Current Computer-Aided Drug Design, 2020, *16*, 6-16

RESEARCH ARTICLE

QSPR Models to Predict Thermodynamic Properties of Cycloalkanes Using Molecular Descriptors and GA-MLR Method

Daryoush Joudaki[1] and Fatemeh Shafiei[1,*]

[1]*Department of Chemistry, Arak Branch, Islamic Azad University, Arak, Iran*

ARTICLE HISTORY

Received: December 04, 2018
Revised: January 28, 2019
Accepted: February18, 2019

DOI:
10.2174/1573409915666190227230744

Abstract: ***Aims and Objectives*:** QSPR models establish relationships between different types of structural information to their observed properties. In the present study the relationship between the molecular descriptors and quantum properties of cycloalkanes is represented.

***Materials and Methods*:** Genetic Algorithm (GA) and Multiple Linear Regressions (MLR) were successfully developed to predict quantum properties of cycloalkanes. A large number of molecular descriptors were calculated with Dragon software and a subset of calculated descriptors was selected with a genetic algorithm as a feature selection technique. The quantum properties consist of the heat capacity (Cv)/ $Jmol^{-1}K^{-1}$ entropy(S)/ $Jmol^{-1}K^{-1}$ and thermal energy(E_{th})/ $kJmol^{-1}$ were obtained from quantum-chemistry technique at the Hartree-Fock (HF) level using the ab initio 6-31G* basis sets.

***Results*:** The Genetic Algorithm (GA) method was used to select important molecular descriptors and then they were used as inputs for SPSS software package. The predictive powers of the MLR models were discussed using Leave-One-Out (LOO) cross-validation, leave-group (5-fold)-out (LGO) and external prediction series. The statistical parameters of the training and test sets for GA–MLR models were calculated.

***Conclusion*:** The resulting quantitative GA-MLR models of Cv, S, and E^{th} were obtained:[r^2=0.950, Q^2=0.989, r^2_{ext}=0.969, $MAE_{(overall,5\text{-}flod)}$=0.6825 $Jmol^{-1}K^{-1}$], [r^2=0.980, Q^2=0.947, r^2_{ext}=0.943, $MAE_{(overall,5\text{-}flod)}$=0.5891$Jmol^{-1}K^{-1}$], and [$r^2$=0.980, Q^2=0.809, r^2_{ext}=0.985, $MAE_{(overall,5\text{-}flod)}$=2.0284 $kJmol^{-1}$]. The results showed that the predictive ability of the models was satisfactory, and the constitutional, topological indices and ring descriptor could be used to predict the mentioned properties of 103 cycloalkanes.

Keywords: Multiple linear regression, molecular descriptors, genetic algorithm, validation, cycloalkanes, GA-MLR.

1. INTRODUCTION

Cycloalkanes are types of alkanes that have one or more rings of carbon atoms in their structure. The physical properties of cycloalkanes are similar to those of alkanes, but they have higher boiling points, melting points and higher densities. Cycloalkanes are non-polar, and they interact only by weak London forces. They can also be used for many different purposes such as motor fuel, natural gas, petroleum gas, kerosene, diesel, and many other heavy oils. Generally, the melting point, the boiling point and the density of cycloalkanes increase as the number of carbons increases. This trend occurs because of the greater number of bonds that are in higher membered ring, thus making the bonds harder to break. They have higher London Dispersion forces because the ring shape allows for a greater area of contact. Ring strain also causes certain cycloalkanes to be more reactive [1, 2].

An important task is to predict biological activities, as well as toxicological or physicochemical properties of chemical compounds from chemical structure data. This domain is usually named quantitative structure activity /property relationship (QSAR/QSPR) and is highly related for example in drug design [3-5].

QSAR and QSPR studies are unquestionably of great importance in Physical, Analytical, Biochemistry, Organic and Inorganic chemistry. The aim of these studies is to convert searches for compounds with the required properties based on chemical intuition and experience into a mathematically quantified and computerized form [6, 7]. QSPR models are obtained through analyzing and calculating the correlation between the property and a variety of structure information of compounds.

The graph theoretical and topological aspects have been used to predict activity coefficients of molecular interaction in binary liquid mixtures of non-electrolytes [8].

Some thermodynamic properties such as molar excess volumes and molar excess enthalpies of butyl acetate with cyclohexane, benzene and toluene in terms of graph theoretical approach have been analyzed [9,10].

Molecular descriptors, tightly connected to the concept of molecular structure, play a fundamental role in scientific research, being the theoretical core of a complex network of knowledge. Indeed, molecular descriptors are based on sev-

*Address correspondence to this author at the Department of Chemistry, Arak Branch, Islamic Azad University, P.O. Box 38135-567, Arak, Iran, Tel /Fax: +9834130039; E-mail: f-shafiei@iau-arak.ac.ir

1875-6697/20 $65.00+.00

eral different theories, such as graph theory, quantum-chemistry, algebraic topology, information theory, organic chemistry, and so on, and are used to model several different properties of chemicals in scientific fields such as toxicology, analytical chemistry, physical chemistry, and medicinal, pharmaceutical, and environmental chemistry [11, 12].

For the use of the molecular descriptors, knowledge of chemoinformatics, chemometrics, statistics and the principles of the QSPR approaches is necessary [13].

The new Neuraminidase inhibitors with the cyclohexene scaffold were investigated using molecular dynamics techniques, and molecular docking. Molecular docking was used to confirm the built 3D- QSAR models of enzyme-inhibitor system [14].

For modeling and predicting the octane number of alkanes and cycloalkanes, Topological Equivalents (TEs) have been used [15].

The boiling points of 343 hydrocarbons (160 paraffins and 183 cycloalkanes) were correlated with three new topological indices, VDI (vertex degree-distance index), OEI (odd-even index), and RDI (ring degree-distance index) based on Vertex, distance, and ring using multiple regression models that were constructed [16].

Excess Gibbs free energies, enthalpies, volumes and comprehensibility of binary mixtures of cycloalkanes have been investigated [17, 18].

QSAR models were applied by using quantum chemical descriptors to predict the toxicity -logEC50 and -logLC50 of 28 alkyl cycloalkane-carboxylates [19, 20].

QSPR models to predict Boiling Point (Bp) of 106 cycloalkanes based on total and local quadratic indices have been researched. The quality of the models was determined by examining the statistical parameters of external prediction series and cross-validation procedures (leave-one-out and leave-group (5-fold)-out) [21].

The relationship between the Monte Carlo method, as the molecular descriptor and vapor pressure at 298.15 K of 84 hydrocarbons (63 alkanes and 21 cycloalkanes) using the van der Waals (vdW) surface area has been studied [22].

However, there has only been limited investigation of the quantitative structure-property relationship of cycloalkanes.

In the present study, QSPR mathematical models have been developed to predict the thermal energy, kJ mol^{-1}, heat capacity, $Jmol^{-1}K^{-1}$and entropy, $Jmol^{-1}K^{-1}$ of 103 cycloalkanes using GA-MLR method based on molecular descriptors calculated from the molecular structure alone.

2. MATERIALS AND METHODS

The thermal energy, heat capacity and entropy of 103 cycloalkanes were taken from the quantum mechanics methodology with the Hartree-Fock (HF) level, using the ab initio 6-31G* basis sets. Various cycloalkanes under study are listed in Table **1**. These data were randomly divided into a training set and an extermal test set consisting of 83, 20 data point, respectively. In order to calculate the theoretical descriptors, the molecular structures were constructed with the aid of Gauss View 5 and Gaussian 98 programs and then the molecular geometries of compounds were better optimized by dragon package 2.1. A total of 1502 theoretical descriptor were calculated for each compound in the data set using Dragon software. The most relevant descriptors are needed to be selected from the remained descriptors. This is the prominent problem in QSPR studies to choose the minimum number of descriptors with high prediction ability of the model. Conventional variable selection methods like stepwise regression are based upon a single solution or a few solutions. To overcome this problem, a Genetic Algorithm (GA) designed for the selection of variables was used. GA is a stochastic method used to solve optimization problems defined by fitness criteria, applying the evolution hypothesis of Darwin and different genetic functions, *i.e*. crossover and mutation. In this work, the number of molecular descriptors was reduced by genetic algorithm analysis and the backward stepwise regression method.

We used Multiple Linear Regression (MLR) technique for a linear relationship between descriptors and quantum properties (heat capacity, entropy and thermal energy) of 103 cycloalkanes. The Genetic Algorithms (GA)-MLR regression are written in MATLAB (version 2010a) environment. In the following, MLR models were performed by the statistical package for social (SPSS) software (version 20).

3. RESULTS AND DISCUSSION

3.1. Statistical Analysis

Structural-property models were generated using the MLR procedure of SPSS version 20. The entropy, thermal energy, and heat capacity as the dependent variable and dragon molecular descriptors as the independent variable were used. The models were assessed with a correlation coefficient (r), coefficient of determination (r^2), adjusted correlation coefficient (r^2_{adj}), Fisher ratio (F), Root Mean Square Error (RMSE), Durbin-Watson statistic (D) and Significance (Sig).

Several linear QSPR models have been created that contain 3-7 descriptors. The suitable descriptors for predicting the above mentioned properties have been selected by using genetic algorithm and Dragon software.

The selection of significant descriptors, which relate the property data to the molecular structure, is an important step in QSPR modelling. Selection of the significant structural descriptors among the 1502 ones was performed as follows: all descriptors with same values for all molecules were omitted, and one of the two descriptors having a pairwise correlation coefficient above 0.9 ($R > 0.9$) was removed. Finally, 233 molecular descriptors remained.

The best molecular descriptors were chosen using the SPSS software which is based on the multivariate backward stepwise.

3.2. QSPR Models for the Entropy

Table **2**, shows the regression coefficient and statistical parameters of models for the entropy of 83 cycloalkanes. It can be seen from Table **2**, that five descriptors are used in the MLR model. These descriptors are: REIG, VE2_A, Mor15u, Vu, and H4u. The regression parameters of the best model of five dragon molecular descriptors are collected in Equation (1).

Table 1. The name of compounds of cycloalkanes used in this study.

S. No.	Compound	S. No.	Compound	S. No.	Compound
1	1,1,2,2- tetramethylcyclopentane	**36**	1,3-diethylcyclohexane	**71**	1-methyl-3- ethylcyclohexane
2	1,1,2,2-tetramethylcycopropane	**37**	1,3-diethylcyclopentane	**72**	1-methyl-3- propylcyclohexane
3	1,1,2,3-tetramethylcyclopentane	**38**	1,3-dimethyl-2- ethylcyclopentane	**73**	1-methyl-3- propylcyclohexane
4	1,1,2,3-tetramethylcyclopropane	**39**	1,3-dimethyl-4- ethylcyclopentane	**74**	1-methyl-4- ethylcyclohexane
5	1,1,2,4-tetramethylcyclopentane	**40**	1,3-dimethylcyclohexane	**75**	1-methyl-4- propylcyclohexane
6	1,1,2-trimethyl-4-ethylcyclopentane	**41**	1,3-dimethylcyclopentane	**76**	1-propylcyclohexane
7	1,1,2-trimethylcyclobutane	**42**	1,4-diethylcyclohexane	**77**	1-propylcyclopentane
8	1,1,2-trimethylcyclohexane	**43**	1,4-dimethylcyclohexane	**78**	2-cyclopropylbutane
9	1,1,2-trimethylcyclopentane	**44**	1-cyclopentyl-1-methylbutane	**79**	2-methyl-1- propylcyclopropane
10	1,1,3,4-tetramethylcyclopentane	**45**	1-cyclopentyl-2-methylbutane	**80**	buthylcyclopentane
11	1,1,3-trimethylcyclohexane	**46**	1-cyclopentyl-3-methylbutane	**81**	cyclobutane
12	1,1-diethylcyclohexane	**47**	1-cyclopropyl-2-methylbutane	**82**	cyclodecane
13	1,1- diethylcyclopentane	**48**	1-cyclopropylbutane	**83**	cycloheptane
14	1,1-diethylcyclopropane	**49**	1-ethyl-1,2- dimethylcyclopropane	**84**	cyclohexane
15	1,1-dimethyl-2-ethylcyclopentane	**50**	1-ethyl-1-methylcyclobutane	**85**	cyclononane
16	1,1-dimethylcyclohexane	**51**	1-ethyl-2,2- dimethylcyclopropane	**86**	cyclooctane
17	1,1-dimethylcyclopentane	**52**	1-ethyl-2,3- dimethylcyclopropane	**87**	cyclopentane
18	1,1-dimethylcyclopropane	**53**	1-ethyl-3-methylcyclobutane	**88**	cyclopropane
19	1,2,2,3-tetramethylcyclopentane	**54**	1-isobutyl-4-methylcyclohexane	**89**	cycloundecane
20	1,2,3,4-tetramethylcyclopentane	**55**	1-isopropyl-1- methylcyclohexane	**90**	ethylcyclohexane
21	1,2,3-trimethyl-4- ethylcyclopentane	**56**	1-isopropyl-1- methylcyclopentane	**91**	ethylcyclopentane
22	1,2,3-trimethylcyclobutane	**57**	1-isopropyl-1- methylcyclopropane	**92**	ethylcyclopropane
23	1,2,3-trimethylcyclohexane	**58**	1-isopropyl-2- methylcyclohexane	**93**	isobuthylcyclohexane
24	1,2,3-trimethylcyclopentane	**59**	1-isopropyl-2- methylcyclopentane	**94**	isobuthylcyclopentane
25	1,2,4-trimethyl-3- ethylcyclopentane	**60**	1-isopropyl-2- methylcyclopropane	**95**	isopropylcyclobutane
26	1,2,4-trimethylcyclohexane	**61**	1-isopropyl-3-methylcyclohexane	**96**	isopropylcyclohexane
27	1,2-diethylcyclohexane	**62**	1-isopropyl-3- methylcyclopentane	**97**	isopropylcyclopentane
28	1,2diethylcyclopentane	**63**	1-isopropyl-4- methylcyclohexane	**98**	methylcyclobutane
29	1,2-diethylcyclopropane	**64**	1-methyl-1-ethylcyclohexane	**99**	methylcyclohexane
30	1,2-dimethyl-1- ethylcyclopentane	**65**	1-methyl-1-propylcyclohexane	**100**	methylcyclopentane
31	1,2-dimethy-3-ethylcyclopentane	**66**	1-methyl-1-propylcyclopentane	**101**	methylcyclopropane
32	1,2-dimethylcyclohexane	**67**	1-methyl-1-propylcyclopropane	**102**	pentylcyclopentane
33	1,2-dimethylcyclopentane	**68**	1-methyl-2-ethylcyclohexane	**103**	propylcyclobutane
34	1,2-dimethylcyclopropane	**69**	1-methyl-2-ethylcyclopentane	**-**	-
35	1,3,5-trimethylcyclohexane	**70**	1-methyl-2-propylcyclopentane	**-**	-

Table 2. Statistical parameters of the models calculated with the SPSS software for entropy, J mol^{-1} K^{-1}.

Model	Independent Variable	r	r^2	r^2_{adj}	RMSE	F
1	**REIG** (first eigenvalue of the R matrix) **Vi**(V total size index / weighted by ionization potential) **Mor15u(**Signal 15 / unweighted)**, H4u**(H autocorrelation of lag 4/unweighted), **Vu** (V total size index/unweighted)**, VE2_A** (Average eigenvector coefficient sum from adjacency matrix)**, RARS**(R matrix average row sum)	0.995	0.991	0.990	0.4105	923.223
2	REIG, Mor15u, H4u, Vu, VE2_A, RARS	0.995	0.991	0.990	0.4108	1.075×10^3
3	REIG, Mor15u, H4u, Vu, VE2_A	0.995	0.991	0.990	0.4133	1.275×10^3

r= Correction coefficient.
r^2= Coefficient of determination.
r^2_{adj}= Adjusted correlation coefficient.
F= Fisher Ration.
RMSE= Root mean sequare error.

S/J mol^{-1} K^{-1} =620.649-226.340(VE2_A)+14.917(Mor15u)+ 2.487(Vu)-10.023(H4u)-215.786(REIG) (1)

n =83, r=0.995, r^2=0.991, r^2_{adj} =0.990, F=1.275×10^3, D= 1.947, Sig =0.000, RMSE=0.4133

3.3. QSPR Models for the Thermal Energy

Table **3** shows the regression coefficients and statistical factors of models for the thermal energy of 83 cycloalkanes. The best linear model for the thermal energy includes three molecular-descriptors SRW06, Mor01m, and MLOGP2. The regression parameters of the best model of the three molecular descriptors are collected in Equation (2).

E_{th} / kJ mol^{-1}=86.007-0.045(SRW06)-3.707(Mor01m) +50. 304(MLOGP2) (2)

n =83, r=1.000, r^2=1.000, r^2_{adj} =1.000, F=274×10^3, D = 1.786, Sig =0.000, RMSE=0.1409

3.4. QSPR Models for the Heat Capacity

Table **4** shows the regression coefficients and statistical parameters of models for the heat capacity of 83 cycloalkanes. It can be seen in Table **4** that six descriptors are used in the suitable MLR model. These descriptors are: R3p, MLOGP, RTp, ALOGP, ALOGP2, and R3e+.

The regression parameters of the best model of the six molecular descriptors are collected in Equation (3).

C_v/J mol $^{-1}$K^{-1} =-75.858+88.134(R3e+)-32.205(R3p)+ 1.445(RTP)+86.583(MLOGP)- 46.681(ALOGP)+3.582(AL OGP2) (3)

n =83, r =0.999, r^2 =0.999, r^2_{adj} =0.998, F= 8160.158, D= 1.339, Sig =0.000, RMSE= 0.1192

The results of the entropy, thermal energy and heat capacity are very satisfactory.

In this study to find the best model to predict the properties mentioned, we will use the following sections.

3.5. Multicollinearity

In statistics, multicollinearity (also collinearity) is a phenomenon in which one-predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. Also, multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable [23].

Multicollinearity can be detected with the help of tolerance and its reciprocal, called variance inflation factor (VIF). If the value of tolerance is less than 0.2 or 0.1 and, simultaneously, the value of VIF 10 and above, then the multicollinearity is problematic.

In all our final models, the multicollinearity has existed, because the values of correlations between independent variables are near to one and VIFs value are not between 1 and 10 (Tables **5-7**).

To study the correlation between the molecular descriptors in the models 1 - 3, we used SPSS program to obtain the Pearson coefficient correlation and collinearity statistics. The results of this study are recorded in Tables **5-7**.

For Equation (1), the Pearson correlation between REIG and VEA2 descriptors is close to unity, and VIF for VEA2 and Vu are bigger than 10 (Table **5**), therefore there is a linearity between these descriptors. After removing REIG from this model, we corrected Equation (1) as follows:

S/J mol^{-1} K^{-1} = 438.668 - 364.129 VEA2 + 14.262Mor15u + 3.389 Vu -8.312 H4u (4)

n =83, r =0.990, r^2 =0.980, r^2_{adj} =0.979, F = 652.102, D = 1.757, Sig =0.001, RMSE=0.5639

For the thermal energy (Eq. (2)), the Pearson correlation between MLOGP2 and Mor01m descriptors is close to unity, and VIF for these descriptors are bigger than 10 (Table **6**), therefore there is a linearity between them. After removing MLOGP2 from this model, we corrected Equation (2) as follows:

E_{th} / kJ mol^{-1} = 288.706 + 0.116 SRW06 + 8.170 Mor01m (5)

n =83, r =0.990, r^2 =0.980, r^2_{adj} =0.980, F =1974.355, D = 1.684, Sig=0.000, RMSE=2.0137

Table 3. Statistical parameters of the models calculated with the SPSS software for thermal energy, kJ mol^{-1}.

Model	Independent Variable	r	r^2	r^2_{adj}	RMSE	F
1	**MLOGP2**(squared Moriguchi octanol-water partition coeff. (logP^2)), **Vi** (V total size index / weighted by ionization potential), **SRW08** (Self-returning walk of order 08), **Mor02m**(signal 02 / weighted by mass), **Mor01m** (Signal 01/Weighted by mass), **SRW06** (Self-returning walk of order 06)	1.000	1.000	1.000	0.1414	136×10^3
2	MLOGP2, SRW08, Mor02m, Mor01m, SRW06	1.000	1.000	1.000	0.1407	165×10^3
3	MLOGP2, Mor02m, Mor01m, SRW06	1.000	1.000	1.000	01405	207×10^3
4	MLOGP2, Mor01m, SRW06	1.000	1.000	1.000	0.1409	274×10^3

r= Correction coefficient.
r^2= Coefficient of determination.
r^2_{adj}= Adjusted correlation coefficient.
F= Fisher Ration.
RMSE= Root mean sequare error.

Table 4. Statistical parameters of the models calculated with the SPSS software for heat capacity, J mol $^{-1}K^{-1}$.

Model	Independent Variable	r	r^2	r^2_{adj}	RMSE	F
1	F02[C-C](Frequency of C - C at topological distance 2), **R3p(** R autocorrelation of lag 3/weighted by atomic polarizabilites), **RTp (**R total index/Weighted by polanizability) **R3e+(**R maximal autocorrelation of lag 3/Weighted by Sanderson electronegativity), **ALOGP2(**squared Ghose-Crippen octanol-water partition coeff. (logP^2), **MLOGP(**Moriguchi octanol-water partition coeff. (logP), **ALOGP** (Ghose-Crippen octanol-water partition coeff. (logP)	0.999	0.999	0.998	0.1194	6980.252
2	ALOGP2, R3p, $R3e^+$, RTp, MLOGP, ALOGP	0.999	0.999	0.998	0.1192	8160.158

r= Correction coefficient.
r^2= Coefficient of determination.
r^2_{adj}= Adjusted correlation coefficient.
F= Fisher Ration.
RMSE= Root mean sequare error.

Table 5. Correlation between the molecular descriptors (Eq. (1)).

Pearson Correlation						Collinearity Statistical		Corrected Model
-	VEA2	Mor15u	Vu	H4u	REIG	Tolerance	VIF	VIF
VEA2	1	0.446	-0.883	-0.735	0.942	0.108	9.265	4.751
Mor15u	-	1	-0.508	-0.425	0.475	0.734	1.362	1.355
Vu	-	-	1	0.763	-0.957	0.078	12.862	5.518
H4u	-	-	-	1	0.789	0.374	2.671	2.504
REIG	-	-	-	-	1	0.039	25.735	-

Table 6. Correlation between the molecular descriptors (Eq. (2)).

Pearson Correlation			-	Collinearity Statistical		
-	SRW06	Mor01m	MLOGP2	Tolerance	VIF	VIF
SRW06	1.000	0.599	0.630	0.539	1.856	1.560
Mor01m	-	1.000	0.994	0.010	99.312	1.560
MLOGP2	-	-	1.000	0.009	105.491	-

Table 7. Correlation between the molecular descriptors (Eq. (3)).

Pearson Correlation							Collinearity Statistical		Corrected Model	
-	R3e+	R3p	RTp	MLOGP	ALOGP	ALOGP2	Tolerance	VIF	VIF	VIF
R3e+	1	-0.83	-0.89	-0.916	-0.786	-0.713	0.076	13.109	12.847	5.275
R3p	-	1	0.883	0.795	0.736	0.678	0.109	9.139	7.327	4.720
RTp	-	-	1	0.964	0.914	0.880	0.025	39.716	32.411	7.411
MLOGP	-	-	-	1	0.941	0.900	0.018	54.956	41.225	-
ALOGP	-	-	-	-	1	0.989	0.006	154.124	-	-
ALOGP2	-	-	-	-	-	1	0.009	110.155	9.369	-

For the heat capacity (Eq.(3)), the Pearson correlation between ALOGP2 and ALOGP descriptors are close to unity, and VIF for some descriptors such as ALOGP, ALOGP2 and MLOGP are bigger than 10 (Table **7**), therefore there is a linearity between them. After removing ALOGP from this model, we corrected Equation (3) as follows:

Cv/J mol -1K-1 = 107.001 -690.541 R3e+ -135.596 R3p + 25.536 RTp (6)

n=83, r =0.975, r^2 =0.950, r^2_{adj} =0.948, F = 652.102, D = 1.447, Sig =0.000, RMSE=0.6876

3.6. Validation

Validation is the process of evaluating software at the end of the development process to determine whether software meets the customer expectations and requirements. It is a statistical methodology used to ensure whether the model created is a good model or a poor model.

Validation is needed to assess the predictive ability and statistical significance of the models [24-27].

In purpose to create and test models, data set of compounds randomly separated into a training set of compounds (80%), that was applied to made model (an internal method) and a prediction set of compounds (20%), that was applied to test the made model (an external method). Statistical factors such as r, r^2, r^2_{adj}, F and RMSE, of these models for training and test sets of the heat capacity, Cv/J $mol^{-1}K^{-1}$ entropy, S/J $mol^{-1}K^{-1}$ and thermal energy, E_{th}/kJ $mol^{-1}$are listed in Table **8**.

3.6.1. Cross-Validation

Cross-validation (CV) is a common method for internally validating a QSPR model [28]. CV process repeats the regression many times on subsets of one molecule (leave one out, LOO) or more than one molecule (leave many out, LMO and leave group out, LGO).

Many authors consider high Q^2_{CV} values as indicator or even as the ultimate proof of the high predictive power of a QSAR/QSPR model [29]. In recent years, some authors demonstrated that a high value of Q^2_{LOO} appears to be necessary but not a sufficient condition for the model to have a high predictive power [21, 30, 31].

In the present paper for the predictive power of the model, squared cross-validation coefficient for leave -one- out (Q^2_{LOO}), leave-group (5-fold)-out (LGO), and external validation through test set were used. The data set of 103 cycloalkanes was randomly separated into a training set of 83 compounds, and test set of 20 compounds.

The Q^2_{LOO} values of the entropy, thermal energy, and heat capacity models (Eqs. (4-6)) calculated 0.947, 0.809, and 0.989 respectively.

Also, the predictive ability and stability of the developed models were assessed using the five-fold cross-validation technique. First, the training set was randomly splitted into five groups of approximately equal size (20%). Each time, one of the five subsets was used as the validation set, and the model was trained with the remaining four subsets (80% of the data). This procedure was repeated five times until each observation has been left out at least once [32-34]. Then, the average error across all five trials was computed. The predictive performance of the models was measured through the Mean Absolute Error (MAE). The MAE for Equations (4-6) had an overall MAE of 0.5891J/molK(0.7256, 0.5023, 0.5526, 0.5275 and 0.6375), 2.0284 kJ/mol (2.1201, 2.2565, 1.7523, 1.9512 and 2.0241), and 0.6825 J/molK(0.7254, 0.6323, 0.6757, 0.7768 and 0.6025) respectively.

Table 8. Statistical parameters obtained by the GA- MLR model for the entropy, thermal energy and heat capacity for training and test sets (Eqs. 4-6).

Data Set	Property	n	r	r^2	r^2_{adj}	RMSE	D	F	Sig
Training	S	83	0.990	0.980	0.979	0.5639	1.757	652.102	0.001
Test	S	20	0.971	0.943	0.928	0.8086	1.877	62.387	0.000
Training	E_{th}	83	0.990	0.980	0.980	2.0137	1.684	1974.355	0.000
Test	E_{th}	20	0.993	0.985	0.984	0.9568	1.737	573.343	0.000
Training	Cv	83	0.975	0.950	0.948	0.6876	1.447	652.102	0.000
Test	Cv	20	0.984	0.969	0.962	0.4112	1.776	133.471	0.000

r= Correction coefficient.
r^2= Coefficient of determination.
r^2_{adj}= Adjusted correlation coefficient.
F= Fisher Ration.
RMSE= Root mean sequare error.
D= Durbin-Watson Statistic
Sig= Significance
N= number

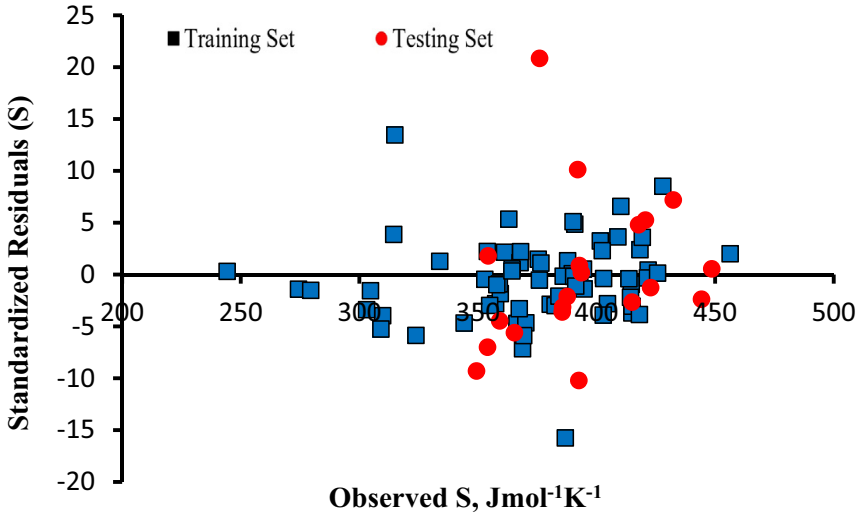

Fig. (1). Residuals (e_i) plotted against the observed entropy of cycloalkanes.

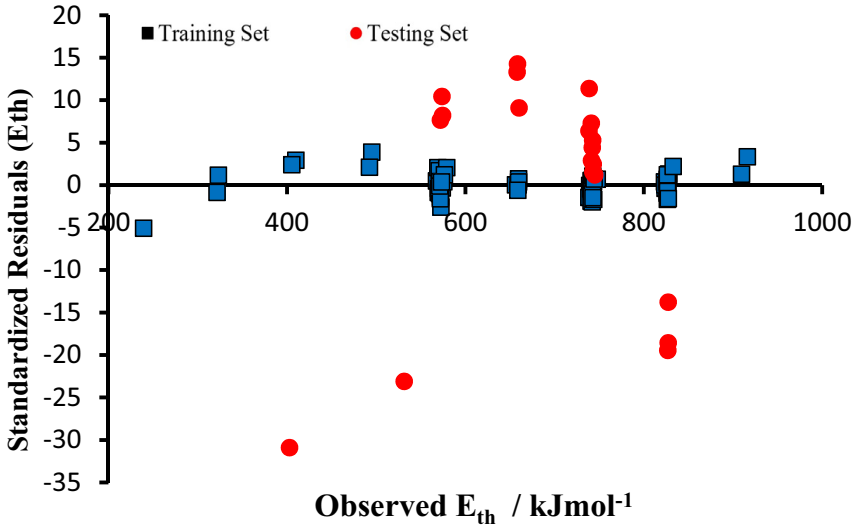

Fig. (2). Residuals (e_i) plotted against the observed thermal energy of cycloalkanes.

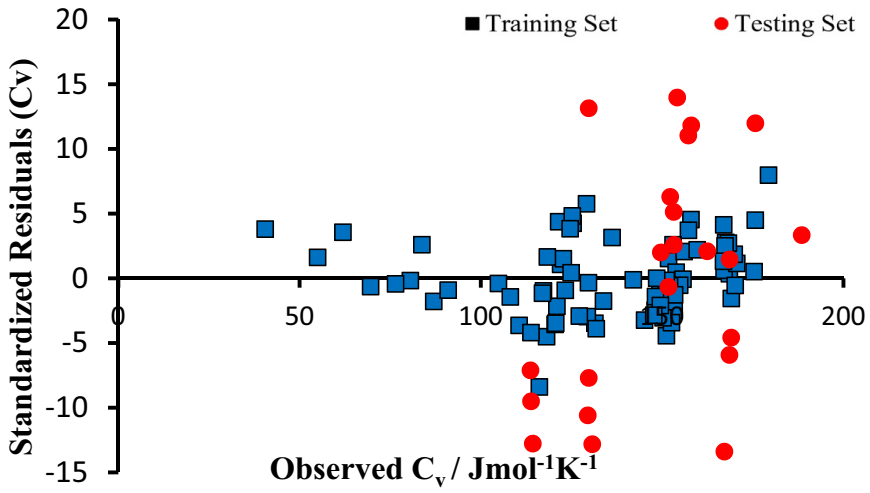

Fig. (3). Residuals (e_i) plotted against the observed heat capacity of cycloalkanes.

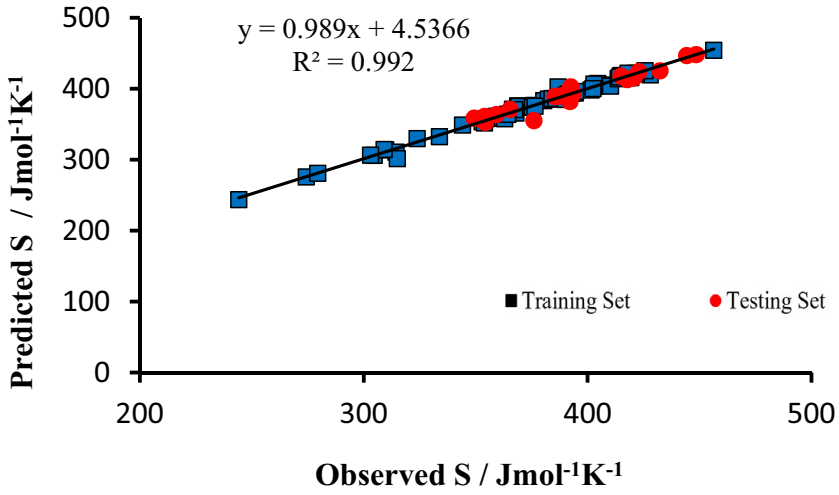

Fig. (4). A plot of observed (experimental) *versus* predicted (calculated) the entropy of cycloalkanes.

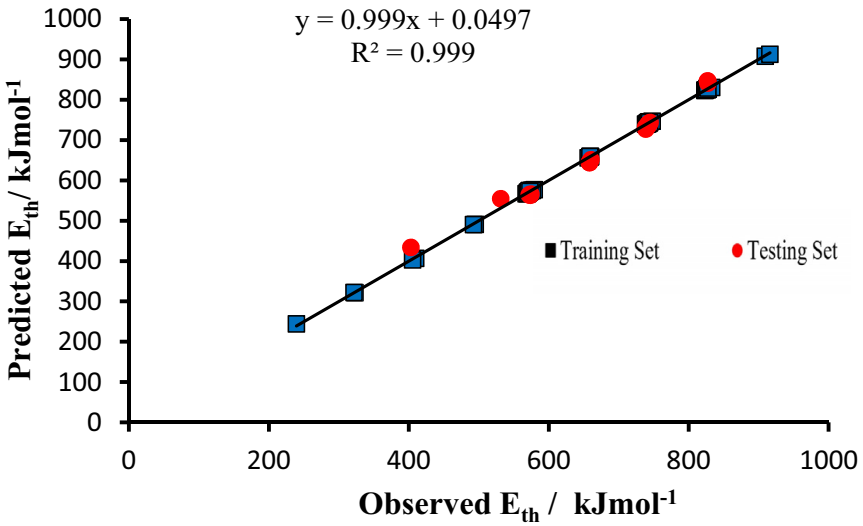

Fig. (5). A plot of observed *versus* predicted (calculated) the thermal energy of cycloalkanes.

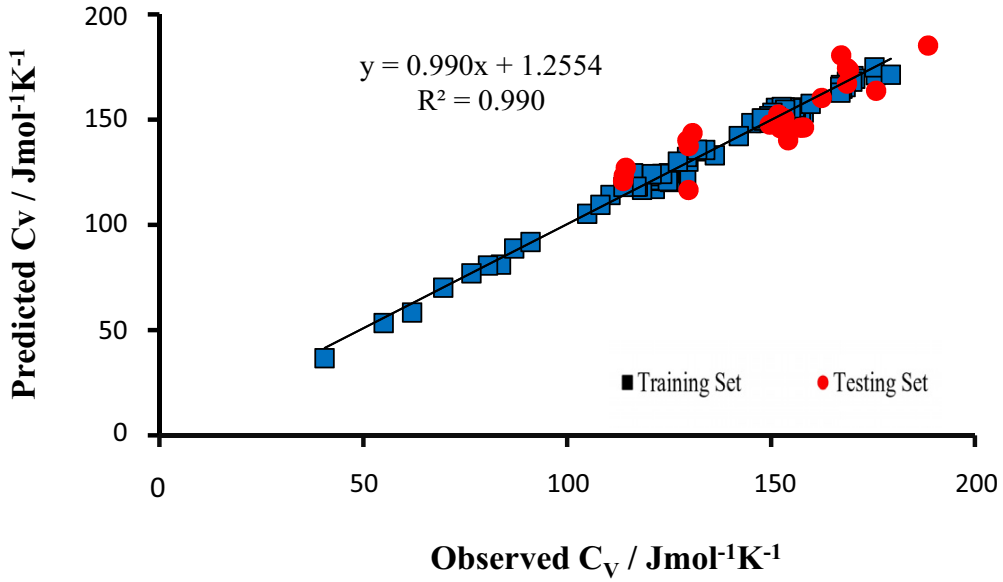

Fig. (6). A plot of observed versus predicted (calculated) the heat capacity of cycloalkanes.

3.7. Durbin-Watson Statistic

The Durbin-Watson (D) Statistic is a number which determines whether there is autocorrelation in the residuals of a time series regression. The value of D always lies between 0 and 4. If the Durbin-Watson statistic is substantially less than 2, there is an evidence of a positive serial correlation. If D>2, it indicates that the successive error terms are, on average, much different in value from one another, *i.e*., they are negatively correlated [28]. In our all models, the value of Durbin-Watson statistic is close to 2 (Eqs. (4-6)) and the bench error is uncorrelated.

3.8. Regular Residuals

The residual is the difference between the observed and predicted values. The residual values of the thermal energy, heat capacity and entropy expressed by Equations (4-6) are shown in Tables (**S1a-S1c**) of the Supplementary material to this paper. A residual plot is a graph that shows the residual values on the vertical axis and independent variables on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. The residual of the MLR calculated values of the entropy, thermal energy, and heat capacity were propagated in both sides of zero line that indicates no systematic error exists in the model development (Figs. **1-3**).

The calculated (predicted) S, Cv and Eth values from Equations (4-6) for the training and test set are listed in Tables (**S1a-S1c**). Figs. (**4-6**) show the linear correlation between the observed and the predicted properties as mentioned above.

CONCLUSION

The results of this study demonstrate that the QSPR method using the GA-MLR technique based on molecular descriptors calculated from molecular structure can generate suitable models for the prediction of entropy, thermal energy, and heat capacity of cycloalkanes and their derivatives. These QSPR models showed high values of multiple correlation coefficient (R>0.97) and Fisher - ratio statistics.

MLR models are proved to be a useful tool in the prediction of S, Eth, and C_v. Leave one out cross - validation, leave-group (5-fold)-out (LGO), and external validation through test set as the evaluation techniques have been designed to evaluate the quality and predictive ability of the MLR models. The validation results suggest that the models possess good predictive ability and robustness. The obtained results and discussion lead us to conclude that combining the two descriptors SRW06 and Mor01m can be used for modeling and predicting the thermal energy of 103 cycloalkanes. These descriptors are classified in walk and path counts, and 3D-MoRSE descriptors, respectively.

The three GETAWAY descriptors, R3p, RTp, and R3e+ can be used for satisfactory prediction of the heat capacity.

The entropy of cycloalkanes derivatives can be better modeled using a combination of the four descriptors, VE2_A, M0r15u, Vu, and H4u. These descriptors are classified in 2D matrix- based, 3D-MoRSE, WHIM and GETAWAY descriptors respectively.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

The authors would like to thank Islamic Azad University Arak for their support on this work.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Vollhardt, K.; Peter, C.; Neil, E. *schore organic chemistry*; Freeman, W. H. NewYork, 5th, ed., **2007**.

[2] McMurry, J.E.; Eric, E.; Simanek, K. *Fundamentals of organic chemistry*; Brooks Cole, 6th, ed., **2006**.

[3] Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breadch Sci-ence Pub.: Amsterdam, The Netherland, **1999**.

[4] Diudea, M.V. *QSAR/QSPR studies by Molecular Descriptors*; Nova Pub.: Huntington, New York, **2001**.

[5] Hessler, G.; Baringhaus, K.H. Artificial Intelligence in Drug Design. *Molecules,* **2018**, *23*(10), 2520-2533.
[http://dx.doi.org/10.3390/molecules23102520] [PMID: 30279331]

[6] Toubaei, A.; Golmohamadi, H.; Dashtbozorgi, Z. QSPR studies for predicting gas to acetone and gas to acetonitrile solvation enthalpies using support vector machine. *J. Mol. Liq.,* **2012**, *175*, 24-32.
[http://dx.doi.org/10.1016/j.molliq.2012.08.006]

[7] Raja, G.; Saravanan, K. Quantum chemical and corrosion inhibition studies of an organic compound: 2,5 dichloroan-iline. *Rasayan J. Chem.,* **2015**, *8*, 8-12.

[8] Singh, p.; Maken, S. Topological aspects of molecular inter-actions in liquid mixtures of non-electrolytes. *Pure Appl. Chem.,* **1994**, *66*(3), 449-454.
[http://dx.doi.org/10.1351/pac199466030449]

[9] Maken, S.; Deshwal, B.R.; Chadha, R.; Singh, K.C.; Kim, H.; Park, J.W. Topological and thermodynamic investigations of molecular interactions in binary mixtures: Molar excess vol-umes and molar excess enthalpies. *Fluid Phase Equilib.,* **2005**, *235*(1), 42-49.
[http://dx.doi.org/10.1016/j.fluid.2005.06.011]

[10] Rani, M.; Maken, S. Topological studies of molecular interac-tions of formamide with propanol and butanol at 298.15 K. *J. Ind. Eng. Chem.,* **2012**, *18*(5), 1694-1704.
[http://dx.doi.org/10.1016/j.jiec.2012.03.011]

[11] Todeschini, R.; Consonni, V. *Handbook of molecular de-scriptors*; Wiley-VCH: Weinheim, **2000**.
[http://dx.doi.org/10.1002/9783527613106]

[12] Prana, V.; Rotureau, P.; André, D.; Fayet, G.; Adamo, C. Development of Simple QSPR Models for the Prediction of the Heat of Decomposition of Organic Peroxides. *Mol. Inform.,* **2017**, *36*(10), 1-9.
[http://dx.doi.org/10.1002/minf.201700024] [PMID: 28402598]

[13] Shafiei, F.; Arjmand, F. Prediction of the normal boiling points and enthalpy of vaporizations of alcohols and phenols using topological. *J. Struct. Chem.,* **2018**, *59*, 748-754.
[http://dx.doi.org/10.1134/S0022476618030393]

[14] Wang, Z.; Cheng, L.P.; Zhang, X.H.; Pang, W.; Li, L.; Zhao, J.L. Design, synthesis and biological evaluation of novel oseltamivir derivatives as potent neuraminidase inhibitors. *Bioorg. Med. Chem. Lett.,* **2017**, *27*(24), 5429-5435.
[http://dx.doi.org/10.1016/j.bmcl.2017.11.003] [PMID: 29141777]

[15] Smolenskii, E.A.; Ryzhov, A.N.; Bavykin, V.M.; Myshenkova, T.N.; Lapidus, A.L. Octane numbers (ONs) of hydrocarbons: a QSPR study using optimal topological in-dices for the topological equivalents of the ONs. *Russ. Chem. Bull.,* **2007**, *56*, 1681-1687.
[http://dx.doi.org/10.1007/s11172-007-0262-2]

[16] Cao, C.; Yuan, H. Topological indices based on vertex, distance, and ring: on the boiling points of paraffins and cycloalkanes. *J. Chem. Inf. Comput. Sci.,* **2001**, *41*(4), 867-877.
[http://dx.doi.org/10.1021/ci000467t] [PMID: 11500103]

[17] Stokes, R.H.; Marsh, K.N.; Tomlins, R.P. Enthalpies of exothermic mixing Enthalpies of exothermic mixing measured by the isothermal displacement calorimeter for cyclo-octane + cyclopentane at 25 °C. *J. Chem. Thermodyn.,* **1969**, *1*, 377-379.
[http://dx.doi.org/10.1016/0021-9614(69)90067-6]

[18] Ewing, M.B.; Marsh, K.N. Thermodynamics of cycloalkane+cycloalkane mixtures: comparison with theory. *J. Chem. Thermodyn.,* **1977**, *9*, 863-871.
[http://dx.doi.org/10.1016/0021-9614(77)90172-0]

[19] Wang, Z.Y.; Zhai, Z.C.; Wang, L.S. Quantitative Structure-activity Relationship of Toxicity of Alkyl (1-phenylsulfonyl) Cycloalkane-carboxylates Using MLSER Model and Ab initio. *QSAR Comb. Sci.,* **2005**, *24*, 211-217.
[http://dx.doi.org/10.1002/qsar.200430873]

[20] Katritzky, A.R.; Slavov, S.H.; Stoyanova-Slavova, I.S.; Kahn, I.; Karelson, M. Quantitative structure-activity relationship (QSAR) modeling of EC50 of aquatic toxicities for Daphnia magna. *J. Toxicol. Environ. Health A,* **2009**, *72*(19), 1181-1190.
[http://dx.doi.org/10.1080/15287390903091863]
[PMID: 20077186]

[21] Ponce, Y.M. Total and Local Quadratic Indices of the Molec-ular Pseudograph's Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. *Molecules,* **2003**, *8*, 687-726.
[http://dx.doi.org/10.3390/80900687]

[22] Olariu, T.; Vlaia, V.; Ciubotariu, C.; Dragos, D.; Ciubotariu, D.; Mracec, M. Quantitative relationships for the prediction of the vapor pressure of some hydrocarbons from the van der Waals molecular surface. *J. Serb. Chem. Soc.,* **2015**, *80*, 659-671.
[http://dx.doi.org/10.2298/JSC140416051O]

[23] Pourbasheer, E.; Ahmadpour, S.; Zare-Dorabei, R.; Nekoei, M.M. Quantitative structure activity relationship study of p38a MAP kinase inhibitors. *Arab. J. Chem.,* **2017**, *10*(1), 33-43.
[http://dx.doi.org/10.1016/j.arabjc.2013.05.009]

[24] Saghaie, L.; Sakhi, H.; Sabzyan, H.; Shahlaei, M.; Shamshiri-an, D. Stepwise MLR and PCR QSAR study of the pharma-ceutical activities of antimalarial3-hydroxypyridinone agents using B3LYP/6-311++ G** descriptors. *Med. Chem. Res.,* **2013**, *22*, 1679-1688.
[http://dx.doi.org/10.1007/s00044-012-0152-5]

[25] Gramatica, P. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.,* **2007**, *26*(5), 694-701.
[http://dx.doi.org/10.1002/qsar.200610151]

[26] Cramer, R.D., III; Bunce, J.D.; Patterson, D.E.; Frank, e. Cross-validation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Mol. Inform.,* **1988**, *7*, 18-25.

[27] Votano, J.R.; Parham, M.; Hall, L.H.; Kier, L.B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis,* **2004**, *19*(5), 365-377.
[http://dx.doi.org/10.1093/mutage/geh043] [PMID: 15388809]

[28] Chatterjee, S.; Simonoff, J. *Handbook of Regression Analysis*; John Wiley & Sons: New York, **2013**.

[29] Neda Ahmadinejad, N.; Shafiei, F.; Momeni. Isfahani, T. Quantitative Structure- Property Relationship (QSPR) Investi-gation of

Camptothecin Drugs Derivatives. *Comb. Chem. High Throughput Screen.,* **2018**, *21*, 1-10.

[30] Shen, J.; Cui, Y.; Gu, J.; Li, Y.; Li, L. A genetic algorithm- back propagation artificial neural network model to quantify the affinity of flavonoids toward P-glycoprotein. *Comb. Chem. High Throughput Screen.,* **2014**, *17*(2), 162-172.
[http://dx.doi.org/10.2174/1386207311301010002]
[PMID: 24206113]

[31] Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.,* **2002**, *20*(4), 269-276.
[http://dx.doi.org/10.1016/S1093-3263(01)00123-1]
[PMID: 11858635]

[32] Xu, J.; Zhu, L.; Fang, D.; Liu, L.; Bai, Z.; Wang, L.; Xu, W. A simple QSPR model for the prediction of the adsorbability of organic compounds onto activated carbon cloth. *SAR QSAR Environ. Res.,* **2013**, *24*(1), 47-59.
[http://dx.doi.org/10.1080/1062936X.2012.728997]
[PMID: 23066906]

[33] Rose, K.; Hall, L.H.; Kier, L.B. Modeling blood-brain barrier partitioning using the electrotopological state. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*(3), 651-666.
[http://dx.doi.org/10.1021/ci010127n] [PMID: 12086527]

[34] Wold, S.; Erikson, L. Statistical Validation of QSAR Results. Validation Tools. *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH Publishers: New York, **1995**, pp. 309-318.