Yang, Z. Yang, J. Sun, B. Hollebone, C. Brown and M. Landriault, J. Environ. Monit. 13 (2011) 3004.

[4] H. I. Abdel-Shafy and M. S. M. Mansour, M. S. M. Egypt. J. petrol. 25 (2016) 107.

[5] T. Bogdanović, J. Pleadin, S. Petričević, E. Listeš, D. Sokolić, K. Marković and V. J. Šimat, Food. Compos. Anal. 75 (2019) 49.

[6] F. Shafiei and F. Arjmand, J. Struct. Chem. 59 (2018) 748.

[7] F. Shafiei and H. Hosseini, MATCH Commun. Math. Comput. Chem. 75 (2016) 583.

[8] N. Ahmadinejad, F. Shafiei and T. Momeni Isfahani, Comb. Chem. High Throughput Screen. 21 (2018) 1.

[9] M. Ghamali, S. Chtita, A. Ousaa, B. Elidrissi, M. Bouachrine, and T. Lakhlifi, J. Taibah. Univ. Sci .11 (2017) 1.

[10] B. Tuzun, C.Z. Yavuz, N. Sabanci and E. Saripinar, Curr. Comput –Aided. Drug. Des. 14 (2018) 370.

[11] S. D. Bolboaca, L. Jantschi and M.V. Diudea, Curr. Comput- Aided. Drug. Des. 9 (2013) 9 (2) 205.

[12] R. P. Schwarzenbach, P. M. Gschwend and D. M. Imboden, Environmental organic chemistry; John Wiley and Sons, Inc, Hoboken, 1993.

[13] F. Esen, Y. Tasdemir, and S. S. Cindoruk, Atmos. Res. 95 (2010) 379.

[14] F. Esen, S.S. Cindoruk and Y. Tasdemir, Environ. Pollut. 152 (2008) 461.

[15] O. G. Apul, P. Xuan, F. Luoand T. Karanfil, RSC. Advances. 3 (2013) 23924.

[16] B. Ma, H. Chen, M. Xu, T. Hayat and Y. He, J. Environ. Pollut. 158 (2010) 2773.

[17] P. Gramatica, P. Piluttiand E. Papa, Environ. Res. 18 (2007) 169.

[18] B. D. Gute, G. D. Grunwald and S. C. Basak, Environ. Res. 10 (1999) 1.

[19] S. Avram, M. Mernea, D. Mihailescu, D. Duda-Seiman and C. Duda-Seiman, Curr. Org. Chem. 17 (2013) 2880.

[20] D. Martin, S. Sild, U. Maran and M. Karelson, J. Phys. Chem. C. 112 (2008) 4785.

[21] G. N. Lu, Z. Dang, X.Q. Tao, P. A. Peng and D.C. Zhang, J. Theor. Comput. Chem. 4 (2005) 811.

[22] N. Bouarra , S. kherouf, A. Bouakkadia and D. Messadi , Res. J. Pharm. Biol. Chem. Sci. 8 (2017) 1.

[23] O. Geiss, C. Senaldi, I. Bianchi, A. Lucena, S. Tirendi and J. Barrero-Moreno, J. Chromatogr. A. 1(2018) 1566.

[24] H. Moriwaki, Y.S. Tian, N. Kawashita and T.J. Takagi, Cheminform. 10 (2018) 1.

[25] S. Ahmadi and E. Habibpour, Med. Chem. 17 (2017) 552.

[26] S. Chatterje and A.S. Hadi, Regression Analysis by Example; 4th Edition, John Wiley & Sonc, Inc., Hoboken. 2006.

[27] S. Chatterjee and J.S. Simonoff, Handbook of regression analysis; John Wiley & Sons, Inc, Hoboken, 2013.

[28] A. O. Aptula, N.G. Jeliazkova, T.W. Schultz and M. T. D. Cronin, QSAR. Comb. Sci. 24 (2005) 385.

[29] C. B. García, J. García, M. M. López Martín and R. Salmerón, J. Appl. Stat. 42 (2015) 648.

[30] D. R. Jensen, D. E. Ramirez, Adv. Decis. Sci, 2013 (2012) 1.

[31] D. J. Dupuis and M. P. Victoria-Feser, Ann. Appl. Stat, 7 (2013) 319.

[32] M. Zhao and D. Wei, Exploring the ligand-protein networks in traditional chinese medicine: current databases, methods and applications. In Advance in Structural Bioinformatics; Springer, Dordrecht, 2015.

[33] D. G. Kleinbaum, Applied regression analysis and other multivariable methods; Australia; Belmont, CA: Brooks/Cole, 2008.

[34] V. Consonni, D. Ballabio and R. Todeschini, J. Chemometr. 24 (2010) 194.

[35] Basak, S. C. Information theoretic indices of neighborhood complexity and their applications. In J. Devillers & A. T. Balaban (Eds.), Topological indices and related descriptors in QSAR and QSPR. Amsterdam: Gordon and Breach Science Publishers.1999, pp. 563–593.

[36] V. Consonni, R. Todeschini, M. Pavan and P. Gramatica, J. Chem. Inform. Comput. Sci. 42 (2002) 693.

[37] R. Todeschini and V. Consonni, Handbook of molecular descriptors; Wiley-VCH: Weinheim, 2000.

[38] O. Devinyak, D. Havrylyuk and R. Lesyk, 2014. J. Mol. Graph. Model. 54 (2014) 194.

مجله شیمی فیزیک و شیمی نظری
دانشگاه آزاد اسلامی واحد علوم و تحقیقات
جلد ۱۶، شماره ۳ و۴، پاییز و زمستان ۱۳۹۸
ISSN ۱۷۳۵-۲۱۲۶

مدل‌های ارتباط کمی ساختار- خاصیت (QSPR) برای پیش‌بینی برخی خواص ترمودینامیکی هیدروکربن‌های آروماتیکی مونو و چند حلقه‌ای با استفاده ازروش الگوریتم ژنتیک- رگرسیون خطی چندگانه (GA-MLR)

فاطمه دیالمه پور و فاطمه شفیعی*

دانشگاه آزاد اسلامی، واحداراک، گروه شیمی، اراک، ایران

چکیده

مدل های ارتباط کمی ساختار- خاصیت (QSPR) برای مدل‌سازی و پیش بینی خواص ترمودینامیکی مانند آنتالپی تبخیر در شرایط استاندارد (ΔH˚vap kJ mol-1) و دمای نقطه جوش نرمال (T˚bp K) ۵۷ هیدروکربن آروماتیکی مونو و چند حلقه‌ای (PAHs) مورد استفاده قرار گرفته است. هیدروکربن‌های آروماتیکی مونو و چند حلقه‌ای به طور تصادفی به دو گروه آموزشی و آزمایشی تقسیم شدند. مجموعه‌ای از توصیف کننده‌های مولکولی با استفاده از نرم‌افزار دراگون برای ترکیبات انتخاب شده محاسبه شده است. از روش الگوریتم ژنتیک (GA) و رگرسیون برگشتی برای انتخاب توصیف کننده‌های مناسب استفاده گردید. برای بدست آوردن رابطه خطی بین توصیف کننده‌ها و خواص شیمیایی از روش رگرسیون خطی چندگانه (MLR) استفاده شده است. توانایی پیش‌بینی مدل‌های GA-MLR با استفاده از روش‌های اعتبار سنجی تقاطعی و اعتبار سنجی خارجی انجام شده است. نتایج و بحث ما را به این نتیجه‌گیری می‌رساند که مدل‌های مجموعه آموزشی ایجاد شده با روش GA-MLR همبستگی خوبی با خصوصیات ترمودینامیکی دارند، به این معنی که می‌توان از مدل‌های QSPR بدست آمده برای تخمین و پیش بینی خواص ذکر شده هیدروکربن‌های آروماتیکی مونو و چند حلقه‌ای (PAHs) استفاده نمود.

کلید واژه‌ها: سامانه‌های رهایش دارو، سینتیک ایزوترمال، هیدروژل، رهایش دارو

[*] مسئول مکاتبات: shafa38@yahoo.com f-shafiei@iau-arak.ac.ir