Iranian J. Math. Chem. 8(2) June (2017) 199−220

Iranian Journal of Mathematical Chemistry

Journal homepage: ijmc.kashanu.ac.ir

# Determination of Critical Properties of Alkanes Derivatives using Multiple Linear Regression

**ESMAT MOHAMMADINASAB**•

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

**ARTICLE INFO**

**Article History:**
Received: 27 July 2016
Accepted: 14 October 2016
Published online: April 11 2017
Academic Editor: Roberto Todeschini

**Keywords:**
Alkanes
MLR
Critical properties
QSPR

**ABSTRACT**

This study presents some mathematical methods for estimating the critical properties of 40 different types of alkanes and their derivatives including critical temperature, critical pressure and critical volume. This algorithm used QSPR modeling based on graph theory, several structural indices, and geometric descriptors of chemical compounds. Multiple linear regression was used to estimate the correlation between these critical properties and molecular descriptors using proper coefficients. To achieve this aim, the most appropriate molecular descriptors were chosen from among 11 structural and geometric descriptors in order to determine the critical properties of the intended molecules. The results showed that among all the proposed models to predict critical temperature, pressure and volume, a model including the combination of such descriptors as hyper-Wiener, Platt, MinZL is the most appropriate one.



## 1. INTRODUCTION

Alkanes are considered as one of the most significant aliphatic hydrocarbons used in such various industries as food, pharmaceutical, petrochemical and oil [1]. Since these chemicals are present in many refining processes (crude oil), it is quite essential to take their various physical and chemical properties– especially critical properties– into consideration. Critical point is a state in which there is no boundary between the two phases of a matter. This state occurs for every matter at the presence of a certain amount of temperature, pressure and combination. Theoretically, it is possible to estimate most of the thermodynamic properties

• Corresponding Author: (Email address: e-mohammadinasab@iau-arak.ac.ir)
DOI: 10.22052/ijmc.2017.58461.1225

of chemicals using their critical properties. Practically, a large amount of theoretical correlations are based on these properties [2–4].

The initial methods for estimating critical properties were experimental and were used for hydrocarbon systems. Due to the fact that experimental values of critical properties are not available for heavy alkanes, it is important to take advantage of appropriate methods for estimating these properties. In this study, the model of multiple linear regression was applied for the first time to find out the most appropriate molecular descriptors in order to estimate critical temperature, critical pressure and critical volume of alkanes and their derivatives [5]. The independent variables in multiple regression model can be obtained through the use of various methods [6–7]. In addition, the graph theory provides us with a suitable means for calculating topological descriptors which function as independent variables [8–9].

The graph theory has a long history in mathematics and its application dates back to about 200 years ago. In 1730, Euler was recognized as the father of graph theory after publishing the "Seven Bridges of Königsberg". This theory is one the most practical branches of mathematics in other disciplines. It has a wide range of applications in such disciplines as biology, chemistry, nanotechnology, operational research and engineering [10].

Chemical Graph Theory is one of the branches of mathematical chemistry which is typically related to theoretical chemistry [11]. According to this theory, a graph indicates a set of elements of a group and their interrelationships. In chemical graphs, the hydrogen atoms are ignored since these atoms do not usually play a significant role in determining the molecular structure. After drawing the chemical graph for a molecule, it would not be difficult to extract topological indices–which are some constant numbers–for that graph. Mathematicians call such constant numbers topological indices. These indices include the structure, size, molecular polymerization, number of atoms and types of freaking molecular atoms. The concept topological indices were initially expressed in 1947. Wiener and Platt were the first to develop graph theory–based quantitative topological variables in 1947 known as Wiener index and Platt index, respectively, and reported Quantitative Structure Property Relationship (QSPR) models on boiling points of hydrocarbones. At that time, this concept was most often used for physical properties such as alkanes and paraffin's boiling points. QSPRs have provided a valuable approach in research into physico–chemical properties of organic compounds [12–13]. In 1994, Ivan Gutman paid much more attention to these issues, specifically the distances and weighted functions, in his paper entitled "on the sum of all distances in composite graphs". In theoretical chemistry, these indices help to predict the chemical and pharmaceutical properties of materials. QSPR is a model that relates the predictor variables of a molecule to its physico–chemical properties. The essential problem in the development of a suitable correlation between chemical structures

and properties can be imputed to the quantitative nature of chemical structures. Graph theory was successfully applied through the translation of chemical structures into characteristic numerical descriptors by resorting to graph invariants.

Hosseini and Shafiei proposed QSPR model for the prediction of gas heat capacity of benzene derivatives through the use of topological indices. The best model was obtained as follows: $C_v = -84.569 + 43.970\,{}^1\chi - 2.298W + 1.463Sz$. This means that ${}^1\chi$, W, Sz descriptors play an important role in affecting heat capacity ($C_v$) of benzene derivatives [14].

QSPR modeling produces predictive model derived from application of statistical tools correlating physico–chemical properties in QSPR models of chemicals with descriptors representative of molecular structure [15–17].

In a nutshell, the aim of present research is to investigate the relationship between all critical properties (as dependent variables) and 2–dimentional and 3–dimentional descriptors (as independent variables) using QSPR and multiple linear regression (MLR) methods for estimating the critical properties of 40 different types of alkanes and their derivatives including critical temperature ($T_c$), critical pressure ($P_c$) and critical volume ($V_c$).

## 2. TOPOLOGICAL INDICES

Considering the research studies in which several two–dimensional indices (topological) were used, the current paper makes an attempt to investigate several three–dimensional (geometrical) indices as molecular descriptors and their application for prediction of critical properties of alkanes [16–17]. As a matter of fact, critical properties are sensitive to molecular geometry; hence, some of the geometric descriptors were employed as independent variables in this research. The statistical formulas used in this regard are presented below.

### 2.1 WIENER INDEX

In 1947, Harold Wiener [18] introduced one of the first molecular descriptors of topological nature for acyclic saturated hydrocarbons. The Wiener index W(G) of a graph G is defined as the half of the sum of the distances between every pair of vertices of G, $D_{ij}$, is the distance of two vertices i and j in the graph G).

$$W = \frac{1}{2}\sum_{i=}^{n}\sum_{j=1}^{n} D_{ij} \qquad (1)$$

### 2.2 HYPER WIENER INDEX

Another related distance–based structural descriptor of the graph G is the hyper–Wiener index [19], WW(G) [5]. WW(G) index is introduced as:

$$WW(G) = \sum d(U,V)^2 + \sum d(U,V)/2 \quad (2)$$

where d(U, V) denotes the distance between the vertices U and V in the graph G and the summations run over all (unordered) pairs of vertices of G.

### 2.3 RANDIĆ INDEX

In 1975, Milan Randić a Croatian–American scientist introduced the Randić index [20–23], the first connectivity index. The Randić index of a chemical graph is the sum of all the bonds contributions:

$$\chi = \sum (\frac{1}{d_i d_j})^{\frac{1}{2}} \quad (3)$$

where $d_i$ and $d_j$ are the degrees of the vertices representing atoms "i", "j".

### 2.4 BALABAN INDEX

Defined by the Romanian chemist Alexandru T. Balaban in 1982, Balaban index is: J=J(G) of a Graph G on n node and m edges and $D_i$ are the sum of all entries in the $^i$th row (or column) of graph distance matrix [24–25]:

$$J = \frac{m}{\mu + 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ (D_i)(D_j) \right]^{-0/5} \quad (4)$$

where $\mu = m - n + 1$ is the cyclomatic number.

### 2.5 HARARY INDEX

The Harary index of a graph G was defined from the inverse of the squared elements of the distance matrix according to the expression [26]:

$$\mathrm{H} = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( D_{ij} \right)^{-2} \quad (5)$$

where $D^{-2}$ is the matrix whose elements are the squares of the reciprocal distances.

### 2.6 GEOMETRIC INDICES

Geometric analysis provides characteristic values related to the geometrical structure of a molecule such as minimal and maximal z length, minimal and maximal projection area, force field energies or van der Waals volume.

## 3. COMPUTATIONAL METHODS

To analyze the relationship between critical properties such as temperature ($T_c$), pressure ($P_c$) and volume ($V_c$) of alkanes derivatives in contrast with molecular descriptors, the research data were collected in two stages as follow:

First, the structure and existing laboratory quantities (experimental) of 40 different types of alkanes and their derivatives in the present investigation were taken from National Institute of Standard and Technology chemistry webbook and were listed in Table 1.

Second, the values of Randić ($\chi$), Harary (H), Balaban (J), Wiener (W), Platt (Platt) and hyper–Wiener (WW) topological indices were calculated by formulas 1–5 using graph theory for 40 different types of alkanes derivatives, and the values of geometry descriptors such as the minimal projection area (MinPA/$A^{\circ 2}$), the maximal projection area (MaxPA/$A^{\circ 2}$), the minimal z length (MinZL/$A^{\circ}$), the maximal z length (MaxZL/$A^{\circ}$), the van der Waals volume (V/$A^{\circ 3}$) were taken for 40 compounds of mentioned training set from the book and webbook [27].

Third, the relationships between $T_c$, $P_c$, $V_c$ with all the used molecular indices were investigated for 40 different types of alkanes derivatives using excel software and relevant equations were extracted.

Fourth, the estimation of critical properties was performed by SPSS software version 16 with MLR method and backward procedure. According to the important determining factors of this method such as correlation coefficient, square correlation coefficient, adjust square correlation coefficient, Fisher statistics, Durbin Watson, the best topological indices were determined for estimating the properties.

The linear regression model is a statistical means for analyzing the correlation between an independent variable and a dependent variable. Now, if we increase the number of independent variables to more than one, the regression model is called multiple regression model [28]. The simple linear regression equation is indicated as y=a+bx, while the multiple regression equation is introduced as:

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + e \tag{6}$$

where, y: dependent variable; $b_0$: regression constant; $b_1$: regression coefficient for first independent variable $x_1$; $b_k$: regression coefficient for $k^{th}$ independent variable $x_k$; e: the observed amount of error.

The simple regression model is linear since the increase of a descriptor x value causes the increase of dependent variable y (if the coefficient $b_i$ is positive). One of the assumptions behind the multiple regression model is that there is a linear correlation (a straight line) between the independent variables and dependent one. Several independent variables form a model which predicts the amount of dependent variable.

This research paper attempts to scrutinize the correlation between critical properties of alkanes derivatives and molecular descriptors through the use of MLR method.

## 4. RESULTS AND DISCUSSION

The experimental data of $T_c$, $P_c$ and $V_c$ of alkanes and their derivatives were shown in Table 1:

**Table 1.** Used compound, experimental data of critical temperature ($T_c$/K), critical pressure ($P_c$/Pa), critical volume ($V_c$/$m^3$) of alkanes derivatives.

| No | Alkane | $T_c$/K | $P_c \times 10^{-6}$/Pa | $V_c \times 10^4$/m |
|---|---|---|---|---|
| 1 | Ethane | 305.3 | 4.9 | 1.47 |
| 2 | Propane | 369.9 | 4.25 | 2.0 |
| 3 | n-Butane | 425.1 | 3.80 | 2.55 |
| 4 | 2-Methylpropane | 407.7 | 3.65 | 2.59 |
| 5 | n-Pentane | 469.8 | 3.36 | 3.11 |
| 6 | 2-Methylbutane | 461.0 | 3.38 | 3.06 |
| 7 | 2,2-Dimethylpropane | 433.8 | 3.20 | 3.07 |
| 8 | n-Hexane | 507.6 | 3.02 | 3.68 |
| 9 | 2-Methylpentane | 497.8 | 3.03 | 3.68 |
| 10 | 3-Methylpentane | 504.0 | 3.11 | 3.68 |
| 11 | 2,2-Dimethylbutane | 489.0 | 3.10 | 3.58 |
| 12 | 2,2-Dimethylbutane | 500.1 | 3.15 | 3.61 |
| 13 | n-Heptane | 540.0 | 2.74 | 4.28 |
| 14 | 2-Methylhexane | 530.5 | 2.74 | 4.21 |
| 15 | 3-Methylhexane | 535.2 | 2.81 | 4.04 |
| 16 | 2,2-Dimethylpentane | 520.5 | 2.77 | 4.16 |
| 17 | 2,3-Dimethylpentane | 537.3 | 2.91 | 3.93 |
| 18 | 2,4-Dimethylpentane | 519.8 | 2.74 | 4.18 |
| 19 | 3,3-Dimethylpentane | 536.4 | 2.95 | 4.14 |
| 20 | 2,2,3-Trimethylbutane | 531.1 | 2.95 | 3.98 |

**Table 1.** (Continued).

| No | Alkane | $T_c$/K | $P_c \times 10^{-6}$/ Pa | $V_c \times 10^4/m^3$ |
|---|---|---|---|---|
| 21 | 3-ethylpentane | 540.6 | 2.89 | 4.16 |
| 22 | n-octane | 568.9 | 2.49 | 4.92 |
| 23 | 2,2,3-trimethylpentane | 563.5 | 2.73 | 4.36 |
| 24 | 2,3,3-trimethylpentane | 573.5 | 2.82 | 4.55 |
| 25 | 2,2,4-Trimethylpentane | 543.9 | 2.57 | 4.68 |
| 26 | 2,2-Dimethylhexane | 549.8 | 2.53 | 4.78 |
| 27 | 3,3-Dimethylhexane | 562.0 | 2.65 | 4.43 |
| 28 | 3-Methyl-3-ethylpentane | 576.5 | 2.81 | 4.55 |
| 29 | 2,3,4-Trimethylpentane | 566.4 | 2.73 | 4.6 |
| 30 | 2,3-Dimethylhexane | 563.5 | 2.63 | 4.68 |
| 31 | 2-Methyl-3-ethylpentane | 567.1 | 2.70 | 4.42 |
| 32 | 3,4-Dimethylhexane | 568.8 | 2.69 | 4.66 |
| 33 | 2,4-Dimethylhexane | 553.5 | 2.56 | 4.72 |
| 34 | 2,5-Dimethylhexane | 550.0 | 2.49 | 4.82 |
| 35 | 2-Methylheptane | 559.7 | 2.50 | 4.88 |
| 36 | 3-Methylheptane | 563.6 | 2.55 | 4.64 |
| 37 | 4-Methylheptane | 561.7 | 2.54 | 4.76 |
| 38 | 3-Ethylhexane | 565.5 | 2.61 | 4.55 |
| 39 | n-Nonane | 595.0 | 2.30 | 5.55 |
| 40 | n-Decane | 617.8 | 2.11 | 6.24 |

The values of used topological indices of 40 different types of alkanes and their derivatives were calculated by formula 1–5, and the values of the geometry descriptors of all the mentioned compounds were taken from the book and webbook [27].

In the first section, in order to apply simple linear regression method the relationship between critical properties of the used compound with all used indices was investigated using excel software (see equations 7–39).

The following equations indicated the relationship between $T_c$ and the values of molecular indices.

| Number | Equation | $R^2$ |
|---|---|---|
| (7) | $T_C$=0.2974 Platt + 7.0538 | 0.5219 |
| (8) | $T_C$=0.0609 X + 1.9244 | 0.8051 |
| (9) | $T_C$=0.0255J + 2.353 | 0.294 2 |
| (10) | $T_C$=0.3181H + 5.2614 | 0.8193 |
| (11) | $T_C$=2.349 W + 5.4462 | 0.7733 |
| (12) | $T_C$=5.9864 WW + 9.9962 | 0.6133 |
| (13) | $T_C$=2.1433 V+84.807 | 0.7969 |
| (14) | $T_C$=0.2754 Min PA+21.597 | 0.4973 |
| (15) | $T_C$=0.1149 Min Z L+7.0897 | 0.4415 |
| (16) | $T_C$=0.6768 Max PA+29.759 | 0.7547 |
| (17) | $T_C$=0.0113 Max Z L +5.5147 | 0.0718 |

According to equations (7 to 17) and the square correlation coefficients ($R^2$), it can be inferred that there was better correlations between $T_c$ and H>X>V of this type of alkanes, respectively.

Also, research results indicated that the correlation coefficients values of equations (8), (10), and (13) are very close to each other and there is a significant distinction between these values and other regression coefficients. On the other hand, the correlation coefficients of the equations (7), (9), (11), (14), (15), (16), and (17) demonstrate that there is not a strong correlation between Tc and J, Platt, W, WW, MinPA, MinZL, MaxPA, MaxZL descriptors. Consequently, the descriptors MaxZL and J which possess a lower correlation compared with other descriptors were not used for predicting $T_c$ of alkanes using the MLR method.

According to the square correlation coefficient of equations (18–28), the following rank was found among $P_c$ and molecular descriptors: V>X>MaxPA>H:

| Number | Equation | $R^2$ |
|---|---|---|
| (18) | $P_C$= −86352 Platt +4×10$^6$ | 0.6427 |
| (19) | $P_C$= −615873 X +5×10$^6$ | 0.8877 |
| (20) | $P_C$=−645673 J + 5×10$^6$ | 0.4705 |
| (21) | $P_C$=−116735 H + 4×10$^6$ | 0.8562 |
| (22) | $P_C$=−13768 W + 4×10$^6$ | 0.6881 |

| | | |
|---|---|---|
| (23) | $P_C$=−4210.3 WW + $3\times10^6$ | 0.5269 |
| (24) | $P_C$= −17730 V+ $5\times10^6$ | 0.9219 |
| (25) | $P_C$= −81662 MinPA+ $5\times10^6$ | 0.5172 |
| (26) | $P_C$= −198960 MinZL+$5\times10^6$ | 0.6020 |
| (27) | $P_C$= −52914 MaxPA+ $5\times10^6$ | 0.8645 |
| (28) | $P_C$= −370030 MaxZL + $5\times10^6$ | 0.1238 |

Therefore, the descriptors MaxZL and J which showed a weak correlation were ignored and the descriptors X, H, V, MaxPA which had a higher correlation were preserved for prediction of alkanes $P_c$ through the use of MLR method. Also, a linear relationship between $V_c$ and X, V, H, MaxPA of this class alkanes was obtained. In accordance with the equations (29–39) it was observed that the Randic index and Volume geometry descriptor had the highest linear relationship with Vc, (R2 > 0.97).

According to the $R^2$ values of equations (30), (32), (35), and (38) the following rank was found among $V_c$ and efficient molecular descriptors: X>V >H >MaxPA.

Also, the descriptors Platt, X, H, W, WW, V, MinPA, MinZL, MaxPA which illustrated a correlation coefficient above 0.5 were used for prediction of alkanes $V_c$ using the MLR method, and the descriptors J, MaxZL which possessed a fairly weak correlation were removed.

| **Number** | **Equation** | **$R^2$** |
|---|---|---|
| (29) | $V_C$=41667 Platt − 3.9284 | 0.6240 |
| (30) | $V_C$=8563X − 0.3378 | 0.9706 |
| (31) | $V_C$=3704.6 J + 1.3583 | 0.3767 |
| (32) | $V_C$=43537 H − 6.0622 | 0.9345 |
| (33) | $V_C$=320744 W − 77.865 | 0.8780 |
| (34) | $V_C$=835066 WW – 229.55 | 0.7267 |
| (35) | $V_C$=303037 V+ 4.5371 | 0.9700 |
| (36) | $V_C$=35070MinPA+ 12.867 | 0.4912 |
| (37) | $V_C$=18439 MinZL+ 1.8871 | 0.6926 |
| (38) | $V_C$=96265 MaxPA+ 4.1775 | 0.9297 |
| (39) | $V_C$=1757.2 MaxZL+ 5.026 | 0.1058 |

In addition, the autocorrelation of descriptors used in the selected model was tested. If the regression coefficients of the diagrams indicating interrelationship between the independent variables were above 0.9, one of the independent variables was ignored.

Following MLR guidelines, the experimental critical properties, i.e. $T_c$, $P_c$, $V_c$ were selected as the dependent variables and the suitable molecular descriptors– as the independent variables– were chosen on SPSS software and backward procedure.

Then, different models were examined and the best model was defined using correlation coefficient (Pearson's r), determination coefficient, Std. Error of estimate, mean square, the Fischer statistic, sum of squares of residual and specifically Fisher statistic and the associated significance values (see Table 2).

**Table 2.** Property, Equations, R, $R^2$, $R^2_{Adjust}$,. RMSE, F statistic, SS, SSE and Sig for estimating of $T_c$, $P_c$, $V_c$.

| Mod. | Prop. | Equation | R | $R^2$ | $R^2_{Adjust}$ | RMSE | F | SS | SSE |
|---|---|---|---|---|---|---|---|---|---|
| 40 | $T_c$ | $T_c$=8.75 Platt – 0.113 WW+ 19.995 MinZL+ 232.130 | 0.959 | 0.920 | 0.913 | 1.8509 E1 | 137.877 | 141718.56 | 12334.32 |
| 41 | $P_c$ | $P_c$= -68615.237 Platt+1413.541 WW – 194862.228 MinZL+ 5494998.974 | 0.961 | 0.923 | 0.917 | 1.4933 E5 | 144.619 | 9.67E12 | 8.028 E11 |
| 42 | $V_C$ | $V_c$=1.006E-5 Platt +1.931E-7 WW+ 2.123E-5 MinZL +5.526 E-5 | 0.986 | 0.972 | 0.970 | 1.5879 E5 | 417.097 | 3.15E-7 | 9.076 E-9 |

### 4.1 STATISTICAL PARAMETERS

**4.1.1.** Significance Level (sig): A coefficient used in the statistical method is significance level. The more the significance level is close to zero, the smaller the significance level and the linear model will be more meaningful. Therefore, the higher the Fisher statistic, the lesser significance level. As it's seen in Table 2, the best three descriptors, as predictors of $T_c$, $P_c$ and $V_C$ in terms of non–standardized coefficients, are represented using the models (40), (41) and (42), respectively.

**4.1.2.** Correlation Coefficients (R): It indicates the correlation between two variables. Statistically, the higher correlation between variables X and Y, the more accurate the prediction will be. R=0.959 in equation (40) illustrates a strong correlation between $T_c$ and Platt, WW, MinZL descriptors using the MLR method.

**4.1.3.** Determination Coefficient ($R^2$): For example, the value of $R^2$=0.972 in equation (42) illustrates that 97.2% of variation is residing in the residual meaning that the fitted line or model is very good.

**4.1.4**. Adjusted Determination Coefficient ($R^2_{Adjust}$): the percentage of adjusted determination coefficient does not represent the influence of all the independent variables, but it only illustrates the real influence of applied independent variables on the dependent variable. Thus, the high value of $R^2_{Adjust}$ (%97.0) can be used to explain the values of the $V_{c(exp)}$ variations in terms of the values of Platt, WW, MinZL independent variables.

**4.1.5.** Also, adjusted determination coefficient $R^2_{adjust}$ indicates the percentage of dependent variable that is justified by the independent variable. The small differences between $R^2_{adjust}$ and $R^2$ indicates that independent variables added to the model have been chosen more appropriately. The slight difference between the above amounts in the proposed model verifies the precision and accuracy of the model for predicting the critical properties. So, in accordance with the unstandardized coefficients, the models (40), (41) and (42) were determined for estimation of $T_c$, $P_c$ and $V_c$, respectively.

**4.1.6.** If the standard deviation of a set of data is close to zero, it indicates that the data are close to the average and have low dispersion.

**4.1.7**. Standard Error of Estimate (STD) or RMSE is used to indicate the spread of values in a distribution. It is a standard method for determining the normal, above–normal and below–normal values. It measures the error rate between the two datasets. RMSE usually compares a predicted value and an observed value.

Finally, the comparison between equations and the values of statistical coefficients showed the best models for predicting $T_c$, $P_c$ and $V_c$ of alkanes using the MLR method which are summarized as follow:

$T_c$=8.75 Platt – 0.113 WW+ 19.995 MinZL+ 232.130 ; DW=2.01

$P_c$=−68615.237 Platt+1413.541 WW – 194862.228 MinZL+ 5494998.974; DW=1.55

$V_c$=1.006E−5 Platt +1.931E-7 WW+ 2.123E-5 MinZL +5.526 E-5; DW=1.85

**4.1.8.** Standard Coefficient β: The values of standard coefficients of β related to effective descriptors used for predicting $T_c$ and $V_c$ in the final equations using MLR method were obtained as follow:

**Table 3.** The standard coefficients β values of Platt, WW, MinZL

| Descriptor/ $T_c$ | **B** | Descriptor / $P_c$ | **β** | Descriptor / $V_c$ | **β** |
|---|---|---|---|---|---|
| Platt | 0.670 | Platt | -0.637 | Platt | 0.531 |
| WW | -0.160 | WW | 0.244 | WW | 0.189 |
| MinZL | 0.643 | MinZL | -0.760 | MinZL | 0.470 |

The standard correlation coefficient β value is a measure of how strongly each predictor variable influences the dependent variable. For example, the standard coefficients β=0.670, 0.531 for the Platt variable which are used for predicting $T_c$ and $V_c$, respectively, illustrate that compared to WW and MinZL predictors, the Platt index has the strongest influence on Tc and $V_c$. Similar to above explanations, the correlation coefficient β=-0.760 reveals that the descriptor MinZL has the highest influence on dependent variable $P_c$ than Platt index. Table 4 indicates the definitive values of $T_{c(pred)}$, $P_{c(pred)}$, $V_{c(pred)}$ of alkanes and their residuals using the equations (40), (41), (42) and MLR method.

**Table 4.** The values of predicted critical properties and residuals of alkanes derivatives.

| **No** | **$T_{c\ Pred}$/K** | **Res($T_c$)/K** | **$P_{c\ Pred}\times10^{-6}$/Pa** | **Res($P_c$)/Pa** | **$V_{c\ Pred}\times10^{4}/m^3$** | **Res ($V_c$) /$m^3$** |
|---|---|---|---|---|---|---|
| **1** | 330.59 | −25.293 | 4.53574 | 364258.3 | 1.60 | −1.31E-05 |
| **2** | 379.43 | −9.535 | 4.09433 | 155665.5 | 2.15 | −1.48E-05 |
| **3** | 422.80 | 2.298 | 3.70817 | 91824.6 | 2.65 | −1.05E-05 |
| **4** | 418.04 | −10.346 | 3.78689 | −136899 | 2.61 | −2.05E-06 |
| **5** | 466.63 | 3.160 | 3.32056 | 39437.23 | 3.20 | −8.85E-06 |
| **6** | 455.33 | 5.663 | 3.46183 | −81833.6 | 3.07 | −1.20E-06 |
| **7** | 453.63 | −19.824 | 3.54328 | −343284 | 3.07 | 4.18E-07 |
| **8** | 500.99 | 6.614 | 3.03014 | −10149.5 | 3.69 | −8.11E-07 |
| **9** | 499.84 | −2.046 | 3.07081 | -35818.8 | 3.65 | 2.61E-06 |
| **10** | 484.50 | 19.499 | 3.21910 | -109106 | 3.48 | 2.02E-05 |
| **11** | 510.43 | −21.433 | 3.02988 | 70110.85 | 3.75 | −1.73E-05 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **12** | 489.79 | 10.305 | 3.19864 | −48641 | 3.53 | 8.19E-06 |
| **13** | 538.36 | 1.637 | 2.71680 | 23192.17 | 4.28 | 4.36E-07 |
| **14** | 535.29 | −4.799 | 2.77432 | −34327.9 | 4.20 | 7.80E-07 |
| **15** | 517.92 | 17.280 | 2.94087 | −130879 | 3.99 | 5.05E-06 |
| **16** | 549.01 | −28.515 | 2.69977 | 70223.33 | 4.30 | −1.44E-05 |
| **17** | 531.62 | 5.673 | 2.83559 | 74406.4 | 4.10 | −1.70E-05 |
| **18** | 515.73 | 4.0718 | 2.99304 | −253049 | 3.96 | 2.23E-05 |
| **19** | 534.92 | 1.478 | 2.83461 | 115385 | 4.13 | 1.11E-06 |
| **20** | 542.61 | −11.514 | 2.79076 | 159233.3 | 4.20 | −2.24E-05 |
| **21** | 500.54 | 40.060 | 3.10743 | −217431 | 3.78 | 3.83E-05 |
| **22** | 572.58 | −3.680 | 2.44304 | 46954.71 | 4.92 | 2.74E-07 |
| **23** | 572.12 | −8.621 | 2.55097 | 179022.4 | 4.68 | −3.17E-05 |
| **24** | 563.77 | 9.725 | 2.63106 | 188937.2 | 4.58 | −2.56E-06 |
| **25** | 579.56 | −35.665 | 2.48220 | 87799.26 | 4.79 | −1.13E-05 |
| **26** | 583.38 | −33.576 | 2.41864 | 111356.9 | 4.89 | −1.07E-05 |
| **27** | 570.81 | −8.811 | 2.53545 | 114551.2 | 4.70 | −2.67E-05 |
| **28** | 547.68 | 28.816 | 2.75675 | 53246.74 | 4.41 | 1.39E-05 |
| **29** | 563.03 | 3.371 | 2.60846 | 121533.7 | 4.59 | 1.33E-06 |
| **30** | 569.55 | −6.052 | 2.51816 | 111837 | 4.71 | −2.63E-06 |
| **31** | 526.94 | 40.157 | 2.92901 | −229019 | 4.21 | 2.10E-05 |
| **32** | 548.57 | 20.227 | 2.71979 | −29789.5 | 4.46 | 2.05E-05 |
| **33** | 549.50 | 3.994 | 2.71478 | −154782 | 4.51 | 2.14E-05 |
| **34** | 563.72 | −13.722 | 2.58063 | −90630.5 | 4.70 | 1.19E-05 |
| **35** | 569.70 | −10.007 | 2.49651 | 3483.528 | 4.82 | 5.61E-06 |
| **36** | 557.60 | 5.997 | 2.60976 | −59768.3 | 4.65 | −8.46E-07 |
| **37** | 556.17 | 5.532 | 2.62218 | −82186.8 | 4.62 | 1.42E-05 |
| **38** | 532.46 | 33.0340 | 2.84845 | −238459 | 4.32 | 2.31E-05 |
| **39** | 602.54 | −7.535 | 2.22211 | 77881.14 | 5.63 | −7.63E-06 |
| **40** | 625.41 | −7.615 | 2.08428 | 25712 | 6.40 | −1.61E-05 |

Figures 1, 2 and 3, show that there was a high linear correlation between the experimental and obtained critical properties and the estimated critical properties using the models. Figure 1 shows a high linear correlation ( $R^2$=0.9199) between the experimental and the obtained Tc using the equation (40). This diagram illustrates the values of $T_{c(pred)}$ variations obtained from equation (40) using a MLR method in terms of the $T_{c(exp)}$. The value of $R^2$=0.9199, in this diagram, indicates the fact that 91.99% of the $T_{c(pred)}$ variations

can appropriately be determined in terms of one unit variation in $T_{c(exp)}$. Figure 2, shows a high linear correlation ( $R^2$=0.9234) between the experimental and the obtained $P_c$ using the equation (41). In this diagram, the high correlation between $P_{c(exp)}$ and $P_{c(pred)}$ was obtained using the MLR. The obtained value of 0.9234 for $R^2$ indicates that %92.34 of the $P_{c(pred)}$ variations can be determined in terms of one unit variation in $P_{c(exp)}$.

According to Figure 3, the indicative equation relationship between $V_{c(pred)}$ obtained from MLR model (42) was calculated as Y=X−1E-18. The value of 0.972 for $R^2$ shows that %97.2 of the $V_{c(pred)}$ variations can be determined in terms of one unit variation in $V_{c(exp)}$.

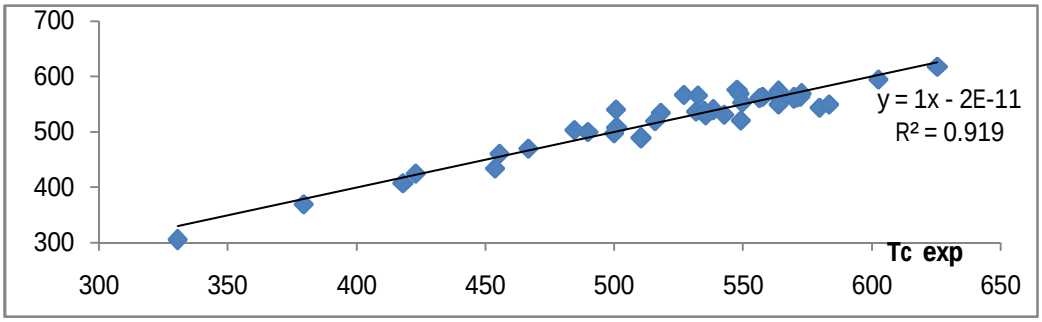


**Figure 1**. The plot of observed $T_c$ vs Predicted $T_c$.

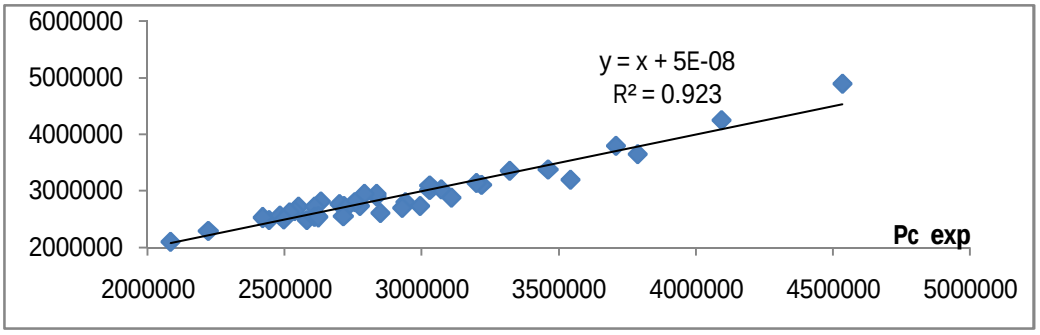


**Figure 2**. The plot of observed $P_c$ vs Predicted $P_c$.

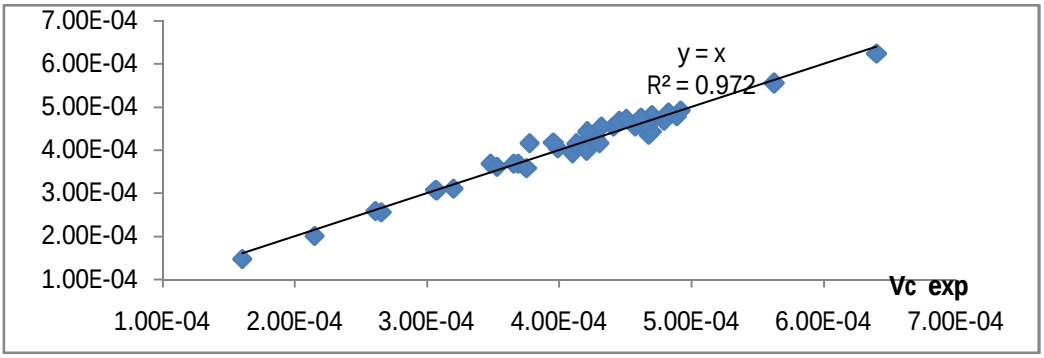


**Figure 3**. The plot of observed $V_c$ vs Predicted $V_c$.

It is obviously determined that the predicted values are so close to the experimental ones. So, it's inferred that the proposed patterns in these models have been selected correctly for determining critical properties of the studied molecules. The residual values are shown at a fairly random pattern (see Figures. 4, 5 and 6). Residuals are used to assess the normality of assumption. Figures 4, 5 and 6 show that the errors around x–axis have almost a uniform distribution. This proves the suitability of the selected pattern for proposed critical properties of alkanes.

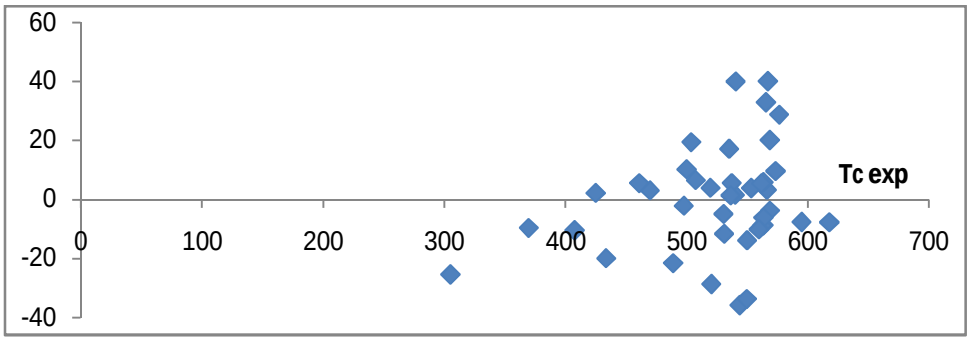


**Figure 4**. The plot of experimental $T_c$ vs the residuals.

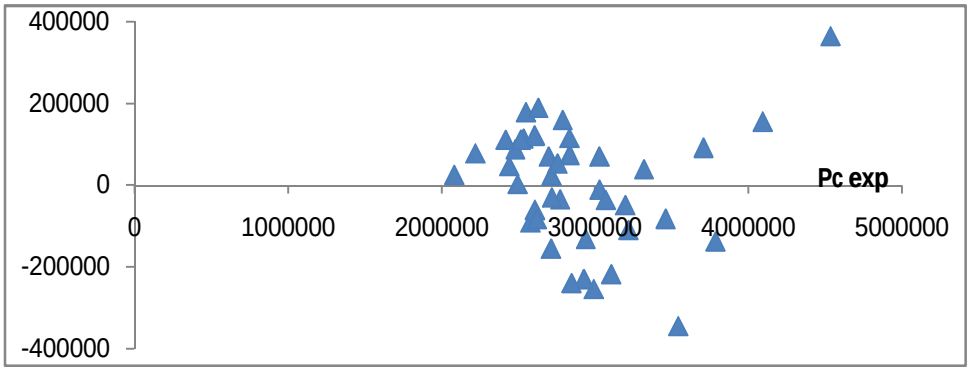


**Figure 5**. The plot of experimental $P_c$ vs the residuals.

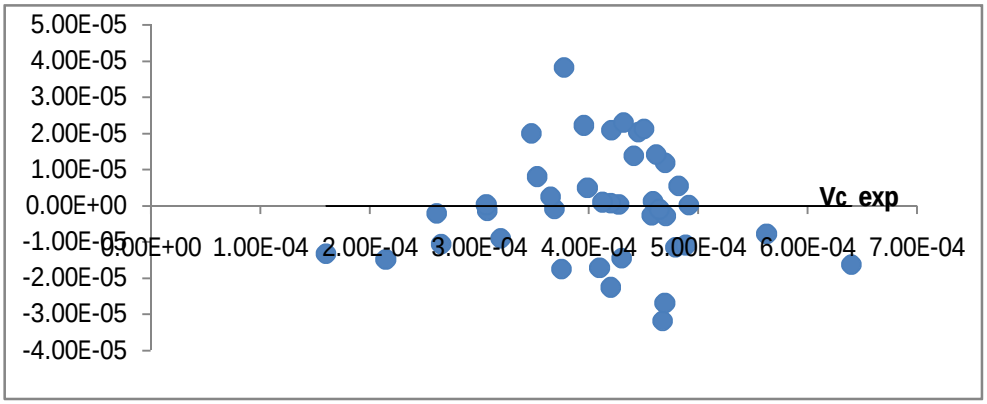


**Figure 6.** The plot of experimental $V_c$ vs the residuals.

## 5. QSAR MODEL VALIDATION

Typically, there are numerous methods for validation of QSAR models. Various statistical tests and coefficients can be used for validation of applied algorithms which, in the following, the most significant ones are represented. The statistical tests and coefficients used for estimation of $T_C$, $P_C$ and $V_C$ are as follow:

### 5.1. LEAVE–ONE–OUT CROSS–VALIDATION (LOOCV)

To determine the LOOCV, at first, a molecule from training set of 40 primary alkanes was removed. Then, QSPR was examined on the 39 remaining molecules. Considering the amounts of RSS and TSS, the amount of $Q^2$ was calculated based on the formula (43) and this cycle was repeated resulting in elimination of 25% of alkanes which were being studied leading to examining and calculating $Q^2$ for the remaining alkanes. Finally, the mean value of $Q^2$ was compared with $R^2$ in the final graphs, the results were shown in Table 5.

$$Q^2 = 1 - \frac{\sum (Y_{\exp(train)} - Y_{pred(train)})^2}{\sum (Y_{\exp(train)} - \overline{Y}_{Training})^2} = 1 - \frac{RSS}{TSS} \tag{43}$$

where, RSS refers to the residual sum of squares and TSS represents the total sum of square. This formula is the most widely used measure of the ability of a QSPR model to reproduce the data in the training set. We have computed the values of $Q^2$ (Eq. 43) using %25 of training set randomly. The values of $Q^2$ are defined as positive and less than one.

The small differences between mean $Q^2$ values of $T_c$, $P_c$, $V_c$ are equal to 0.9295, 0.9286, 0.9761, respectively, and the $R^2$ values of them verify the precision and accuracy of the model for predicting the critical properties.

**Table 5.** The values of $Q^2$ LOO of $T_c$, $P_c$, $V_c$.

| No. | $Q^2$LOO($T_c$) | $Q^2$LOO($P_c$) | $Q^2$LOO($V_c$) | Number | $Q^2$LOO($T_c$) | $Q^2$LOO($P_c$) | $Q^2$LOO($V_c$) |
|---|---|---|---|---|---|---|---|
| 1 | 0.9206 | 0.9250 | 0.9720 | 6 | 0.9350 | 0.9295 | 0.9755 |
| 2 | 0.9209 | 0.9252 | 0.9729 | 7 | 0.9353 | 0.9307 | 0.9757 |
| 3 | 0.9216 | 0.9254 | 0.9744 | 8 | 0.9355 | 0.9316 | 0.9802 |
| 4 | 0.9268 | 0.9259 | 0.9745 | 9 | 0.9359 | 0.9339 | 0.9803 |
| 5 | 0.9277 | 0.9283 | 0.9754 | 10 | 0.9360 | 0.9351 | 0.9805 |

### 5.2. MULTICOLLINEARITY TEST

Multicollinearity test was performed to avoid habits in the decision making process regarding the partial effect of independent variables on the dependent variable. A good regression model is a model in which there is not a high correlation between the independent variables. Multicollinearity test is performed through the use of SPSS software and the value of variance inflation factor (VIF) to avoid linear correlation between the independent variables. If the VIF value line is a number between 1 and 10, then there is no multicollinearity, and if VIF<1 or >10, then there is multicollinearity. In all our final models, the multicollinearity did not exist, because the values of correlations between independent variables are not close to one, and VIF values line between the numbers 1 to 10. The analysis of VIF values for all the descriptors indicated that the best models for predicting $T_c$, $P_c$ and $V_c$ values are: Platt, WW, and MinZL.

### 5.3. TEST FOR AUTOCORRELATION USING THE DURBIN–WATSON STATISTIC

From a statistical regression analysis lens of view, Durbin–Watson (DW) statistic is a number to examine autocorrelation in the residuals. The DW values 2.01, 1.55, 1.85 in final models are considered acceptable indicating that there is poor correlation between the errors and the independence of residuals. These numerical values indicate that our final models are perfect.

### 5.4. SKEWNESS AND KURTOSIS TEST

The normality of residuals represents whether the distribution function is symmetrical or asymmetrical. For a completely symmetrical distribution, the skewness and kurtosis are equal to zero. In a non–symmetrical distribution, when most of the scores “scrunch up” towards a few high scores it is positively skewed, and when most of the scores cluster towards a few low scores it is negatively skewed. Generally, if the skewness and kurtosis are placed at an interval between [2,-2], the data follow a normal distribution. The observed values for residuals skewness are 0.094, 0.076, 0.060, and the observed values for residuals kurtosis of variables $T_c$, $P_c$ and $V_c$, are 0.624, 0.500, -0.578, respectively. These indicate the normality of them.

### 5.5. APPLICABILITY DOMAIN

The applicability domain (AD) of QSAR model was used to verify the prediction reliability, identify the problematic compounds and predict the compounds with acceptable activity that fall within this domain. The most common methods used for determination of the AD of QSAR models have been described by Gramatica that used the leverage values for each compound. The leverage approach allows the determination of the position of new chemical in the QSAR model, *i.e.*

whether a new chemical will lie within the structural model domain or outside of it. The leverage approach along with the Williams Plot are used to determine the applicability domain in all QSAR models.

To construct the William Plot, the leverage $h_i$ for each chemical compound– in which QSAR model was used to predict its property– was calculated according to the following equation:

$$h_i=x^T(X^TX)^{-1}x \quad (44)$$

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

$$h^*=3(p+1)/n \quad (45)$$

where n is the number of training compounds and p is the number of predictor variables.
In this research, in each of the three models, the descriptor vector x includes the Platt, WW, MinZL descriptors and X is Platt, WW, MinZL descriptors matrix related to training set of alkanes. (The leverage values are shown in Table 6). Also, according to equation (45), the value of $h^*$ is equal to 0.3 (n=40 and p=3). Then, the defined applicability domain (AD) was visualized using a Williams plot, the plot of the standardized residuals versus the leverage values ($h$). A compound with $h_i > h^*$ seriously influences the regression performance and may be excluded from the applicability domain (See Figs. 7, 8, 9). The results indicated that among 40 different types of alkanes, there is just one outlier.

**Table 6.** The leverage values of used alkanes

| Alkane | $h_i$ | Alkane | $h_i$ | Alkane | $h_i$ | Alkane | $h_i$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.0836 | 11 | 0.0494 | 21 | 0.0234 | 31 | 0.0444 |
| 2 | 0.1047 | 12 | 0.0363 | 22 | 0.0946 | 32 | 0.0350 |
| 3 | 0.1146 | 13 | 0.0832 | 23 | 0.0952 | 33 | 0.0363 |
| 4 | 0.0569 | 14 | 0.0545 | 24 | 0.1050 | 34 | 0.0358 |
| 5 | 0.1193 | 15 | 0.0303 | 25 | 0.0864 | 35 | 0.0483 |
| 6 | 0.0558 | 16 | 0.0443 | 26 | 0.0458 | 36 | 0.0383 |
| 7 | 0.0453 | 17 | 0.0338 | 27 | 0.0512 | 37 | 0.0364 |
| 8 | 0.0979 | 18 | 0.0285 | 28 | 0.0673 | 38 | 0.0287 |
| 9 | 0.0556 | 19 | 0.0493 | 29 | 0.0560 | 39 | 0.2072 |
| 10 | 0.0417 | 20 | 0.0843 | 30 | 0.0329 | 40 | 0.5754 |

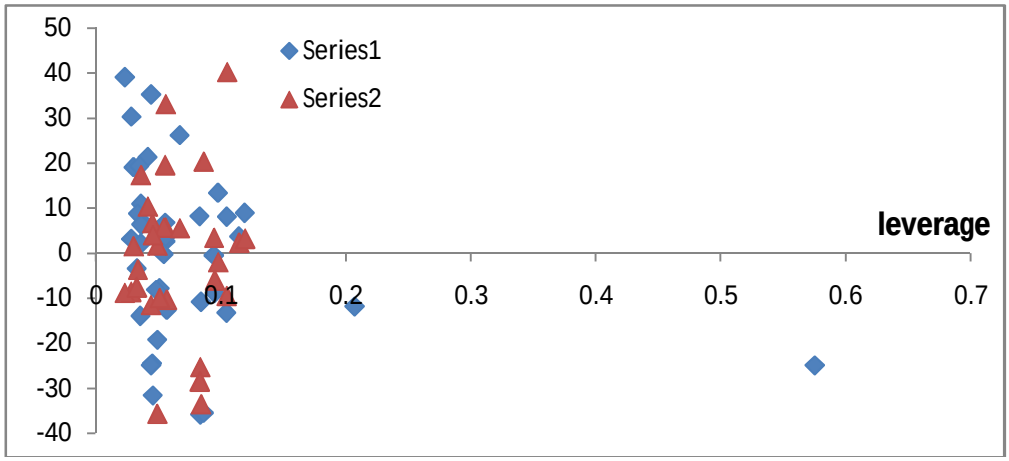


**Figure 7**. Williams plot of residual $T_c$ vs leverage.

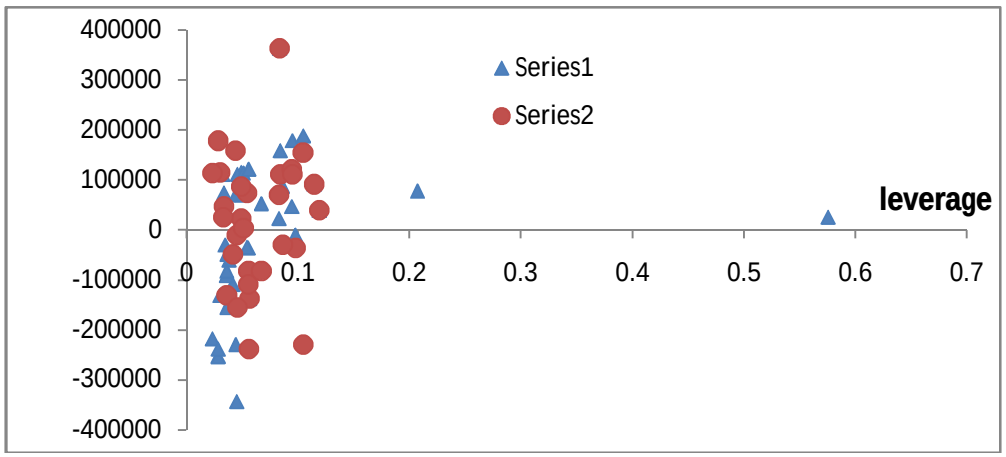


**Figure 8.** Williams plot of residual $P_c$ vs leverage.

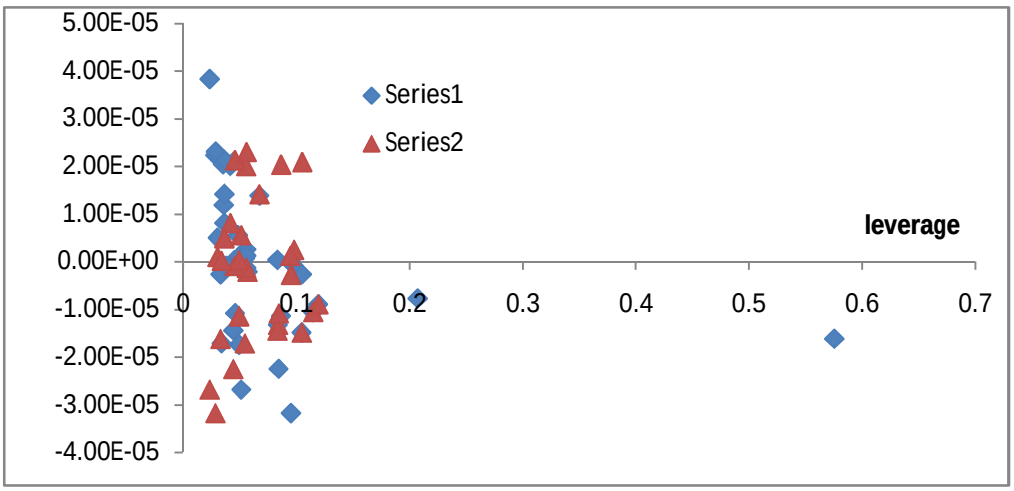


**Figure 9.** Williams plot of residual $V_c$ vs leverage.

Thus, the analyses of various statistical coefficients, tables, diagrams and QSPR model validation through the use of MLR method show that they possess the necessary and sufficient validity for predicting the critical properties of alkanes.

## 6. Conclusion

The afore–mentioned computational methods involve methods which all focus on the molecular structures and properties. The underlying concept of these methods is based on the fact that the molecular and geometric structures are responsible for all the physical and chemical properties of molecules including t. The results of the present study indicate that the simple linear regression model with dispersion coefficient (alone) is not sufficient for determining the critical properties of alkanes. However, the multiple linear and regression model benefiting from various descriptors, factors and efficient coefficients can suggest the best algorithm for determining these properties. It was also witnessed that among the proposed models to predict the critical properties, the model including a combination of descriptors Hyper–Wiener, Platt, and MinZL is the most appropriate one. And the last but not least, this was the first time that the relationship between critical properties with molecular descriptors of alkanes and their derivatives was investigated through the use of SPSS software and MLR method.

## REFERENCES

1. R. T. Morison and R. Neilson Boyd, *Organic Chemistry*, Allyn & Bacon, 2003.
2. H. Wiener, Correlation of heats of isomerization and differences in heats of vaporization of isomers, among the paraffin hydrocarbons, *J. Am. Chem. Soc.* **69** (1947) 2636–2638.
3. A. A. Gakh, E. G. Gakh, B. G. Sumpter and D. W. Noid, Neural network–graph theory approach to the prediction of the physical properties of organic compounds, *J. Chem. Inf. Comput. Sci.* **34** (1994) 832–839.
4. O. Ivanciuć, The neural network MolNet prediction of alkane enthalpies, *Anal. Chem. Acta.* **384** (1999) 271–284.
5. D. C. Montgomery, E. A. Peck and G. G. Vining, *Introduction to Linear Regression Analysis*, John Wiley & Sons, Inc, 2006.
6. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry,* Wiley–VCH Verlag GmbH, 2008.
7. I. Gutman and B. Furtula (eds), *Novel Molecular Structure Descriptors–Theory and Applications I and II*, University of Kragujevac and Faculty of Science Kragujevac, 2010.

8. I. Gutman, A formula for the Wiener number of trees and its extension to graphs containing cycles, *Graph Theory Notes N. Y.* **27** (1994) 9−15.
9. R. B. King, *Chemical Applications of Topology and Graph Theory*, Elsevier, Amsterdam, 1983.
10. I. Gutman and O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer–Verlag, Berlin, 1986.
11. M. Randić, *Chemical Graph Theory–Facts and Fiction*, NISCAIR–CSIR, India, 2003.
12. M. Randić, Quantitative Structure–property relationship: boiling points of planar 1009–benzenoids, *New. J. Chem*. **20** (1996) 1001–1009.
13. M. Shamsipur, B. Hemmateenejad and M. Akhond, Highly Correlating Distance/Connectivity–Based Topological Indices. 1: QSPR Studies of Alkanes, *Bull. Korean Chem. Soc*. 25 (2004) 253–259.
14. H. Hosseini and F. Shafiei, Quantitative Structure Property Relationship Models for the Prediction of Gas Heat Capacity of Benzene Derivatives Using Topological Indices, *MATCH Commun. Math. Comput. Chem*. **75** (2016) 583–592.
15. M. Goodarzi and E. Mohammadinasab, Theoretical investigation of relationship between quantum chemical descriptors, topological indices, energy and electric moments of zig–zag polyhex carbon nanotubes $TUHC_6[2p,q]$ with various circumference [2p] and fixed lengths, *Fullerenes, Nanotubes Carbon Nanostructures.* **21** (2013) 102–112.
16. A. Alaghebandi and F. Shafiei, QSPR modeling of heat capacity, thermal energy and entropy of aliphatic aldehydes by using topological indices and MLR method, *Iranian J. Math. Chem*. **7** (2016) 235–251.
17. M. Pashm Forush, F. Shafie and F. Dialamehpour, QSPR study on benzene derivatives to some physico chemical properties by using topological indices, *Iranian J. Math. Chem*. **7**(1) (2016) 93–110.
18. H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc*. **69** (1947) 17–20.
19. G. Cash, S. Klavžar and M Petkovsek, Three Methods for Calculation of the Hyper–Wiener Index of Molecular Graphs, *J. Chem. Inf. Comput. Sci*. **42** (2002) 571–576.
20. X. Li and Y. Shi, A survey on the Randić index, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 127–156.
21. M. Randić, Characterization of atoms, molecules and classes of molecules based on paths. enumerations, *MATCH Commun. Math. Comput. Chem*. **7** (1979) 5–64.
22. B. Liu and I. Gutman, On general Randić indices, *MATCH Commun. Math. Comput. Chem.* **58** (2007) 147–154.

23. M. Randić, Charactrization of molecular branching, *J. Am. Chem. Soc*. **97** (1975) 6609–6615.
24. A. T. Balaban and T. S. Balaban, New Vertex Invariant and topological indices of chemical graphs based on information on distance, *J. Math. Chem*. **8** (1991) 383–397.
25. A. T. Balaban, Topological index based on topological distances in molecular graph, *Pure Appl. Chem.* **55** (1983) 199–206.
26. K. C. Das, B. Zhou and N. Trinajstić, Bounds on Harary index, *J. Math. Chem*. (2009) 1377–1393.
27. www.chemicalize.org
28. M. Randić and S. C. Basak, Multiple regression analysis with optimal molecular descriptors, *SAR QSAR Environ. Res.* **11** (2000) 1–23.
29. G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, Hoboken, NJ: Wiley–Interscience, 2003.
30. K. Roy and I. Mitra, On various metrics used fo r validation of predictive QSAR models with applications in virtual screening and focused library design, *Comb. Chem. High Throughput Screen.* **14** (2011) 450–474.

# *Determination of Critical Properties of Alkanes Derivatives using Multiple Linear Regression*

**ESMAT MOHAMMADINASAB**

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

# تعیین خواص بحرانی مشتقات آلکان‌ها با استفاده از رگرسیون خطی چندگانه

ادیتور رابط : علیرضا اشرفی

## چکیده

در این پژوهش، برخی روش‌های محاسباتی ریاضی برای تخمین دما و فشار و حجم بحرانی ۴۰ نوع مختلف از آلکانها و مشتقاتشان ارائه شده است. در این مدل، رابطه کمی ساختار-خاصیت بر اساس نظریه گراف با برخی شاخصهای ساختاری و هندسی ترکیبات شیمیایی به کار گرفته شده است. برای بررسی همبستگی میان خواص بحرانی و توصیف‌گرهای مولکولی، مدل رگرسیون خطی چندگانه با کمک ضرایب مناسب بکار گرفته شده است. برای این منظور، از بین یازده توصیف‌گر ساختاری و هندسی مورد مطالعه، مناسب‌ترین آنها برای تعیین خواص بحرانی آلکانها انتخاب شدند. نتایج نشان دادند که از بین مدلهای پیشنهادی برای پیش بینی دما و فشار و حجم بحرانی آلکانهای مورد مطالعه، مدل شامل ترکیبی از توصیف گرهای hyper-Wiener، Platt، MinZL مناسب‌ترین مدل می باشد.

**لغات کلیدی**: آلکانها، MLR، خواص بحرانی، QSPR