Journal of Physical and Theoretical Chemistry
of Islamic Azad University of Iran, 17 (3, 4) 93-105: Fall 2020 & Winter 2021
(J. Phys. Theor. Chem. IAU Iran)
ISSN 1735-2126

QSPR Models to Predict Thermodynamic Properties of Alkenes Using Genetic Algorithm and Backward- Multiple Linear Regressions Methods

Fatemeh Ghaemdoost and Fatemeh Shafiei*

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

Received July 2020; Accepted September 2020

ABSTRACT

Quantitative structure–property relationship (QSPR) models establish relationships between different types of structural information to their properties. In the present study the relationship between the molecular descriptors and quantum properties consist 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}) of 100 alkenes is represented. Genetic algorithm (GA) and backward-multiple linear regressions (BW-MLR) were successfully developed to predict quantum properties of alkenes. Molecular descriptors were calculated with Dragon software and the genetic algorithm (GA) method was used to selected important molecular descriptors. The quantum properties were obtained from quantum-chemistry technique at the Hartree-Fock (HF) level using the ab initio 6-31G* basis sets. The predictive powers of the BW-MLR models were discussed by using leave-one-out (LOO) cross-validation and external test set. Results showed that the predictive ability of the models was satisfactory, and the 2D matrix-based descriptors, topological, edge adjacency and Connectivity indices could be used to predict the mentioned properties of 100 alkenes.

Keywords: Backward-Multiple linear regression; Molecular descriptors, Genetic algorithm; validation; QSPR; alkenes

INTRODUCTION

Quantitative structure–property relationships (QSPR) and quantitative structure–activity (QSAR) models are mathematical equations that relate properties or activities of compounds to a wide range of molecular descriptors [1]

QSAR and QSPR studies are unquestionably of great importance in Biochemistry, analytical chemistry, physical chemistry, pharmaceutical, environmental chemistry and toxicology [2, 3]. The aim of these studies is to search for new compounds with the required properties and activities by mathematical and computer methods [4, 5].

Molecular descriptors are closely related to the concept of molecular structure and they are developed for the purpose of obtaining correlations with physicochemical properties and biological activities of chemical substances have been applied for a very extensive range [3, 6, 7].

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

The experimental properties namely the octanol-water partition coefficient (logP), melting point, boiling point, aqueous solubility (Sw) and polarizability (α) of linear alkanes and alkenes have been investigated using a novel index based on connectivity and distances in the graph of a molecular structure [8].

A QSPR analysis has been applied to derive a quantitative relationship between the chemical structures of 91 alkenes and their physicochemical properties such as enthalpy of vaporization at standard condition (ΔH°_{vap}/kJ.mol^{-1}) and normal temperature of boiling points (T°_{bp} /K)[9].

Artificial neural networks (ANNs) have been used to construct QSPR models for predicting the normal boiling point, density, and refractive index of 66 alkenes [10].

General regression neural network (GRNN) and stepwise multiple linear regression (MLR) techniques were applied to develop QSAR models for the prediction reaction rate constants of ozone of 95 alkenes [11].

The novel information theoretic topological index, I_k, is derived from the edge signed graphs has been applied to predict three properties of 24 unsaturated hydrocarbons (Alkenes) using multiple regression analysis (MRA)[12].

A PCR analysis was applied to find a multiparametric QSPR model between 15 different properties of 149 alkanes and eleven topological indices (Sh indices)[13].

A QSPR model has been used to estimate critical volume of unsaturated hydrocarbon alkenes and alkynes using simple connectivity indices [14].

QSPR study has been devoted to predict physical and chemical properties such as density (D), boiling point (BP) and melting point (MP) of 162 mono alkenes using ad hoc descriptors and molecular connectivity indices [15].

In the present study, QSPR mathematical models have been developed to predict the thermal energy (E_{th}/kJ mol^{-1}) heat capacity (Cv/J mol^{-1} K^{-1}) and entropy(S/J mol^{-1} K^{-1}) of 100 alkenes using BW-MLR method based on molecular descriptors calculated from the molecular structure by using Dragon software, and also several methods have been used for testing the predictive ability of the models.

MATERIALS AND METHODS

The thermal energy, heat capacity and entropy of 100 alkenes were taken from the quantum mechanics methodology with the ab initio Hartree-Fock theory, using 6-31G* basis sets [16].

These compounds and their quantum properties are listed in Table **1**. In order to build and test QSPR models, a data set of 100 alkenes was randomly separated into 2 groups: a training set of 80 compounds, which was used to build a model and a test set of 20 compounds, which was applied to evaluate the built model. In order to calculate the theoretical descriptors, first, the molecular structures were constructed using Gauss View 5 software and then, the molecular geometries of compounds were better optimized with Gaussian 98 programs [17]. These optimized structures were entered in Dragon package 2.1. A total of 1896 theoretical descriptor were calculated for each compound in the data set using Dragon software.

DRAGON software is a very important tool for the calculation of a wide range of descriptors including different groups: topological, 2D autocorrelations, aromaticity indices, geometrical GETAWAY, radial distribution function (RDF), 3D-MoRSE, Galvez topological charge, weighted holistic invariant molecular (WHIM), empirical, functional groups, atom-centered fragments, and constitutional descriptors [18, 19].

The Genetic Algorithm (GA) is implemented in MATLAB (2010a) software and backward stepwise-linear

multiple regression method using the Statistical Package for the Social Science (SPSS) software Version 20 were used to reduce the number of molecular descriptors and build QSPR models [20].

The QSPR models were evaluated with regression parameters: 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 (DW) and significance (Sig) [21-23].

RESULTS AND DISCUSSION

QSPR models and statistical coefficients

Table 1. The name of 100 alkenes and their thermal energy(E_{th}/kJ mol^{-1}), heat capacity(Cv/J mol^{-1} K^{-1})and entropy(S/J mol^{-1} K^{-1})used in this study

No.	Compound	Cv J mol^{-1} K^{-1}	S J mol^{-1} K^{-1}	E_{th} kJ mol^{-1}	No.	Compound	Cv J mol^{-1} K^{-1}	S J mol^{-1} K^{-1}	E_{th} kJ mol^{-1}
1	1-Butene	61.14	279.81	317.06	51	2-pentene	78.86	302.56	397.61
2	1-heptene	120.43	370.43	568.00	52	3,3dimethyl 1-butene*	116.29	325.48	481.05
3	1-hexene	99.26	339.86	484.33	53	3,3dimethyl 1-heptene	164.62	415.37	732.77
4	1-nonene*	165.83	431.7	735.35	54	3,3dimethyl 1-hexene	141.24	384.68	649.1
5	1-octene	145.81	401.06	651.68	55	3,3dimethyl 1-pentene	123.93	354.24	565.37
6	1-Pentene	80.47	309.63	400.67	56	3,4diethyl 2-Hexene	169.77	468.72	821.44
7	1-propene	32.77	254.03	233.41	57	3,4dimethyl 1-pentene*	120.13	372.94	568.49
8	2,3,3trimethyl 1-butene*	132.24	335.19	562.09	58	3,4dimethyl 2-Hexene	133.93	423.14	652.69
9	2,3dimethyl 1-butene	110.14	351.9	485.42	59	3,4dimethyl 2-pentene	119.62	387.28	568.71
10	2,3dimethyl 1-heptene	164.23	442.13	736.7	60	3,5dimethyl 1-heptene	164.34	432.73	736.48
11	2,3dimethyl 1-hexene	139.69	411.48	653.02	61	3,6dimethyl 1-octene*	187.92	473.05	820
12	2,3dimethyl 1-pentene*	118.68	380.84	569.33	62	3,7dimethyl 1-octene	203.12	468.63	819.66
13	2,3dimethyl 2-butene	108.53	346.78	478.64	63	3ethyl 1-heptene	160.44	439.86	737.58
14	2,3dimethyl 2-heptene	160.22	415.73	730.55	64	3ethyl 1-hexene	134.86	409.25	653.89
15	2,3dimethyl 2-hexene	136.48	386.96	646.89	65	3ethyl 1-pentene	114.64	378.31	570.24
16	2,3dimethyl 2-pentene	118.82	358.42	563.28	66	3ethyl 2-heptene	155.63	407.19	732.8
17	2,4,4trimethyl 2-pentene*	146.78	386.4	647.27	67	3ethyl 2-pentene	111.43	347.74	565.6
18	2,4dimethyl 1-heptene	166.09	438.02	736.36	68	3-heptene*	108.74	355.68	564.91
19	2,4dimethyl 1-hexene*	138.60	407.27	652.66	69	3-hexene	96.05	326.09	481.28
20	2,4dimethyl1-pentene	126.79	376.66	568.68	70	3methyl 1-butene*	86.37	319.57	402.2
21	2,4dimethyl 2-heptene	162.07	446.9	735.73	71	3methyl 1-heptene	141.42	411.46	653.5
22	2,4dimethyl 2-hexene*	135.70	416.19	652.02	72	3methyl 1-hexene	119.50	380.85	569.83
23	2,4dimethyl 2-pentene	124.38	385.18	568.12	73	3methyl 1-octene	169.07	442.02	737.18
24	2,5,5trimethyl 2-hexene	169.36	399.36	727.75	74	3methyl 1-pentene	101.80	350.16	486.13
25	2,5dimethyl 1-hexene	147.31	411.22	652.54	75	3methyl 2-heptene	137.41	399.49	647.80
26	2,5dimethyl 2-heptene	164.48	442.75	735.69	76	3methyl 2-hexene	116.29	368.23	564.13
27	2,5dimethyl 2-hexene	144.10	412.4	651.71	77	3methyl 2-pentene*	94.99	341.15	480.50
28	2,6dimethyl 2-octene*	188.23	478.2	819.68	78	3-octene	139.38	385.19	648.57
29	2-butene	60.34	271.53	313.96	79	4,4dimethyl1-pentene	130.64	352.14	564.57
30	2ethyl3methyl 1-butene	120.43	372.94	569.46	80	4,4dimethyl 2-hexene	138.83	376.44	645.89
31	2ethyl 1-butene	100.19	314.14	482.48	81	4,4dimethyl 2-pentene	127.43	347.74	561.58
32	2ethyl 1-hexene	138.21	373.95	649.68	82	4,5dimethyl 2-heptene*	152.23	434.91	735.61
33	2ethyl 1-pentene	117.09	343.77	566.02	83	4ethyl 2methyl 1-hexene	160.32	434.07	736.88
34	2-heptene	117.21	362.55	564.89	84	4methyl 1-octene	168.27	439.74	737.17
35	2-hexene	96.86	332.41	481.23	85	4methyl 1-pentene	106.20	348.04	485.82
36	2-methyl 2butene*	84.92	305.23	396.45	86	4methyl 2-heptene	137.41	403.54	650.26
37	2-methyl 3nonene	201.33	474.03	820.15	87	4methyl 2-hexene	117.09	382.24	569.12
38	2methyl 4ethyl 1-heptene	187.27	464.36	820.54	88	4methyl 2-octene	163.45	443.48	736.49
39	2methyl 4ethyl 2-hexene	156.31	443.28	736.1	89	4methyl 2-pentene	103.80	351.67	485.21
40	2methyl 1-butene*	85.64	294.25	397.62	90	4methyl 1-hexene*	115.22	378.6	569.80
41	2methyl 1-hexene	122.40	353.55	564.9	91	5,5dimethyl 1-hexene	150.46	386.00	648.45
42	2methyl 1-pentene	104.60	323.6	481.25	92	5,5dimethyl 2-hexene	146.45	374.53	645.11
43	2methyl 2-heptene*	135.90	387.19	647.49	93	5ethyl 1-heptene	163.66	438.17	737.86
44	2methyl 2-hexene	119.99	358.42	563.82	94	5ethyl 2-heptene	158.84	437.00	737.04
45	2methyl 2-pentene	102.99	328.83	480.21	95	5methyl 1-heptene	143.03	412.21	653.75
46	2methyl 3-heptene	140.76	412.83	652.8	96	5methyl 1-hexene	124.81	380.07	569.75
47	2methyl 3-hexene	119.99	382.34	569.16	97	5methyl 2-heptene	139.01	411.32	652.79
48	2Methyl 1-propene	80.89	267.61	313.28	98	5methyl 2-hexene	121.60	380.10	568.85
49	2methyl 1-octene*	165.07	414.81	732.24	99	6methyl 1-heptene*	141.69	411.10	653.43
50	2-octene	119.99	393.22	648.57	100	6methyl 2-heptene	143.98	412.14	652.78

* Compounds selected for test set in external validation procedure

QSPR models for the thermal energy

The BW–MLR analysis led to the derivation of 13 models for thermal energy (E_{th}), with 13- 21descriptors. Table **2**, shows the regression coefficient and statistical parameters of models for thermal energy (E_{th}) of 80 alkenes.

The results of the models were observed to be very satisfactory. The statistical coefficients of the 13 models were almost similar; so, the model 13, which had the lowest number of descriptors, was selected. The QSPR model and statistical parameters for nine molecular descriptors are shown as follows (Equation (1)):

E_{th}= 172.427-186.025 (HNar) +76.023 (PCR) -15.264 (X4) +124.588 (X1v) -16.747 (X3sol) +4.871 (RDSQ) +200.246 (HDcpx) -4.602 (EEig06d) -5.906 (ESpm14x) (1)

N_{train}=80, R_{train}=0.998, R^2_{train}=0.995, $R^2_{adj,train}$=0.995, RMSE=2.939, F_{train}=1618.649, DW_{train}=2.113, Sig_{train}=0.000

QSPR models for the entropy

Table 3, shows the statistical parameters of 14 models for the entropy of 80 alkenes. The statistical coefficients of the 14 models were almost similar; so, the model 14, which had the lowest number of descriptors, was selected. The QSPR model and statistical parameters for nine molecular descriptors are shown as follows (Equation (2)):

S= 138.404 - 1663.337 (HNar) + 1419.576 (GNar) + 173.091 (MSD) -35.406 (Har) + 26.039 (Jhetv) + 53.335 (MAXDN) + 16.818 (S2K) + 28.901 (XMOD) - 11.999 (ESpm10d) (2)

N_{train}=80, R_{train}=0.991, R^2_{train}=0.981, $R^2_{adj,train}$=0.979, $RMSE_{train}$=2.639, F_{train}=412.221, DW_{train}=1.523, Sig_{train}=0.000.

Table 2.Statistical parameters of the models calculated with the SPSS software for the thermal energy (E_{th}/kJ mol^{-1}).

Model	Independent Variable	R	R^2	R^2_{adj}	RMSE	F
1	ESpm10d, HNar, PCR, EEig06d, X4Av, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, QW, HDcpx, RDSQ, WA, X1v, XMOD, GNar	0.998	0.996	0.994	3.076	628.923
2	ESpm10d, HNar, PCR, EEig06d, X4Av, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, HDcpx, RDSQ, WA, X1v, XMOD, GNar	0.998	0.996	0.994	3.053	671.754
3	ESpm10d, HNar, PCR, EEig06d, X4Av, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, HDcpx, RDSQ, WA, X1v, GNar	0.998	0.996	0.994	3.050	719.094
4	HNar, PCR, EEig06d, X4Av, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, HDcpx, RDSQ, WA, X1v, GNar	0.998	0.996	0.994	3.038	771.641
5	HNar, PCR, EEig06d, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, HDcpx, RDSQ, WA, X1v, GNar	0.998	0.996	0.994	3.025	830.321
6	HNar, PCR, EEig06d, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, Jhetm, HDcpx, RDSQ, WA, X1v	0.998	0.996	0.995	3.014	895.378
7	HNar, PCR, EEig06d, X4, BIC1, X3sol, CSI, X0Av, X2A, IDE, ESpm14x, HDcpx, RDSQ, WA, X1v	0.998	0.996	0.995	3.003	969.752
8	HNar, PCR, EEig06d, X4, BIC1, X3sol, X0Av, X2A, IDE, ESpm14x, HDcpx, RDSQ, WA, X1v	0.998	0.996	0.995	2.932	1054.676
9	HNar, PCR, EEig06d, X4, BIC1, X3sol, X0Av, IDE, ESpm14x, HDcpx, RDSQ, WA, X1v	0.998	0.996	0.995	2.932	1150.652
10	HNar, PCR, EEig06d, X4, BIC1, X3sol, X0Av, ESpm14x, HDcpx, RDSQ, WA, X1v	0.998	0.996	0.995	2.938	1235.763
11	HNar, PCR, EEig06d, X4, BIC1, X3sol, X0Av, ESpm14x, HDcpx, RDSQ, X1v	0.998	0.995	0.995	2.936	1351.685
12	HNar, PCR, EEig06d, X4, X3sol, X0Av, ESpm14x, HDcpx, RDSQ, X1v	0.998	0.995	0.995	2.938	1484.039
13	HNar, PCR, EEig06d, X4, X3sol, ESpm14x, HDcpx, RDSQ, X1v	0.998	0.995	0.995	2.939	1618.649

Table 3. Statistical parameters of the models calculated with the SPSS software for the entropy (S/J mol^{-1} K^{-1})

Model	Independent Variable	R	R^2	R^2_{adj}	RMSE	F
1	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTI, EEig02x, GMTIV, Jhetv, BIC1, MSD, ESpm14u, X1A, TIE, ECC, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.978	2.678	159.551
2	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTI, EEig02x, GMTIV, Jhetv, BIC1, MSD, ESpm14u, X1A, ECC, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.978	2.656	170.081
3	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTI, EEig02x, GMTIV, Jhetv, BIC1, MSD, ESpm14u, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.979	2.655	181.663
4	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTI, GMTIV, Jhetv, BIC1, MSD, ESpm14u, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.979	2.643	194.460
5	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTI, GMTIV, Jhetv, MSD, ESpm14u, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.979	2.633	208.655
6	ESpm10d, Hnar, EEig06d, MAXDN, CIC2, GMTIV, Jhetv, MSD, ESpm14u, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.980	2.622	224.501
7	ESpm10d, Hnar, MAXDN, CIC2, GMTIV, Jhetv, MSD, ESpm14u, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.980	2.619	239.795
8	ESpm10d, Hnar, MAXDN, CIC2, GMTIV, Jhetv, MSD, X1A, S2K, ESpm07x, Xu, w, GNar, XMOD, Har	0.992	0.984	0.980	2.618	255.803
9	ESpm10d, Hnar, MAXDN, CIC2, GMTIV, Jhetv, MSD, X1A, S2K, Xu, w, GNar, XMOD, Har	0.992	0.983	0.980	2.618	274.117
10	ESpm10d, Hnar, MAXDN, GMTIV, Jhetv, MSD, X1A, S2K, Xu, w, GNar, XMOD, Har	0.992	0.983	0.980	2.616	296.185
11	ESpm10d, Hnar, MAXDN, GMTIV, Jhetv, MSD, S2K, Xu, w, GNar, XMOD, Har	0.991	0.983	0.979	2.629	314.121
12	ESpm10d, Hnar, MAXDN, GMTIV, Jhetv, MSD, S2K, w, GNar, XMOD, Har	0.991	0.982	0.979	2.626	344.059
13	ESpm10d, Hnar, MAXDN, Jhetv, MSD, S2K, w, GNar, XMOD, Har	0.991	0.982	0.979	2.637	372.094
14	ESpm10d, Hnar, MAXDN, Jhetv, MSD, S2K, GNar, XMOD, Har	0.991	0.981	0.979	2.639	412.221

QSPR models for the heat capacity

Table 4, shows the regression coefficients and statistical factors of models for the heat capacity of 80 alkenes. The regression parameters of the suitable linear model for the heat capacity incudes fourteen molecular descriptors are collected in Equation (3).

Cv = 237.448+18.294 (Ss) - 212.266 (RBF) + 0.149 (GMTI) - 0.179 (GMTIV) - 15.288 (RHyDp) +16.700 (Jhetv) + 7.996 (MAXDN) -11.037 (S2K) - 353.329 (X1A) + 14.114 (XMOD) + 29.668 (BIC1) - 53.121 (ESpm06X) + 9.522 (ESpm14x) - 4.829 (ESpm10d) (3)

N_{train}=80, R_{train}=0.997, R^2_{train}=0.994, $R^2_{adj,train}$=0.992, $RMSE_{train}$=1.678, F_{train}=710.480, DW_{train}=1.588, Sig_{train}=0.000.

In the present study, to find the best BW-MLR models for predicting the mentioned properties of alkenes, we used the following sections.

Multicollinearity

The collinearity, reliability, stability and robustness of the models are influenced by the autocorrelation and multicollinearity properties of the descriptors contributed in the models. These parameters in the models were examined by calculating the variance inflation factor (VIF) and Durbin-Watson (DW) statistics [24-26]. The VIF shows us how much the variance of the coefficient estimate is increased by multicollinearity. If the VIF value lies between1-10, then there is no multicollinearity, and if the VIF<1 or >10, then there is multicollinearity.

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

Model	Independent Variable	R	R^2	R^2_{adj}	RMSE	F
1	ESpm10d, RBF, EEig06d, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14u, ESpm14x, ECC, HDcpx, Ss, UNIP, ESpm04u, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.735	466.631
2	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14u, ESpm14x, ECC, HDcpx, Ss, UNIP, ESpm04u, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.638	499.481
3	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14u, ESpm14x, ECC, HDcpx, Ss, ESpm04u, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.631	534.783
4	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14x, ECC, HDcpx, Ss, ESpm04u, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.637	572.405
5	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14x, ECC, HDcpx, Ss, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.631	617.074
6	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14x, HDcpx, Ss, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.679	660.263
7	ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14x, Ss, GMTIV, RHyDp, ESpm06x, XMOD	0.997	0.994	0.992	1.678	710.48

Good regression model should not have happened multicollinearity.

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 .(see 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 (PCC) and collinearity statistics in the ANOVA table. The results of this study are recorded in Tables 5 to7.

The suitable linear model for prediction of the thermal energy (Equation **1**) includes nine molecular descriptors (HNar, PCR, EEig06d, X4, X3sol, ESpm14x, HDcpx, RDSQ and X1v).

From Table 5, the Pearson correlation between RDSQ and X1v descriptors is close to unity, and VIF for RDSQ, HDcpx, X1v,and HNar are bigger than 10(see Table 5), therefore there is a linearity between these descriptors. After removing X1v from this model, and the next step PCR and HDcpx, we corrected Equation 1 as follows:

$$E_{th} = 366.464 + 15.537\ (RDSQ) - 16.303\ (EEig06d) - 8.567\ (ESpm14x) \quad (4)$$

N_{train}=80, R_{train}=0.991, R^2_{train}=0.981, $R^2_{adj,train}$=0.981, F_{train}=1337.423, DW_{train}=1.911, Sig_{train}=0.000, RMSE=2.430

The suitable linear model for prediction of the entropy (Equation **2**) includes nine molecular descriptors (ESpm10d, HNar, MAXDN, Jhetv, MSD, S2K, GNar, XMOD and Har). From Table 6, the Pearson correlation between (GNar, HNar) and (XMOD and Har)descriptors are close to unity, and VIF for ESpm10d, HNar, Jhetv, MSD, S2K, GNar, XMOD and Har are bigger than 10 (see Table 6), therefore there is a linearity between these descriptors. After removing XMOD from this model, and the next step GNar and Jhetv, we corrected Equation 2 as follows:

$$S = 241.191 + 21.765\ Har + 15.762\ MAXDN - 6.529\ ESpm10d \quad (5)$$

N_{train}=80, R_{train}=0.964, R^2_{train}=0.930, $R^2_{adj,train}$=0.927, F_{train}=335.457, DW_{train}=1.893, Sig_{train}=0.000, RMSE=2.383

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

	Pearson Correlation for E_{th}									Collinearity Statistical			Corrected model
	HNar	**PCR**	**EEig06d**	**X4**	**X3sol**	**ESpm14x**	**HDcpx**	**RDSQ**	**X1v**	**Tolerance**	**VIF**	**VIF**	**VIF**
HNar	1	-0.276	-0.035	-0.100	-0.017	0.381	-0.217	0.793	-0.766	0.038	26.465	-	-
PCR		1	-0.028	0.149	0.351	-0.380	0.367	0.005	-0.047	0.562	1.780	1.308	-
EEig06d			1	0.000	-0.042	0.046	0.183	-0.110	0.070	0.696	1.436	1.427	1.2
X4				1	0.278	-0.091	0.157	-0.080	-0.100	0.202	4.959	-	-
X3sol					1	-0.408	0.289	0.157	-0.311	0.156	6.422	-	-
ESpm14x						1	-0.781	-0.166	0.230	0.102	9.802	1.608	1.475
HDcpx							1	0.176	-0.334	0.040	24.941	8.202	-
RDSQ								1	-0.958	0.006	169.470	6.841	1.579
X1v									1	0.003	292.944	-	-

The suitable linear model for prediction of the heat capacity (Equation **3**) includes fourteen molecular descriptors (ESpm10d, RBF, MAXDN, GMTI, BIC1, X1A, Jhetv, S2K, ESpm14x, Ss, GMTIV, RHyDp, ESpm06x and XMOD). From Table 7, the Pearson correlation between (RHyDp, XMOD) and (ESpm06x, ESpm14x) descriptors are close to unity, and VIF for these descriptors are bigger than 10 (see Table 6), therefore there is a linearity between these descriptors. After removing RHyDp from this model, and the next step ESpm06x, XMOD, RBF and Jhetv, we corrected Equation 3 as follows:

$$C_v = 277.653 + 0.201\ (GMTIV) - 350.058\ (X1A) \quad (6)$$

N_{train}=80, R_{train}=0.965, R^2_{train}=0.931, $R^2_{adj,train}$=0.929, F_{train}=521.719, DW_{train}=1.782, Sig_{train}=0.000, RMSE=1.922

Validation

Validation is the important step in QSAR/QSPR modeling in order to ensure the model created is a good model or a poor model [27-29]. There are several techniques to approximate the quality and accuracy of the QSPR model [31].

In this section, for verification, validity of the regression models and the predictive ability and statistical significance of the QSPR models, internal validation and external validation technique was applied to made model by splitting of set of chemical compounds into a training set (80%) and a test set(20%)[30]. From the internal validation technique, the leave one- out cross-validation (LOOcv) method was used to validate the selected QSPR models; the value of Q^2 LOO can be calculated as the following:

$$Q^2 = 1 - \frac{\sum(Y_i - \hat{Y}_{i|i})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{PRESS}{TSS} \qquad Q^2 \leq 1 \qquad (7)$$

In the Equation (7), 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 and PRESS is the sum of squares of the prediction errors and TSS represents the total sum of squares [31].

The Q^2 LOO values of the thermal energy (E_{th}/kJ mol^{-1}), heat capacity (Cv/J mol^{-1} K^{-1}) and entropy(S/J mol^{-1} K^{-1}) models (Equations (4-6)) were calculated as 0.978, 0.987 and 0.929, respectively.

Statistical factors such as R, R^2, R^2adj, F, and RMSE of the best models (Equations (4-6)) 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 reported in Table 7.

Table 6. Correlation between the molecular descriptors (Eq. (3))

1	Pearson Correlation for cv														Collinearity Statistical	
	ESpm10d	**RBF**	**MAX**	**GMTI**	**BIC1**	**X1A**	**Jhetv**	**S2K**	**ESpm14x**	**Ss**	**GMTIV**	**RHyDp**	**ESpm06x**	**XMOD**	**VIF**	**VIF**
ESpm10d	1	-0.286	0.008	0.450	-0.155	0.498	0.477	0.380	0.647	0.168	0.352	0.463	0.345	0.456	123.263	-------
RBF		1	0.208	0.207	-0.294	0.155	0.366	0.558	0.104	0.615	0.010	0.233	0.156	0.153	40.737	17.362
MAXDN			1	0	-0.401	0.305	0.135	0.059	0.231	0.208	0.001	0.060	0.297	0.103	3.263	2.528
GMTI				1	0.037	0.023	0.332	0.099	0.074	0.477	-0.542	-0.556	0.105	0.187	516.225	------
BIC1					1	0.352	0.358	0.296	0.288	0.169	0.007	-0.186	-0.210	0.011	8.802	------
X1A						1	0.306	0.325	0.806	.0.275	-0.145	-0.144	0.836	0.143	59.702	15.057
Jhetv							1	0.072	0.196	0.116	-0.126	-0.383	-0.079	0.122	31.611	19.936
S2K								1	0.184	0.463	-0.039	0.093	0.182	0.779	124.216	89.129
ESpm14x									1	0.100	-0.037	-0.107	-0.915	0.168	540.984	------
Ss										1	-0.613	-0.723	0.046	0.048	1070.68	386.081
GMTIV											1	0.526	-0.118	0.104	590.408	58.803
RHyDp												1	-0.142	0.943	1457.24	------
ESpm06x													1	0.034	696.531	402.275
XMOD														1	741.388	-----

Table 6. continued

2	Pearson Correlation for cv								Collinearity Statistical			
	ESpm06x	**RBF**	**MAXDN**	**GMTIV**	**Jhetv**	**X1A**	**S2K**	**Ss**	**VIF**	**VIF**	**VIF**	**VIF**
ESpm06x	1	-0.071	0.304	-0.082	-0.566	0.655	0.526	-0.069	28.302	18.770	-----	-----
RBF		1	-0.106	0.782	0.717	-0.304	-0.597	-0.812	14.584	3.061	2.913	-----
MAXDN			1	-0.019	-0.070	0.274	0.514	-0.137	2.509	1.508	1.494	-----
GMTIV				1	0.652	-0.459	-0.538	-0.938	55.675	2.946	2.945	1.390
Jhetv					1	-0.487	-0.531	-0.641	19.033	8.903	4.101	-----
X1A						1	0.424	0.416	12.689	9.267	4.396	1.390
S2K							1	0.311	13.445	------	-----	-----
Ss								1	120.60	-----	------	-----

Table 7. Statistical parameters obtained by the BW- MLR model for the entropy, thermal energy and heat capacity for training and test sets (Eqs.(4)-(6))

Data set	properties	N	R	R^2	R^2_{adj}	RMSE	DW	F	sig
training	Cv J $mol^{-1}K^{-1}$	80	0.965	0.931	0.929	1.922	1.782	521.719	0.000
test	Cv J $mol^{-1}K^{-1}$	20	0.948	0.899	0.887	1.065	1.884	375.770	0.000
training	E_{th} kJ mol^{-1}	80	0.991	0.981	0.981	2.430	1.911	1337.423	0.000
test	E_{th} kJ mol^{-1}	20	0.997	0.994	0.993	3.249	1.996	883.491	0.000
training	S J $mol^{-1}K^{-1}$	80	0.964	0.930	0.927	2.383	1.893	335.457	0.000
test	S J $mol^{-1}K^{-1}$	20	0.975	0.951	0.942	1.282	1.953	103.423	0.000

Durbin-Watson Statistic

The Durbin-Watson (DW) Statistic is a test for presence of autocorrelation in the residuals from a regression analysis. The DW test reports a test statistic, with a value from 0 to 4. A value near 2 indicates nonautocorrelation; a value toward 0 indicates positive autocorrelation; and a value toward 4 indicates negative autocorrelation [32]. In our model, the Durbin-Watson values are near 2 (Table 7); therefore, there is no autocorrelation.

Residuals

The residual is the difference between the experimental (observed) value of the dependent variable (y) and the calculated (predicted) value ($\hat{y}$) is called the residual.

The residual of the BW-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 development of the BW- MLR models (see Figs 1-3).

Figures (4-6) show the linear correlation between observed and predicted values of the entropy, thermal energy, and heat capacity obtained using Equations (4-6) respectively.

Interpretation of the best descriptors

As can be seen from Table **8**, the four block of descriptors, namely, 2D matrix-based descriptors (Har index), Connectivity (X1A and RDSQ indices), Topological(MAXDN index) and Edge adjacency indices (EEig06d and ESpm14x indices) are useful to predict the mentioned properties than the other block of descriptors. This means that these descriptors have more effect on the Cv, S and E_{th} of alkenes 2D Matrix-based descriptors are calculated based on the elements of so-called graph-theoretical matrices [33] by using several algebraic operations. The Balaban-like indices inferred from the adjacency matrix [34, 35] are important examples of this category.

Connectivity indices are calculated from the vertex-degree of a molecular graph [36, 37]. The Randić index [38] is a prominent example of this category.

Topological indices are defined by various structural features into account, e.g., distances and eigenvalues. The term topological index has been firstly coined by Hosoya [39].

Edge adjacency indices are based on the edge adjacency matrix of a graph. The resulting descriptor-value is the sum of all edge entries of the adjacency matrix of a graph. Balaban developed several indices by using graph-theoretical matrices [40].

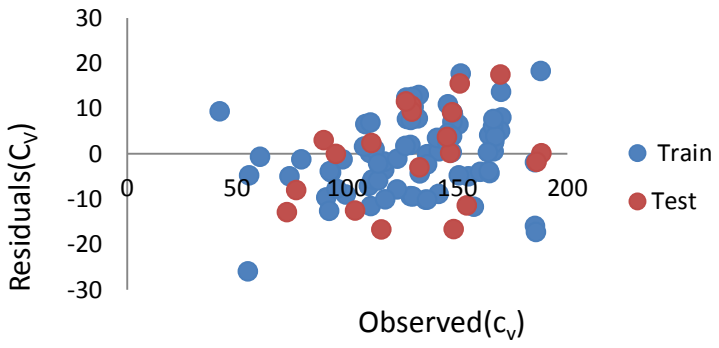

Fig. 1. Residuals plotted against the observed heat capacity (Cv/J $mol^{-1}K^{-1}$)for training and test sets of alkenes.

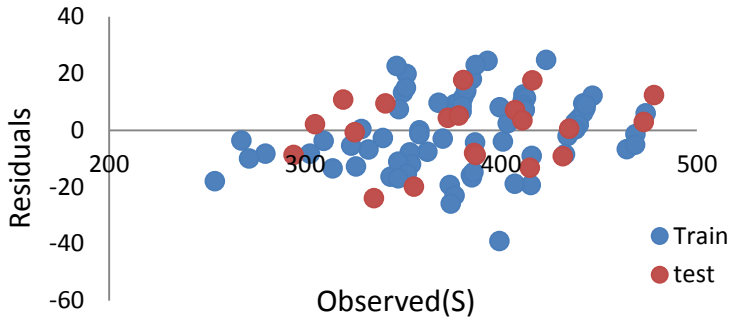

Fig. 2. Residuals plotted against the observed entropy (S/J $mol^{-1}K^{-1}$) for training and test sets of 100 alkenes.

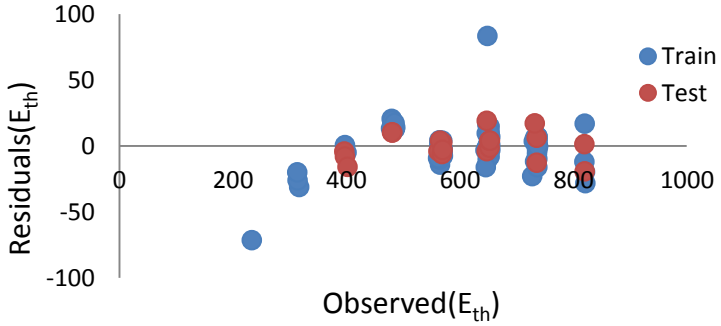

Fig. 3. Residuals plotted against the observed thermal energy (E_{th}/kJ mol^{-1}) for training and test sets of 100 alkenes.

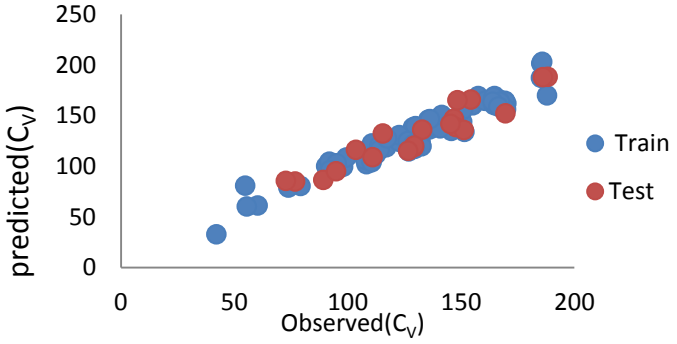

Fig. 4. Comparison between predicted and experimental values of the heat capacity (Cv/J $mol^{-1}K^{-1}$)of 100 alkenes using BW-MLR method.

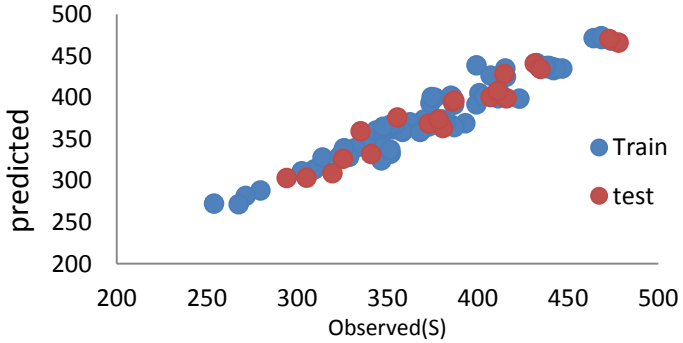

Fig. 5. Comparison between predicted and experimental values of the entropy(S/J $mol^{-1}K^{-1}$) of 100 alkenes using BW-MLR method.

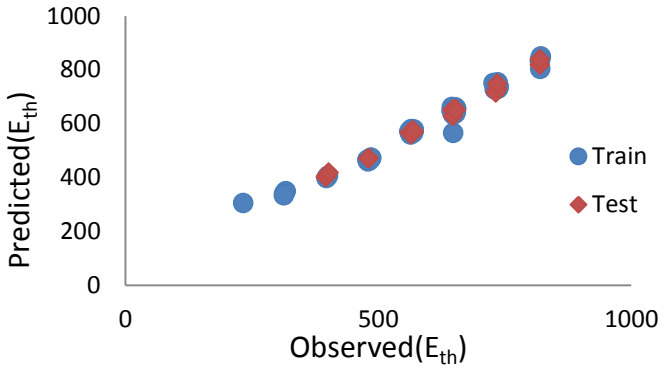

Fig. 6. Comparison between predicted and experimental values of the thermal energy (E_{th}/kJ mol^{-1}) of 100 alkenes using BW-MLR method.

Table8. List of the best selected molecular descriptors that appear in the final models

N	Property	Symbol	description	Block description
80	Cv	X1A	average connectivity index of order 1	Connectivity indices
		GMTIV	Gutman Molecular Topological Index by valence vertex degrees	Topological indices
80	S	ESpm10d	Spectral moment 10 from edge adj. matrix	Edge adjacency indices
		MAXDN	maximal electrotopological negative variation	Topological indices
		Har	Reciprocal squared distance matrix (H2)	2D matrix-based descriptors
80	E_{th}	EEig06d	Eigenvalues	Edge adjacency indices
		ESpm14x	Spectral moment 14 from edge adj. matrix weighted by edge degrees	Edge adjacency indices
		RDSQ	reciprocal distance sum inverse Randic-like index	Connectivity indices

CONCLUSION

QSPR studies are mathematical relationships between the properties studied and their molecular descriptors.

In the present study, QSPR models have been developed to predict the thermal energy (E_{th}/kJ mol^{-1}), heat capacity (Cv/J$mol^{-1}K^{-1}$) and entropy(S/J $mol^{-1}K^{-1}$) of 100 alkenes. These properties were obtained from quantum-chemistry technique at the Hartree-Fock (HF) level using the ab initio 6-31G* basis sets. The Backward stepwise regression and Genetic Algorithm (GA) technique was applied to select the most important molecular descriptors and BW-MLR method was used to build QSPR models for the prediction of the studied properties. Molecular descriptors calculated with Dragon software. The statistical parameters such as the squared correlation coefficient (R^2), adjusted correlation coefficient (R^2_{adj}), Fisher ratio (F) and Root Mean Square Error (RMSE) have been used to evaluate the quality and predictive ability of proposed BW-MLR models. The leave one-out cross-validation (LOOcv) and external validation methods were used to validate the selected QSPR models. The validation results suggest that the models possess good predictive ability and robustness. The BW-MLR results indicated that the statistical coefficients are very satisfactory and there is suitable linear relationship between the quantum properties and molecular descriptors of 100 alkenes.

REFERENCES

[1] E. Pourbasheer, R. Aalizadeh, M. R. Ganjali and P. Norouzi, Med. Chem. Res. 23 (2014) 57.

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

[3] Das, K. C.; Zhou, B.; Trinajstić, N. Bounds on Harary index. J. Math. Chem. 49 (2009) 1369.

[4] S. Ahmadi and E. Habibpour, Anti-Cancer Agents Med. Chem. 17 (2017) 552.

[5] S. D. Bolboaca, L. Jantschi and M. V. Diudea, Curr. Comput- Aided. Drug. Des. 9 (2013) 195.

[6] N. Cai-hua, L. Liang-chao and F. Zhi-yun, Wuhan Univ. J. Nat. Sci. 5 (2000) 464.

[7] M. Shamsipur, R. Ghavamib, B. Hemmateenejadc and H. Sharghib, QSAR. Comb. Sci. 23 (2004) 734.

[8] G. Selcuk and T. Lemi, Chem. J. 1 (2015) 103.

[9] F. Ghaemdoost and F. Shafiei, Curr. Comput- Aided. Drug. Des. 16 (2020) 1.

[10] S. Liu, R. Zhang, M. Liu and Z. Hu, J. Chem. Inf. Comput. Sci. 37 (1997) 1146.

[11] Y. Xu, X. Yu and S. Zhang, J. Braz. Chem. Soc. 24 (2013) 1781.

[12] K. P. Sahu and S. L. Lee, Chem. Phy. Lett. 396 (2004) 465.

[13] S. D. Nelson and P. G. Seybold, J. Mol. Graph. Model. 20 (2001) 36.

[14] O. Zuas and D. Styarini, Reaktor. 12 (2009) 260.

[15] M. Shamsipur, B. Hemmateenejad, R. Ghavami and H. Sharghi, Pol. J. Chem. 81 (2007) 269.

[16] J. S. Binkley, J. A. Pople and W. J. Hehre, J. Am. Chem. Soc. 102 (1980) 939.

[17] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb and J. A. Pople, J. A. Gaussian; Inc, Wallingford CT, 2009.

[18] R. Todeschini and V. Consonni, Handbook of Molecular Descriptors; Wiley-VCH: Weinheim, 2000.

[19] R. Todeschini and V. Consonni, Molecular descriptors for chemoinformatics Alphabetical listing (2nd ed., Vol. 1); Weinheim: Wiley-VCH, 2009.

[20] M. V. Diudea, QSPR/QSAR studies for molecular descriptors; Ed Nova Science Hunting don: New York, 2000.

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

[22] B. L. Podlgar and D. M. Ferguson, Drug Des. Discov. 17 (2000) 4.

[23] F. Hadizadeh, S. Vahdani and M. Jafarpour, Iran J. Basic. Med. Sci. 16 (2013) 910.

[24] T. A. Craney and J. G. Surles, Qual. Eng. 14 (2002) 391.

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

[26] A. Fisher, Ronald. Statistical Methods for Research Workers; Oliver and Boyd: Edinburgh, UK, 1925.

[27] P. Gramatica, QSAR. Comb. Sci. 26 (2007) 694.

[28] R. D.III. Cramer, J. D. Bunce and D. E. Patterson, Quant. Struct. Act. Relat.7 (1988) 18.

[29] S. Velilla, Am. Stat. 74 (2018) 114.

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

[31] V. Consonni, D. Ballabio and R. Todeschini, J. Chem. Inf. Model. 49 (2009) 1669.

[32] P. Pratim Roy, S. Paul, I. Mitra and K. Roy, Molecules. 14 (2009) 1660.

[33] E. Estrada, J. Chem. inf. Comput. Sci. 38 (1998) 23.
[34] A. Balaban, Chem. Phys. Lett. 89 (1982) 399.
[35] J. Devillers, A. T. Balaban, Topological indices and related descriptors in QSAR and QSPR; Amsterdam, Gordon & Breach: The Netherlands.1999.
[36] L. B. Kier, L. H. Hall, Molecular Connectivity in Structure-Activity Analysis; RSP-Wiley: Chichetser (UK), 1986.
[37] H. S. Ramane, A. S. Yalnaik, J. Appl. Math. Comput. 55 (2017) 609.
[38] M. Randić, J. Am. Chem. Soc.97 (1975) 6609.
[39] H. Hosoya, Bull. Chem. Soc. Jpn. 44 (1971) 2332.
[40] R. Todeschini, R. Cazar, E. Collina, Chemom. Intell. Lab. Syst. 15 (1992) 51.