ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

QSPR Study of the Boiling Point of Diverse Hydrocarbons: Hybrid (GA/ MLR) Approach.

Nour-eddine Kertiou, Amel Bouakkadia, and Djelloul Messadi*.

Environmental and Food Safety Laboratory, Badji Mokhtar University, Annaba 23000, Algeria.

ABSTRACT

A quantitative structure- property relationship (QSPR) was performed for the prediction of the boiling points of hydrocarbons which consists of alkanes, alkenes, dienes, alkynes, cycloalkanes, and cycloalkenes. The entire set of 165 compounds was divided into a training set of 125 hydrocarbons and a test set of 40 compounds. A five descriptor model, with squared correlation coefficient (R^2) of 99.80% and standard error of estimation (s) of 4.67, was developed by applying multiple linear regression analysis using the ordinary least square regression method and genetic algorithm- variable subset selection. The reliability of the proposed model was further illustrated using various evaluation technics: leave- one- out cross- validation, bootstrap, randomization tests, and validation through the test set.

Keywords: Hydrocarbons-Boiling points- QSPR-Molecular Descriptors- Multiple Linear Regression.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Boiling point Bp is one of the most important physical property, used to describe the volatility of a compound (its presence in the atmospheric environment), defined as the temperature at which the vapor pressure of a pure saturated liquid is 760 mmHg [1]. Also, to estimate other properties such as critical temperatures, vapor pressure and flash points [2-4].

For many hydrocarbons, the values of boiling point are not available in the literature. Their experimental measurement is expensive, consumes a long time and it requires pure compounds. Moreover, the compounds of high molecular weight decompose before reaching their boiling points and require measures under reduced pressure and subsequent correction for atmospheric pressure. Therefore, the direct measurement of the boiling point of the organic compound is laborious [5].

The aim of the present work is to develop a robust QSPR[6] model that could predict the boiling point values for a diverse set of hydrocarbons (which consists of alkanes, alkenes, dienes, alkynes, cycloalkanes, and cycloalkenes) using the general molecular descriptors computed with the help of DRAGON software [7].

In this study, we present a new QSPR model for the prediction of the boiling point of a set of 165 hydrocarbons. Our goal is to develop an accurate, simple, fast, and less expensive method for calculation of boiling point values. The predictive power of resulting model is demonstrated by testing it on test data that were not used during model generation

METHODS

Experimental Data

The experimental Bp values (K) of 165 selected, structurally heterogeneous, hydrocarbons were taken from the literature [8]. The boiling point values span between 111.6 and 628.12K (Table 1). The detailed structures of all studied compounds are available as Supporting Information.

Descriptor Generation

The chemical structure of each compound was sketched on a PC using the HYPERCHEM program [9] and preoptimized using MM+ molecular mechanics method (Polack- Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of 0.01 kcal. $Å^{-1}.mol^{-1}$ as a stopping criterion. Then the geometries were used as input for the generation of 1664 descriptors from 20 different classes such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), and Molecular Walk Counts using Dragon software (version 5.4) [7].

Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 98% pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 1230 descriptors.

Selection of the training and test sets It is important to rationally define a training set from which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set.

Several procedures can be adopted for the selection of the training and test sets, the later should contain between 15 and 40% of the compounds in the full data set.

The Duplex algorithm [10] was applied in this study to separate data into two independent subsets: a training set of – compounds to build the model and a test set of the remained- compounds to evaluate its prediction ability.

ISSN: 0975-8585

The algorithm begin with a list of the n (=165) observations where the k regressors are standardized to unit length; that is,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_{jj}^{1/2}}, i = 1, 2, \dots, n ; j = 1, 2, \dots, k \quad (1)$$

Where $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the corrected sum of squares of the jth regressor. The standardized regressors are then orthonormalized. This can be done by factoring the $Z'Z$ matrix as:

$$Z'Z = T'T \quad (2)$$

Where T' is unique k×k upper triangular matrix. The elements of T can be found using the square root or cholesky method [11]. Then make the transformation

$$W = Z T^{-1} \quad (3)$$

Resulting in a new set of variables (the w's) that are orthogonal and have unit variance. Then the Euclidian distance between all possible pairs of points is calculated. The two points which are farthest apart are assigned to the estimation set. The two points in the remaining list which are farthest are assigned to the prediction set. At the third step the point which is farthest from the two points in the estimation set is added to the estimation set. At the fourth step the point which is farthest from the two points in the prediction set is included in the prediction set. The alternation between the estimation and the prediction set continues until all points in the list have been assigned to one of the two sets. Of course, once a point is assigned to a set, it is deleted from further consideration.

This algorithm was applied in the present study to separate data into two independent subsets: a training set of 125 compounds to build the model and a test set of the remained 40 compounds to evaluate its prediction ability.

Model Development and Validation

Multiple linear regression analysis (MLR) and variable selection were performed by the software MobyDigs [12] using the Ordinary Least Square regression (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [13].

The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q^2. The models with lower Q^2 are those with fewer descriptors. First of all, models with 1-2 variables were developed by the all – subset – method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and, at the same time, protect against any over parameterization, which would lead to a loss of predictive power for molecules outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [14]. The GA was stopped when increasing the model size did not increase the Q^2 value to any significant degree. Particular attention was paid to the collinearity of the selected molecular descriptors: by applying the QUIK rule (Q Under Influence of K) [15] a necessary condition for the model validity. Acceptable model is only that with a global correlation of [x + y] block (K_{xy}) greater than the global correlation of the x block (K_{xx}) variable, x being the molecular descriptors and y the response variable.

The collinearity in the original set of molecular descriptors results in many similar models that more or less yield the same predictive power (in MOBYDIGS software 100 models of different dimensionality). Therefore, when there were models of similar performance, those with higher ΔK (K_{xy}- K_{xx}) were selected and further verified.

In this work, the "breaking point" rule was used to manage this problem. This method consists of analysing the improvement in the correlation with the number of variables in the model. By plotting the R^2 values as functions of the number of descriptors, asymptotic behavior was observed, and the improvement in the correlation became less significant after a certain rank (ΔR^2<0.02-0.03). At this point (the "breaking point"), the model is considered to be optimal, representing the best compromise between correlation and parameterization.

The models were justified by the R^2, the adjusted R^2, the external Q^2_{ext}, the F ratio values, the standard error of estimation s and the significance level value p. the R^2 and adjusted R^2 were calculated using the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad (4)$$

$$R^2_{adj} = 1 - \left[\frac{N-1}{N-M-1}(1 - R^2)\right] \quad (5)$$

Where N is the number of members of the training set and M is the number of descriptors involved in the correlation. The adjusted R^2 is a better measure of the proportion of variance in the data explained by the correlation than R^2, because R^2 is somewhat sensitive to changes in N and M. The adjusted R^2 corrects for the artificiality introduced when M approaches N through the use of a penalty function which scales the result. A variance inflation factor (VIF) was calculated to test if multicollnearities existed among the descriptors, which is defined as:

$$VIF = \frac{1}{1 - R_j^2} \quad (6)$$

Where R_j^2 is the squared correlation coefficient between the jth coefficient regressed against all the other descriptors in the model. Models would not be accepted if they contain descriptors with VIFs above a value of five.

Randomization tests were also carried out to prove the possible existence of chance correlation. To do this, the dependent variable was randomly scrambled and used in the experiment. Models were then investigated with all members in the descriptor pool to find the most predictive models. The resulting models obtained on the training set with the randomized IR values should have significantly lower R^2 values than the proposed one because the relationship between the structure and property is broken. This is a proof of the proposed model's validity as it can be reasonably excluded that the originally proposed mode was obtained by chance correlation.

Validation of the models was further performed by using the external test set composed of data not used to develop the prediction model. The Q^2_{ext} is determined with Eq. (7):

$$Q^2_{ext} = 1 - [(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_{(i)})^2 / n_{ext}) / (\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2 / n_{tr})] \quad (7)$$

Here n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

According to Golbraikh et al, [16,17] a QSPR model can provide an acceptable prediction if it verifies the following conditions:

$$Q^2_{EXT} > 0.5 \quad (8\text{-a})$$

$$r^2 > 0.6 \quad (8\text{-b})$$

$$(r^2 - r_0^2)/r^2 < 0.1 \text{ or } (r^2 - r_0'^2)/r^2 < 0.1 \quad (8\text{-c})$$

$$0.85 < k < 1.15 \text{ or } 0.85 < k' < 1.15 \quad (8\text{-d})$$

r^2 is the correlation coefficient between the calculated and experimental values in the test set; r^2_0 (calculated versus observed versus) and r'^2_0 (observed versus calculated values) are the coefficients of determination ; k, k' are slopes of the regression lines through the origin of calculated versus observed and observed versus respectively.

Here

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad \text{(9-a)}$$

$$r_0^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \hat{y}_i^{r0})^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \quad \text{(9-b)}$$

$$r'^2_0 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^{r0})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{(9-c)}$$

$$k = \frac{\sum y_i \hat{y}_i}{\sum y_i^2} \quad \text{(9-d)}$$

$$k' = \frac{\sum y_i \hat{y}_i}{\sum \hat{y}_i^2} \quad \text{(9-e)}$$

Where $\hat{y}^{r0}$ and y^{r0} are defined as $\hat{y}^{r0} = ky$ and $y^{r0} = k'\hat{y}$, respectively.

The reason to use r_0^2 and require k values that are chose to 1 is that when actual versus predicted retention indices are compared, an exact fit is required, not just a correlation.

The robustness of the models and their predictivity were evaluated by both Q^2_{LOO} and bootstrap. In this last procedure K n-dimensional groups are generated by a randomly repeated selection of n- objects from the original data set.

The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 8000 times.

The proposed model was also checked for reliability and robustness by permutation testing: new models are recalculated for randomly recorded response (Y- scrambling) by using the same original independent variable matrix. After repeating this test several times (100 times in this work) it is expected to obtain new models that have significantly lower R^2 and Q^2 than the original model. If this condition is not verified the original model is not acceptable, as it was due to a chance correlation or a structural redundancy in the training set.

Obtaining a robust model does not give real information about its prediction power. This is evaluated by predicting the compounds included in the test set.

ISSN: 0975-8585

Analysis

The applicability domain (AD) [18,19] is a theoretical region in the space defined by the descriptors of the model and the modeled response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (h_{ii}) approach [20].

The warning leverage h^* is, generally, fixed at $3(m + 1)/n$, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [21], the plot of standardized residuals versus leverage values.

Table 1: Experimental and calculated Bp for the studied compounds

		Descriptors					Boiling Point		
N	Sample	MAXDN	VEv1	HATS5u	H6m	R1p+	Expt.	Calc.	Residual
1	Methane	0	1	0	0	0	111.6	102.67	8.93
2	Ethylene	0	1.414	0	0	0.059	169.4	188.41	-19.01
3	Ethane	0	1.414	0	0	0.059	184.5	188.41	-3.91
4	Acetylene	0	1.414	0	0	0.057	188.4	187.62	0.78
5	Propylene	0.25	1.715	0	0	0.076	225.5	232.59	-7.09
6	Propane	0.25	1.716	0	0	0.061	231	227.46	3.54
7	Propadiene	0.25	1.716	0	0	0.072	238.7	231.37	7.33
8	Cyclopropane	0	1.732	0	0	0.072	240.3	243.14	-2.84
9	Propyne	0.347	1.714	0	0	0.099	249.9	236.76	13.14
10	Isobutane	0.5	1.972	0	0	0.059	261.4	258.11	3.29
11	isobutylene	0.5	1.969	0	0	0.068	266.2	260.74	5.46
12*	But-1-ene	0.417	1.978	0.613	0	0.067	266.9	270.04	-3.14
13*	Buta-1,3-diene	0.361	1.985	0.331	0	0.062	268.7	268.98	-0.28
14	Butane	0.181	1.974	0.658	0	0.056	272.6	274.72	-2.12
15	E-But-2-ene	0	1.964	1.195	0	0.067	274	288.01	-14.01
16	Z-But-2-ene	0	1.964	1.062	0	0.068	276.9	287.36	-10.46
17*	vinylacetylene	0.597	1.987	0.288	0	0.097	278.1	272.25	5.85
18	But-1-yne	0.653	1.978	0.508	0	0.107	281.2	274.2	7
19	2,2-Dimethylpropane	0.75	2.204	0	0	0.051	282.6	282.79	-0.19
20	3,3-Dimethylhexane	0.658	2.779	0.997	0.002	0.044	284.00	286.17	-2.17
21*	Cyclobutane	0	2	0	0	0.057	285.7	280.21	5.49
22*	3-Methyl but-1-ene	0.685	2.206	0.9	0	0.072	293.3	300.14	-6.84
23	Penta-1,4-diene	0.583	2.217	0.348	0	0.059	299.1	296.69	2.41
24	But-2-yne	0.181	1.959	1.956	0	0.071	300.1	288.43	11.67
25	Isopentane	0.449	2.202	0.88	0	0.057	301	303.06	-2.06
26	Pent-1-ene	0.347	2.209	0.644	0	0.063	303.1	308.02	-4.92
27	2-Methyl but-1-ene	0.412	2.2	1.011	0	0.062	304.3	306.93	-2.63
28	2-Methyl buta-1,3-diene	0.648	2.207	0.903	0	0.082	307.2	305.29	1.91
29	Pentane	0.156	2.204	0.806	0.001	0.054	309.2	312.31	-3.11
30*	E-pent-2-ene	0.337	2.196	0.918	0.002	0.063	309.5	308.16	1.34
31	Z-pent-2-ene	0.337	2.196	0.795	0.001	0.06	310.1	306.19	3.91
32	2-Methyl but-2-ene	0.287	2.191	1.36	0	0.063	311.7	313.41	-1.71
33	Pent-1-yne	0.583	2.21	0.419	0	0.098	313.3	309.6	3.7
34	3-Methyl buta-1,2-diene	0.523	2.194	1.453	0	0.086	314	313.86	0.14

ISSN: 0975-8585

35*	3,3-Dimethyl but-1-ene	0.944	2.417	0.862	0	0.069	314.4	322.72	-8.32
36	E-penta-1,3-diene	0.25	2.202	0.748	0	0.083	315.2	318.25	-3.05
37	cyclopentene	0.181	2.234	0	0	0.07	317.4	315.1	2.3
38	Penta-1,2-diene	0.455	2.202	0.635	0.001	0.083	318	309.33	8.67
39	Cyclopentane	0	2.236	0	0	0.049	322.4	314.88	7.52
40*	2,2-Dimethylbutane	0.708	2.411	0.914	0	0.051	322.9	324.57	-1.67
41*	2,3-Dimethyl but-1-ene	0.676	2.411	1.385	0	0.068	328.8	335.57	-6.77
42	Z-4-Methyl pent-2-ene	0.616	2.407	0.739	0.002	0.058	329.6	327.97	1.63
43	2,3-Dimethylbutane	0.481	2.412	1.295	0	0.052	331.1	336.84	-5.74
44*	E-4-Methyl pent-2-ene	0.616	2.407	0.689	0.002	0.065	331.7	329.79	1.91
45	Hexa-1,5-diene	0.441	2.426	0.732	0.001	0.057	332.6	337.28	-4.68

		Descriptors					Boiling Point		
N	Sample	MAXDN	VEv1	HATS5u	H6m	R1p+	Expt. Bp	Calc. Bp	Residual
46	2-Methylpentane	0.435	2.412	0.762	0.001	0.054	333.4	334.51	-1.11
47*	3- Methylpentane	0.398	2.409	1.226	0.001	0.054	336.4	339.14	-2.74
48	Hex-1-ene	0.323	2.418	0.721	0.001	0.062	336.6	341.85	-5.25
49	Z-Hex-3-ene	0.326	2.403	0.637	0.004	0.052	339.6	334.66	4.94
50*	E-Hex-3-ene	0.326	2.403	0.613	0.003	0.056	340.3	335.91	4.39
51	2-Methyl pent-2-ene	0.331	2.401	0.858	0.002	0.061	340.5	339.59	0.91
52*	Z-3-Methyl pent-2-ene	0.309	2.4	1.429	0.001	0.051	340.9	341.69	-0.79
53	E-Hex-2-ene	0.267	2.408	0.635	0	0.06	341	341.32	-0.32
54	Hexane	0.145	2.412	0.712	0.003	0.047	341.9	341.92	-0.02
55*	Z-Hex-2-ene	0.267	2.408	0.661	0.003	0.056	342	339.44	2.56
56	E-3-Methyl pent-2-ene	0.309	2.4	1.368	0.001	0.058	343.6	343.74	-0.14
57	Methylcyclopentane	0.287	2.431	0.558	0	0.053	344.9	341.16	3.74
58	2,3-Dimethyl but-2-ene	0.241	2.4	1.74	0	0.051	346.4	347.21	-0.81
59	1-Methylcyclopentene	0.175	2.423	0.83	0	0.057	348.95	347.7	1.25
60	2,3,3-Trimethyl but-1-ene	0.944	2.607	1.423	0	0.052	351	351.62	-0.62
61	2,2-Dimethylpentane	0.7	2.604	0.654	0.001	0.048	352.3	352.08	0.22
62	2,4-Dimethylpentane	0.458	2.607	0.671	0.004	0.048	353.6	360.76	-7.16
63*	Cyclohexane	0	2.449	0.554	0	0.046	353.9	352.03	1.87
64	2,2,3-Trimethylbutane	0.75	2.607	1.376	0	0.047	354	356.67	-2.67
65	Cyclohexene	0.181	2.448	0.515	0	0.063	356.1	350.84	5.26
66	3,3-Dimethylpentane	0.667	2.601	1.291	0	0.045	359.2	357.39	1.81
67*	1,1Dimethylcyclopentane	0.556	2.62	0.812	0	0.053	361	363.05	-2.05
68	2,3-Dimethylpentane	0.468	2.603	1.282	0.001	0.049	362.9	366.12	-3.22
69	2-Methylhexane	0.431	2.605	0.594	0.004	0.052	363.2	362.34	0.86
70	E-1,2-dimethylcyclopentane	0.319	2.616	1.055	0	0.048	365	371.67	-6.67
71*	3-Methylhexane	0.384	2.601	0.964	0.002	0.051	365	366.52	-1.52
72*	3-Ethylpentane	0.347	2.598	1.241	0.003	0.052	366.6	369.85	-3.25
73	Hept-1-ene	0.312	2.61	0.701	0.003	0.061	366.8	371.76	-4.96
74	Heptane	0.139	2.604	0.642	0.004	0.046	371.6	371.31	0.29
75*	2,2,4-Trimethylpentane	0.728	2.788	0.589	0.005	0.044	372.4	377.25	-4.85
76	Z-1,2-Dimethylcyclopentane	0.319	2.616	1.166	0	0.051	372.7	373.64	-0.94
77	Methylcyclohexane	0.297	2.629	0.776	0	0.048	374.1	372.19	1.91
78	Ethylcyclopentane	0.236	2.614	0.798	0.001	0.055	376.6	374.43	2.17
79	1,1,3-Trimethylcyclopentane	0.584	2.792	0.789	0.001	0.046	378	386.53	-8.53
80	1-Ethylcyclopentene	0.222	2.605	0.943	0.001	0.058	379.45	375.78	3.67
81*	2,2,3,3-Tetramethylbutane	0.813	2.789	1.564	0	0.037	379.6	381.06	-1.46

ISSN: 0975-8585

82	2,2-Dimethylhexane	0.698	2.784	0.464	0.006	0.048	380	377.81	2.19
83	2,5-Dimethylhexane	0.447	2.788	0.525	0.011	0.043	382.3	385.32	-3.02
84*	2,4-Dimethylhexane	0.454	2.784	0.778	0.005	0.047	382.6	389.37	-6.77
85*	2,2,3-Trimethylpentane	0.741	2.784	1.269	0.001	0.045	383	382.97	0.03
86	3,3-Dimethylhexane	0.658	2.779	0.997	0.002	0.044	385.1	382.56	2.54
87*	2,3,4-Trimethylpentane	0.491	2.785	1.246	0.002	0.048	386.6	392.98	-6.38
88	1,1,2-Trimethylcyclopentane	0.597	2.795	1.386	0	0.048	386.9	392.29	-5.39
89	2,3,3-Trimethylpentane	0.708	2.783	1.536	0	0.042	387.9	385.67	2.23
90	2,3-Dimethylhexane	0.463	2.782	1.007	0.004	0.045	388.8	390.19	-1.39

		Descriptors					Boiling Point		
N	Sample	MAXDN	VEv1	HATS5	H6m	R1p+	Expt. Bp	Calc.	Residual
91*	3-Ethyl-2-methylpentane	0.454	2.779	1.221	0.005	0.046	388.8	391.96	-3.16
92*	2-Methylheptane	0.429	2.785	0.511	0.007	0.05	390.8	388.58	2.22
93	3,4-Dimethylhexane	0.417	2.78	1.273	0.003	0.044	390.9	393.71	-2.81
94*	4-Methylheptane	0.37	2.779	0.791	0.004	0.049	390.9	392.72	-1.82
95	3-Ethyl-3-methylpentane	0.625	2.777	1.479	0.001	0.051	391.4	390.17	1.23
96*	Cycloheptane	0	2.646	0.902	0	0.046	391.6	386.23	5.37
97	3-Ethylhexane	0.333	2.776	0.949	0.004	0.051	391.7	395.63	-3.93
98*	3-Methylheptane	0.37	2.779	0.791	0.004	0.049	392.1	392.72	-0.62
99	E-1,4-Dimethylcyclohexane	0.314	2.797	0.865	0.001	0.041	392.5	396.22	-3.72
100*	1,1-Dimethylcyclohexane	0.571	2.804	0.943	0	0.047	392.7	390.6	2.1
101*	Z-1,3-Dimethylcyclohexane	0.321	2.802	0.877	0	0.048	393.3	399.33	-6.03
102*	Oct-1-ene	0.306	2.788	0.641	0.005	0.06	394.4	398.62	-4.22
103	1-Ethyl-1-methylcyclopentane	0.514	2.791	1.174	0.001	0.055	394.7	395.29	-0.59
104	2,2,4,4-Tetramethylpentane	0.766	2.96	0.547	0.009	0.039	395.4	400	-4.6
105	E-1,2-Dimethylcyclohexane	0.33	2.802	1.142	0	0.045	396.6	400.36	-3.76
106	2,2,5-Trimethylhexane	0.718	2.958	0.397	0.015	0.043	397.2	400.14	-2.94
107*	Z-1,4-Dimethylcyclohexane	0.314	2.797	0.865	0.001	0.041	397.5	396.22	1.28
108	E-1,3-Dimethylcyclohexane	0.321	2.802	0.963	0.001	0.047	397.6	399.64	-2.04
109	E-Oct-2-ene	0.232	2.784	0.643	0.005	0.048	398.1	396.78	1.32
110*	Octane	0.135	2.782	0.582	0.006	0.043	398.8	397.41	1.39
111	Isopropylcyclopentane	0.391	2.795	0.927	0.001	0.048	399.6	395.99	3.61
112*	2,2,4-Trimethylhexane	0.727	2.955	0.617	0.009	0.046	399.7	403.65	-3.95
113	Z-1,2-Dimethylcyclohexane	0.33	2.802	1.172	0	0.048	402.9	401.47	1.43
114	Propylcyclopentane	0.222	2.789	0.622	0.003	0.051	404.1	399.27	4.83
115*	Ethylcyclohexane	0.247	2.797	0.913	0.001	0.049	404.9	401.85	3.05
116	2,2-Dimethylheptane	0.699	2.954	0.375	0.01	0.046	405.8	402.25	3.55
117	2,2,3,4-Tetramethylpentane	0.77	2.956	1.177	0.004	0.042	406.2	406.74	-0.54
118*	2,2,3-Trimethylhexane	0.74	2.951	0.986	0.005	0.046	406.8	406.59	0.21
119	2,2,5,5-Tetramethylhexane	0.743	3.122	0.337	0.026	0.035	410.6	419.05	-8.45
120	2,2,3,3-Tetramethylpentane	0.804	2.954	1.564	0.001	0.043	413.4	409.49	3.91
121	1,E-3,5-Trimethylcyclohexane	0.344	2.967	0.845	0.001	0.046	413.7	423.53	-9.83
122	2,3,3,4-Tetramethylpentane	0.75	2.954	1.561	0.002	0.038	414.7	409.5	5.2
123	2-Methyloctane	0.429	2.954	0.459	0.008	0.048	416.4	414.1	2.3
124	3,3-Diethylpentane	0.583	2.943	1.441	0.004	0.049	419.3	416.23	3.07
125	Non-1-ene	0.302	2.956	0.592	0.007	0.056	420	423.22	-3.22
126	Cyclooctane	0	2.828	0.933	0	0.044	422	414.57	7.43
127	Nonane	0.133	2.951	0.536	0.007	0.041	424	422.88	1.12

ISSN: 0975-8585

128	Isopropylcyclohexane	0.398	2.965	0.947	0.003	0.044	427.7	420.91	6.79
129	3,3,5-Trimethylheptane	0.685	3.111	0.82	0.011	0.046	428.8	431.09	-2.29
130	Propylcyclohexane	0.233	2.959	0.72	0.002	0.048	429.9	425.77	4.13
131	2,2,3,3-Tetramethylhexane	0.803	3.11	1.224	0.006	0.043	433.5	430.13	3.37
132	Deca-1,3-diene	0.286	3.114	0.514	0.012	0.069	442	451.41	-9.41
133	Dec-1-ene	0.299	3.115	0.548	0.008	0.054	443.7	447.17	-3.47
134	Isobutylcyclohexane	0.414	3.122	0.602	0.007	0.045	444.5	441.62	2.88
135*	tert-Butylcyclohexane	0.68	3.127	0.949	0.005	0.042	444.7	434.94	9.76

		Descriptors					Boiling Point		
N	Sample	MAXDN	VEv1	HATS5	H6m	R1p+	Expt. Bp	Calc. Bp	Residual
136	Decane	0.131	3.11	0.502	0.008	0.04	447.3	447.38	-0.08
137	sec-Butylcyclohexane	0.347	3.119	0.947	0.011	0.045	452.5	445.57	6.93
138	Butylcyclohexane	0.228	3.116	0.586	0.005	0.048	454.1	448.92	5.18
139	Undec-1-ene	0.298	3.266	0.512	0.01	0.051	465.8	469.26	-3.46
140	Undecane	0.13	3.261	0.471	0.01	0.038	469.1	469.86	-0.76
141	Hexylcyclopentane	0.216	3.268	0.404	0.006	0.052	476.3	473	3.3
142	Dodec-1-ene	0.297	3.41	0.481	0.013	0.049	486.5	490.39	-3.89
143	Dodecane	0.129	3.406	0.448	0.012	0.037	489.5	491.8	-2.3
144	Heptylcyclopentane	0.216	3.414	0.375	0.008	0.05	497.3	494.66	2.64
145	Tridec-1-ene	0.296	3.549	0.455	0.021	0.045	505.9	508.68	-2.78
146	Tridecane	0.128	3.544	0.425	0.022	0.035	508.6	510.38	-1.78
147	Octylcyclopentane	0.217	3.553	0.351	0.011	0.051	516.9	516.02	0.88
148	Tetradec-1-ene	0.295	3.682	0.433	0.035	0.044	524.3	525.85	-1.55
149	Tetradecane	0.128	3.678	0.407	0.036	0.034	526.7	527.5	-0.8
150	Nonylcyclopentane	0.217	3.687	0.341	0.017	0.048	535.3	534.65	0.65
151	Pentadec-1-ene	0.295	3.811	0.412	0.051	0.041	541.5	541.2	0.3
152	Pentadecane	0.128	3.807	0.389	0.051	0.032	543.8	543.68	0.12
153	Decylcyclopentane	0.218	3.816	0.324	0.021	0.048	552.5	553.9	-1.4
154	Hexadec-1-ene	0.294	3.935	0.395	0.067	0.04	558	556.76	1.24
155	Hexadecane	0.127	3.932	0.375	0.066	0.031	560	559.43	0.57
156*	Decylcyclohexane	0.228	3.94	0.383	0.031	0.046	570.8	570.32	0.48
157	Heptadecane	0.127	4.052	0.36	0.082	0.03	575.2	574.11	1.09
158*	Dodecylcyclopentane	0.218	4.062	0.303	0.047	0.048	584.1	586.16	-2.06
159	Octadec-1-ene	0.294	4.173	0.364	0.099	0.036	588	585.1	2.9
160	Octadecane	0.127	4.17	0.348	0.096	0.029	589.5	588.97	0.53
161	Tridecylcyclopentane	0.219	4.179	0.3	0.063	0.046	598.6	600.32	-1.72
162*	1-Cyclopentyltetradecane	0.219	4.294	0.286	0.08	0.046	599	614.06	-15.06
163	Nonadecane	0.127	4.284	0.336	0.111	0.028	603.1	602.96	0.14
164	Eicosane	0.126	4.395	0.325	0.125	0.026	617	616.42	0.58
165	1-Cyclopentylpentadecane	0.22	4.405	0.284	0.094	0.045	625	628.12	-3.12

RESULTS AND DISCUSSION

Results of the MLR Model

A multiple linear regression (MLR) was employed to describe the relation between critical properties and their molecular descriptors. The best model and the number of descriptors (p) in the final QSPR model was determined on the basis of the correlation coefficient R^2. At first, the optimal p is tested using p= 2 to 8. An increase of the R^2 value less than 0.02 was chosen as a threshold. Figure. 1 shows the application of the

breaking point criterion [22] in the present case suggest a best five-parameters equation was obtained, which is as the following:

$$Bp = -56.17 + 157.85\ \mathbf{VEv1} - 34.74\ \mathbf{MAXDN} + 7.62\ \mathbf{HATS5u} - 231.67\ \mathbf{H6m} + 360.75\ \mathbf{R1p+} \quad (10)$$

$R^2 = 99.77\%$ $R^2_{adj} = 99.80\%$ $Q^2_{LOO} = 99.73\%$ $Q^2_{EXT} = 99.57\%$ $Q^2_{BOOT} = 99.61\%$ $s = 4.79$

$F = 10423.55$ $K_{xx} = 41$ $K_{xy} = 50.27$

Here, VEV1 is the eigenvector coefficient sum from van der Waals weighted distance matrix; MAXDN is the maximal electrotopological negative variation [23,24] ; HATS5u is the leverage-weighted autocorrelation of lag 5 / unweighted [25,26] ; H6m is the H autocorrelation of lag 6 / weighted by atomic masses ; R1p+ is the R maximal autocorrelation of lag 1 / weighted by atomic polarizabilities [25,26]

More information about these descriptors can be found in [27] and the references therein.

The results for the randomized models can be compared with the real starting one only by representing in a plot the statistical coefficients R^2 and Q^2. This is depicted in figure. 2. The statistics for the modified Bp vectors are clearly lower than the real QSPR model. This ensures that a real structure-property relationship has been found out.

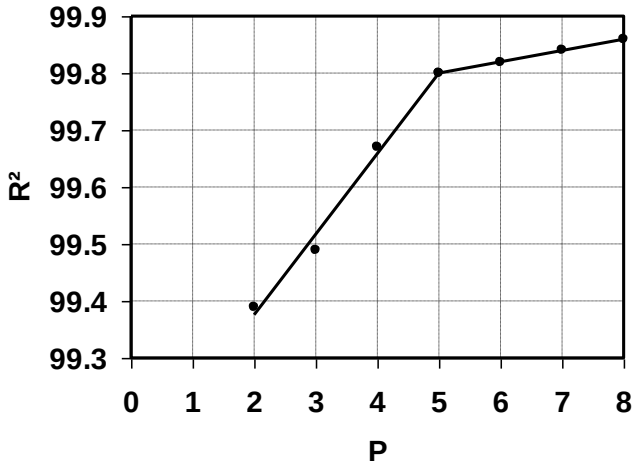

Figure 1: Breaking point rule for determination of the optimum number of the descriptors

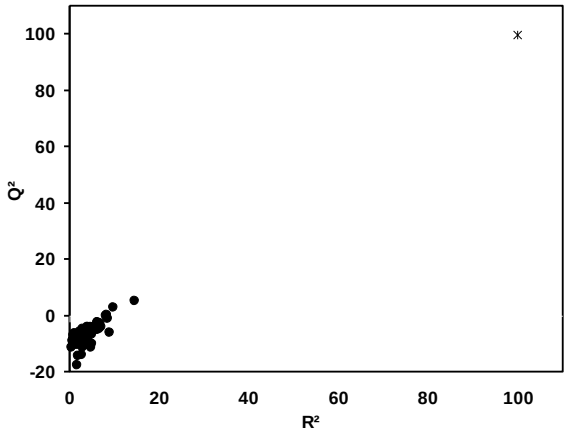

Figure 2: Randomization test associated to previous QSPR model. Black circles represent the randomly ordered, and star corresponds to the real boiling points.

ISSN: 0975-8585

Some important statistical parameters (as given in table 2) were used to evaluate the involved descriptors. The t-value of a descriptor measures the statistical significance of the regression coefficients. The high absolute t-values shown in table 2 express that the regression coefficients of the descriptors involved in the MLR model are significantly larger than the standard deviation. The t-probability of a descriptor can describe the statistical significance when combined together within an overall collective QSPR model (i. e., descriptor's interactions). Descriptors with t-probability values below 0.05 (95% confidence) are usually considered statistically significant in a particular model, which means that their influence on the response variable is not merely by chance [28]. The smaller t-probability suggests the more significant descriptor. The t-probability values of the five descriptors are very small, indicating that all of them are highly significant descriptors. The VIF values suggest that these descriptors are weakly correlated with each others. Thus, the model can be regarded as an optimal regression equation.

For the training and test set are showed in table 1 and figure. 2. Regression lines were used for comparing the values obtained by this model with experimental values. As can be seen from figure. 3, the calculated slope and intercept (a=0.998 ; b=0.88) did not differ greatly from the "ideal" values of 1 and 0, respectively, and most of the predicted Bp values agreed, for all the training and testing sets. Thus, model has been developed that calculate the Bp values for hydrocarbons with accuracy comparable to experiment.

The distribution of errors for the entire data set is given in figure. 4. Residuals are distributed normally around zero (the mean value) as can be clearly seen from the histogram in the right side of the plot,

Table 2: Characteristics of the selected descriptors in the best MLR model

Descriptor	Descriptor type	X	Dx	t- value	t- probability	VIF
Constant		-56.17	3.96	-14.20	0	
VEv1	Eigenvalue-based indices	157.85	1.13	140.32	0	3.03
MAXDN	Topological descriptors	-34.74	2.21	-15.72	0	1.24
HATS5u	GETAWAY descriptors	7.62	1.13	6.75	0	1.30
H6m	GETAWAY descriptors	-231.67	30.08	-7.70	0	2.86
R1p+	GETAWAY descriptors	360.75	35.64	10.12	0	1.50

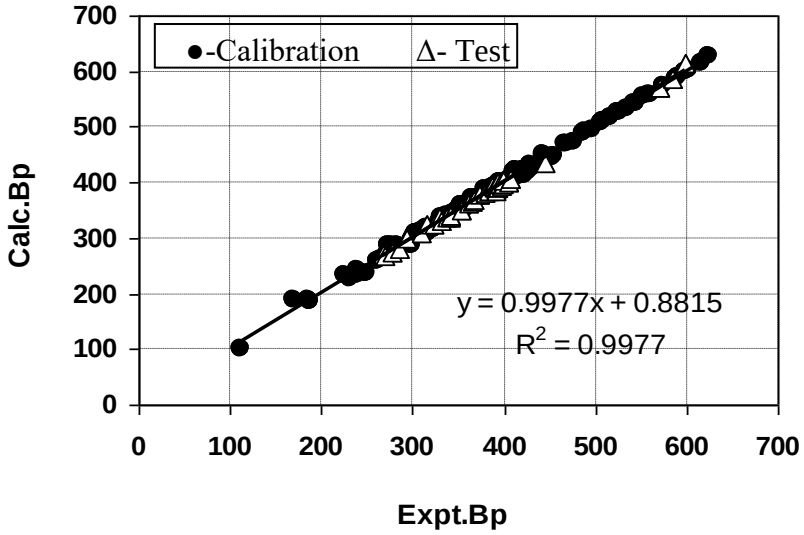

Figure 3: Plot of predicted *vs*. experimental Bp for the entire data set.

ISSN: 0975-8585

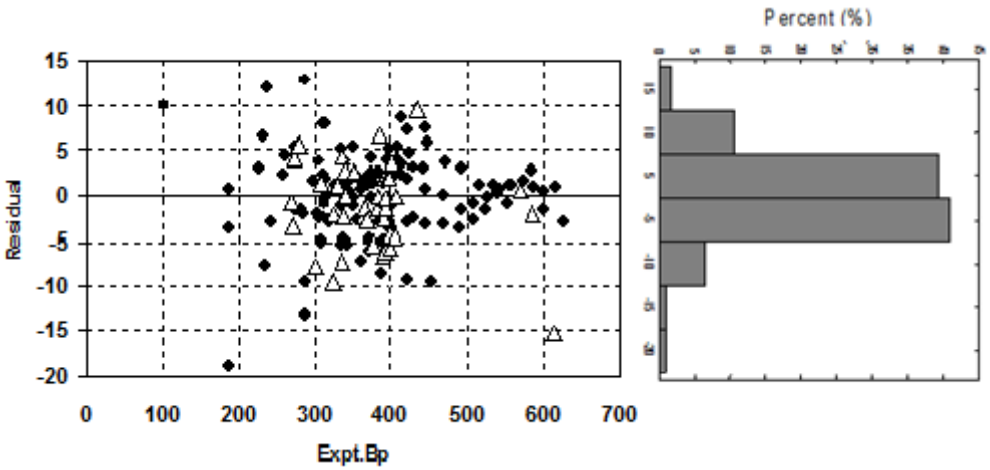

Figure 4: Plot of residual *vs*. experimental Bp for the entire data set.

Descriptor Contribution Analysis and Interpretation

Based on a previously described procedure [29, 30], the relative contribution of the five descriptors to the model were determined and they decrease in the following order: VEv1(67.01%) > MAXDN(11.36%) > HATS5u (07.50%) > H6m (07.29%) > R1p+ (06.84%) . It should be noted that the difference in the descriptor contribution between the three last descriptors used in the model is not significant, but the first one had a very high contribution indicating that these descriptor is indispensable in generating the predictive model (Figure.5).

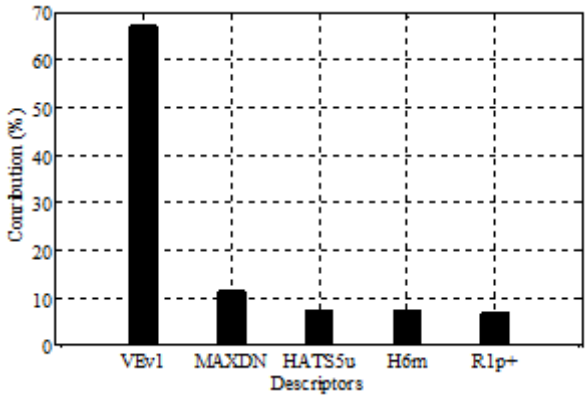

Figure 5: Relative contributions of the selected descriptors to the MLR model.

The first important descriptor is VEv1, which has a relatively very high positive correlation with the experimental Bp values (R= 99.95%). The positive coefficient of VEv1 indicates that the hydrocarbons with larger values for this descriptor would have higher Bp values.

The second important descriptor is MAXDN, a topological descriptor, which has a smaller negative correlation coefficient with the experimental Bp values (R= -10%). The electrotopological state indices are atomic indices calculated from a H-depleted molecular graph as:

$$S_i = I_i + \Delta I_i = I_i + \sum_{j=1}^{A} \frac{I_i - I_j}{(d_{ij}+1)^k} \qquad (11)$$

where I_i is the intrinsic state of the *i*th atom and ΔI_i is the field effect on the *i*th atom calculated as perturbation of the intrinsic state of *i*th atom by all other atoms in the molecule, the MAXDN is calculated as the maximum negative value of ΔI_i in the molecule; d_{ij} is the topological distance between the *i*th and the *j*th

atoms; A is the number of non-hydrogen atoms in the molecule. The exponent *k* is a parameter to modify the influence of distant or nearby atoms for particular studies. In DRAGON it is taken as *k* = 2.

The last three descriptors are HATS5u, H6m and R1p+, there are a GETAWAY descriptors and correlates with the experimental Bp values of -5.40 (p=0.5), 74 and -54.4% respectively. The GETAWAY descriptors [25,26] have been proposed as chemical structure descriptors derived from a new representation of molecular structure, the molecular influence matrix. These descriptors, as based on spatial autocorrelation, encode information on molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

HATS5u, H6m and R1p+ are calculated by Eq. (12), (13) and (14) respectively.

$$HATSkw = \sum_{i=1}^{A}\sum_{j>1}^{A}(w_i.h_i)(w_j.h_j).\delta(k;d_{ij}) \qquad \text{for k=1,2,3,...D} \qquad (12)$$

$$RTw+ = \max_{ij}\left(\frac{\sqrt{h_{ii}-h_{jj}}}{r_{ij}}.w_i.w_j.\delta(k;d_{ij})\right) \qquad i \neq j \quad \text{and} \quad k=1,2,3,\ldots,D \qquad (13)$$

$$Hkw = \sum_{i=1}^{A}\sum_{j>1}^{A}h_j.w_i.w_j.\delta(k;d_{ij};h_{ij}) \qquad \text{for k=1,2,3,...D} \qquad (14)$$

where *A* is the number of atoms, w is an atomic weighting scheme, d_{ij} is the topological distance, $\delta(k, d_{ij})$ is a Dirac- delta function ($\delta=1$ *if* $d_{ij}=k$, zero otherwise), r_{ij} is the interatomic distance. D is the molecule topological diameter that is the maximum topological distance in the molecule. The coefficient of R1p+ is positive, meaning that the hydrocarbons with larger values for this descriptor have larger Bp values.

The following statistical parameters obtained for the external tests set verify the well-accepted conditions (8-a to 8-d), which reinforces the predictive capabilities of the present model.

$$Q^2_{EXT} = 0.9971 > 0.5 \qquad r^2 = 0.996 > 0.6$$
$$(r^2 - r_0^2)/r^2 = (0.996 - 0.9997)/0.996 = -0.004 < 0.1$$
$$\text{or} \quad (r^2 - r_0'^2)/r^2 = (0.996 - 0.9997)/0.996 = -0.004 < 0.1$$

$$0.85 < k = 0.9965 < 1.15 \quad \text{or} \quad 0.85 < k' = 1.003 < 1.15$$

Applicability Domain of the MLR Model

Before a QSPR model is put into use for screening compounds, its applicability domain must be defined and predictions for only those compounds that fall in this domain can be considered as reliable.

The AD of the MLR model was analyzed in the Williams plot (shown in figure.6). There are three X outliers (Compounds 1, 64 and 65) with leverage higher than the warning limit of 0.14 is a structurally influential compound, and one Y outlier with residual higher than ± 3 (Compound 66) in the training set. Deleting these observations could alter slightly R^2 between the experimental Bp values and the selected descriptors to 99.75% (Q^2 = 99.71%) and decrease the standard error to 4.58, while utilization of a higer energy conformation geometry for this observation alter negatively the calculated model.

Validation

In order to estimate the predictive power of MLR, in this case we used two validation procedures. Firstly, using the leave-one-out procedure; a Q^2_{LOO} =99.70% and the bootstrap procedure a Q^2_{BOOT} =99.67%, reveal the high predictive ability of the model. Secondly the external validation procedure; by using a set of 40 compounds which have not been explored for training set. The external predictive power is confirmed by a

high Q^2_{ext} value (Q^2_{ext}=99.70%) that reveals model applicability also to predict the boiling points of unknown series compounds. The plot of predicted versus experimental values for data set is shown in figure. 3(Δ).

Remains to be noted that there is a single Y (Compound 162) outlier with residual higher than ± 3

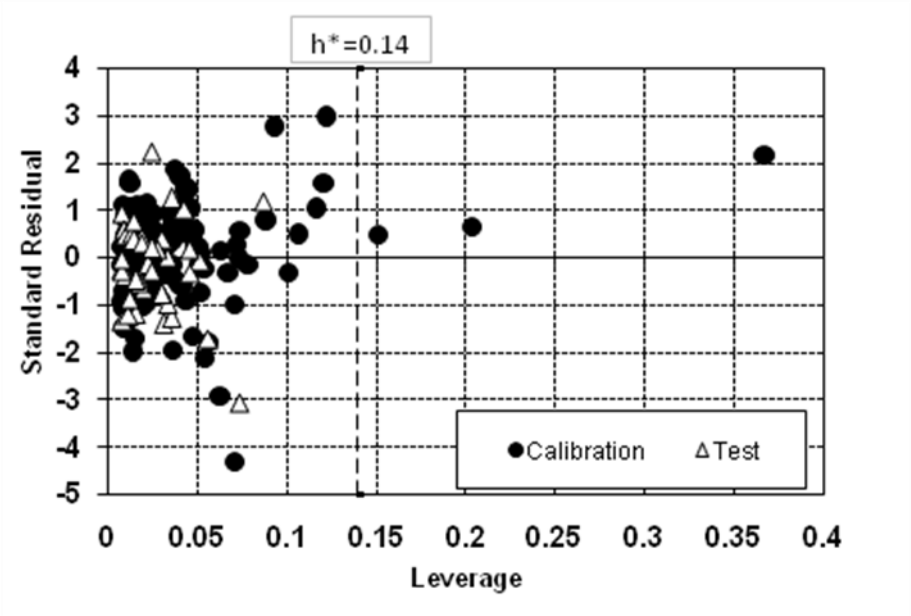

Figure 6: Williams plot of the MLR model for the entire data set.

CONCLUSION

In this paper, the QSPR method was applied to the prediction of the boiling points of organic compounds. A five-parameter linear model was developed by hybrid GA/ MLR approach with R^2 of 99.80 and s of 4.67 for the training set. The selected descriptors express many factors influencing boiling points, to name: molecular size and shape, specific atomic properties. Several validation techniques, including leave-one-out cross-validation and bootstrap, randomization tests, and validation through the test set, illustrated the reliability of the proposed model. All of the descriptors can be directly calculated from the molecular structure of the compound, thus the proposed model is predictive and could be used to estimate the boiling points of hydrocarbons. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

REFERENCES

[1] F. Gharagheizi, S. A. Mirkhani, P. Ilani-Kashkouli, et al, Fluid Phase Equil., 2013, (354), 250.
[2] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Chemical Reviews., 2010, (110), 5714.
[3] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Chemical Reviews., 2012, (112), 2889.
[4] W.J. Lyman, W.F. Reechl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, American Chemical Society, Washington, DC, 1990.
[5] D. Yi-min, Z. Zhi-ping, C. Zhong, Z. Yue-fei, Z. Ju-lan, and L. Xun, J. Mol. Graphics Modell., 2013, (44), 113.
[6] Katritzky A. R., Fara D. C., Energy Fuel., 2005, (19), 922.
[7] Todeschini R., Consonni V., Mauri A., Pavan M., 2005. DRAGON Software – version 5.4-TALETEsrl
[8] R.C. Reid, J.M. Prausnitz, B.E. Poling, The Properties of Gases & liquids,Fourth Edition, Mc Graw-Hill Book Company, New York, 1987.
[9] Hyperchem™. Release 6.02 for windows. 2000. Molecular Modeling system
[10] Snee R D., *Technometrics*, 1977, (19), 415.
[11] See Graybill, 1976/ Graybill, F. A. Theory and Application of the Linear Model, Duxbury, North Scituate, Mass. pp. 231-236.

[12] Todeschni R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).
[13] Leardi R., Boggia R., Tarrile M., 1992. *J. Chemom*, 1992, (6), 267.
[14] Xu J., Zhang H., Wang Lei., Liang G., Wang Luoxin., Shen X., Xu W., *Spectrochimica Acta Part A*, 2010, (76), 239.
[15] Todeschini R., Maiocchi A., Consonni V., The K Correlation Index: Theory Development and its Application in Chemometrics. *Chemom,* Int. Lab. Syst, 1999, (46), 13.
[16] Golbraikh A, Tropsha A. J Comput Aided Mol Des 2002; 16: 357-369.
[17] Golbraikh A, Tropsha A. J Mol Graph Model 2002; 20: 269-276.
[18] Tropsha A., Gramatica P., Gombar V K., *QSAR Comb. Sc*i, 2003, (22), 69.
[19] Shen M., Béguin C., Golbraikh A., Stables J P., Kohn H., Tropsha A., *J. Med. Chem*, 2004, (47), 2356.
[20] Weisberg S., 2005. Applied Linear Regression, 3rd edn. (John Wiley and sons, Inc., New Jersey,)
[21] SCAN- Software for Chemometric Analysis- 1995. version 1.1- for Windows, Minitab USA.
[22] Katritzky, A. R., Dobchev, D. A., Tulp, I., Karelson, M., Carlson, D. A. Med. Chem. Lett. 2006, (16), 2306.
[23] A.T. Balaban, D. Ciubotariu, M. Medeleanu, J.Chem.Inf.Comput.Sci. 1991, (31), 517.
[24] P.Gramatica, Corradi M., Consonni V., Chemosphere 2000, (41), 763.
[25] Consonni V., Todeschini R., Pavan M., *J. Chem. Inf. Comput. Sci*, 2002, (42), 682.
[26] Consonni V., Todeschini R., Pavan M., Gramatica P., *J. Chem. Inf. Comput. Sci*, 2002, (42), 693.
[27] Todeschini R., Consonni V. , 2009. Molecular Descriptors for Chemoinformatics Volumes I & II. (WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009).
[28] Ramsey F. L., Schafer D. W., 2002. The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd edn. (Wadsworth group, USA).
[29] Zheng F., Bayram E., Sumithran S P., Ayers J T., Zhen C G., Schmitt J D., Dwoskim L P., Crooks P A., Bioorg. Med. Chem, 2006, (14), 3017.
[30] Guha R., Jurs P C., *J. Chem. Inf. Model*, 2005, (45), 800.