Multiple Linear Regression-Based Predictions of Physicochemical Properties of Aliphatic Hydrocarbons Using Zagreb Vector Index and VL-Index

ABSTRACT

In the present paper, we study the variation of physical properties in a series of aliphatic hydrocarbons using the Zagreb vector index and VL-index. A multiple linear regression technique is applied to predict physical properties such as melting point and boiling point of various chemicals.

Keywords: Zagreb vector index, VL-index, QSPR, Maxima, Scilab, R programming.

I. INTRODUCTION

The various topological indices introduced in the literature serve as a crucial bridge between mathematics and chemistry, providing quantitative measures that describe molecular structure in terms of graph-theoretical parameters. These indices enable researchers to model, compare, and predict various physicochemical and biological properties of chemical compounds through purely mathematical formulations. Over the years, a wide range of indices, based on degree sequences, distance matrices, Zagreb matrices, and eccentricity matrices, have been developed and successfully applied in Quantitative Structure–Property Relationship (QSPR) and Quantitative Structure–Activity Relationship (QSAR) studies[14].

Among these, Zagreb-based indices have received special attention because of their computational simplicity and strong discriminating power in predicting molecular properties. Building upon this foundation, the Zagreb vector indices were introduced as an extension to capture more structural information simultaneously. Unlike traditional scalar indices, Zagreb vector indices provide a multidimensional representation of molecular topology that considers interrelated parameters in a unified framework, thereby improving the accuracy of molecular discrimination and prediction[16].

This study is motivated by the need to explore the relationship between Zagreb vector indices and the physical properties of selected chemical compounds. The primary objective is to evaluate how variations in these topological indices correspond to the variations observed in physicochemical properties such as boiling points, melting points, and other measurable characteristics. The strength and nature of these relationships are quantified using correlation coefficients[11], which measure the degree of linear association between variables.

According to standard statistical theory, when variables exhibit linear dependency, linear regression analysis serves as an effective modeling technique[10]. The correlation coefficient ***r*** is interpreted in terms of its absolute value to determine the strength of association, classified as very weak (0–0.19), weak (0.20–0.39), moderate (0.40–0.59), strong (0.60–0.79), and very strong (0.80–1.00)[11]. To model and predict physical properties, both simple linear regression and multiple linear regression are employed, where the dependent variable represents a selected physical property and the independent variables correspond to one or more Zagreb vector indices. In contrast with other approximation techniques such as discriminant analysis, logistic regression, decision tree methods, Fourier transformations, or neural networks, linear regression remains one of the most practical and interpretable approaches, particularly when the underlying relationship is linear[13].

II. PRELIMINARIES

This work encounters two main types of challenges: technical and mathematical[12].

Technical Challenges

Collecting accurate records of physical properties and chemical structures, particularly for compounds with more than 10 carbon atoms. These often deviate from ideal straight-chain structures, leading to imprecise experimental data such as boiling points.Variations due to changing positions of double bonds, which produce multiple isomers and irregular data patterns.

Mathematical Challenges

Although regression is a simple technique for predicting unknown values from known variables, establishing a mathematical model relating dependent and independent variables remains challenging. Multiple linear regression is employed here for predictions, with all calculations performed using R programming. Key difficulties include selecting appropriate approximation methods (focusing on simple programming techniques), finding the best-fitting curve for data as a function of single or multiple variables, detecting relationships that correlation factors may miss, and precisely choosing dependent and independent variables.

Advantages of Vector Index

The vector index proves more useful than traditional topological indices for computing physical properties of organic compounds, particularly in distinguishing isomers from straight-chain variants. However, it remains constant for similar compounds like methane and CCl_4, yields large values requiring computational aids as carbon atoms increase, and thus has limitations.

Advantages of VL- Index

The VL-index is considered more effective than traditional topological indices for computing the physical properties of organic compounds, particularly due to its ability to distinguish between isomers and their straight-chain counterparts. This capability makes it a valuable tool in molecular

characterization and property prediction. However, a limitation of the VL-index is that it remains constant for structurally very similar compounds, which can reduce its effectiveness in differentiating such closely related molecules.

III. MAIN RESULTS

Regression Approach

This paper emphasizes multiple linear regression alongside quantitative structure analysis and vector indices. Simple regression uses one predictor, while multiple regression employs two or more; the latter better captures consistent relationships across variables [12].

R programming facilitates multiple regressions and data frame construction. Chemical properties such as surface tension, boiling point, melting point, or molar volume serve as dependent variables, with Zagreb vector index and other parameters as independents. The analysis targets families like hydrocarbons, where properties increase with carbon atoms, incorporating both statistical and polynomial approximations[12].

Multiple Linear Regression Analysis in R

In this section, multivariate regression techniques are employed using the R programming language to model various physicochemical properties of hydrocarbons. These properties include boiling point, melting point, molar volume, and surface tension, which serve as response variables in the analysis[5]. The predictors utilized comprise the first Zagreb vector index along with other relevant molecular parameters that influence these physical characteristics. The study particularly focuses on hydrocarbons, whose physicochemical properties demonstrate predictable variations in relation to their carbon atom count. By integrating statistical regression with polynomial approximation methods, the analysis aims to develop robust models capable of accurately estimating these thermal and volumetric properties [5][12].

Regression Formula for Alkanes Melting Point and Boiling Points

This subsection presents the development of regression formulas to predict the melting and boiling points of several lower alkanes, as shown in Table 1. Utilizing physicochemical parameters and R programming techniques, these formulas aim to accurately estimate the melting and boiling points of alkanes[5][12][2].

The MV_1-index of alkanes ($n \geq 3$) is computed using Maxima as follows [12]:

$$\mathbf{MV_1(C_nH_{(2n+2)}) = (n-3)(8,4,4) + 2(8,3,5) + (n-2)(5,2,3) + 6(5,1,4)}$$

In the same context, the VL-index of alkanes is evaluated via SciLab and expressed as[12]:

$$\mathbf{VL(C_nH_{(2n+2)}) = 0.5((n-1)(6+14+4) + 2(n+1)(3+2+4))}$$

A table 1 presenting the VL-index and variations of the Zagreb vector index for alkanes is provided below.

Table 1 : The MV_1 and VL-indices of variation for alkanes.

Sl. No.	Formula	Alkane	MV_1-index	VL-index	MP (°C)	BP (°C)
1	CH_4	Methane	28.30	18	−184.6	−160.7
2	C_2H_6	Ethane	49.07	39	−164.8	−69.62
3	C_3H_8	Propane	64.54	60	−147.8	−40.58
4	C_4H_{10}	n-Butane	80.20	81	−130.7	−9.30
5	C_5H_{12}	n-Pentane	95.94	102	−113.6	22.87
6	C_6H_{14}	n-Hexane	111.75	123	−96.4	55.89
7	C_7H_{16}	n-Heptane	127.59	144	−79.2	89.30
8	C_8H_{18}	n-Octane	143.44	165	−62.0	122.70
9	C_9H_{20}	n-Nonane	159.32	186	−44.8	156.60
10	$C_{10}H_{22}$	n-Decane	175.20	207	−27.6	175.20

For alkanes ranging from C_2 to C_{10}, the observed boiling points show strong consistency with standard references, such as the NIST Chemistry Web Book and the CRC Handbook [3][8].

Beyond C_{10}, however, accurate experimental measurements are limited. As a result, the tabulated values from C_{11} onward are based primarily on extrapolated predictions rather than direct measurements.

Starting from C_{11} and onwards, the dataset exhibits an almost uniform increase in boiling point of approximately 16–17°C for each additional carbon atom.

To quantify this trend, a linear regression analysis was performed by correlating the boiling point, bp, with the number of carbon atoms, ***n***[10].

The regression equation is given by

$$\mathbf{bp(n) = a\,n + b}$$

where bp(n) represents the boiling point of an alkane with n carbons, a is the slope corresponding to the average rise in temperature per carbon, and b is the intercept.

From the statistical analysis, the parameters were obtained as

$$\mathbf{a = 16.84,} \qquad \mathbf{b = 13.12}$$

leading to the predictive model

$$\mathbf{bp(n) = 16.84n + 13.12}$$

This regression-based method offers a reliable approach for predicting the boiling points of higher alkanes (n > 10), in cases where experimental measurements are not available [3][10].

The generalized regression equation for the melting point (Mp) of alkanes (n > 11) is:

$$\mathbf{Mp(n) = 16.84n + b}$$

where b is the intercept determined from statistical fitting to available data[3][8].

Regression Analysis Results

The R programming code for multiple linear regression is as follows[12]:

```
mvi <- c(28.3, 49.07, 64.54, 80.20, 95.94, 111.75, 127.59, 143.44, 159.32, 175.20)

vli <- c(18, 39, 60, 81, 102, 123, 144, 165, 186, 207)

mp <- c(-188, -186, -183, -38, -130, -95, -91, -57, -54, -30)

modula <- lm(formula = mp ~ mvi + vli)
```

Summary(modula)

The output begins with:

Call: lm(formula = mp ~ mvi + vli)

Residuals Summary

Residuals	Min	1Q	Median	3Q	Max
Values	−35.177	−15.220	−6.245	0.512	92.741

Coefficients

Term	Estimate	Std. Error	t value	*p*-value
Intercept	−207.2195	183.5722	−1.129	0.296
mvi	0.5349	10.3136	0.052	0.960
vli	0.4146	7.8795	0.053	0.960

Residual standard error[9]: 39.32 on 7 degrees of freedom

Multiple R-squared[9]: 0.6949, Adjusted R-squared[9]: 0.6078

F-statistic[9]: 7.974 on 2 and 7 DF, p-value[9]: 0.01568

The regression function is defined as:

mp_calc(mvic,vlic) = −207.22 + 0.535 × mvic + 0.415 × vlic

The input vectors are:

mvic=[159.32,175.20,191.10,207.00,…,1768.47]

mvic=[159.32,175.20,191.10,207.00,…,1768.47]

vlic=[186,207,228,249,…,2307]

vlic=[186,207,228,249,…,2307]

Using this function, melting points are computed for the compounds. The calculated values range from 834.26 to 1696.31 and are presented in red in the table. Note that an increase in the Zagreb vector index (mvi) corresponds to a decrease in surface tension, as shown in the separate linear model:

St = 15.04336 − 0.03939 ×mvi + 0.09824 × bpc

Boiling Point Regression Results:

A similar R program for boiling point approximation yields the following results:

Call: lm(formula = bp ~ mvi + vli)

Residuals:

Residuals	Min	1Q	Median	3Q	Max
Values	**−19.3844**	**−5.1792**	**0.7507**	**8.9144**	**13.2099**

Coefficients:

Term	Estimate	Std. Error	t-value	*p*-value

Term	Estimate	Std. Error	t-value	*p*-value
Intercept	−361.781	60.924	−5.938	0.000577
mvi	11.713	3.423	3.422	0.011108
vli	−7.246	2.615	−2.771	0.027668

Model Statistics[9]:

Residual standard error: 13.05 on 7 degrees of freedom

Multiple R-squared: 0.989, Adjusted R-squared: 0.9859

F-statistic: 314.5 on 2 and 7 DF, p-value: 1.398e-07

The approximation equation is:

$$\mathbf{Bp = -361.78 + 11.713 \times mvi - 7.246 \times vli}$$

The analysis is extended to unsaturated hydrocarbons such as alkenes and alkynes, calculating both Zagreb vector indices and VL-index using the same methodology.

Regression Formula for Alkenes Boiling Points:

In this section, we examine alkenes, which are unsaturated hydrocarbons containing one or more double bonds. The MV_1- and VL-indices are calculated for selected lower members, as shown in Table 2, to validate the computational approach before applying it to higher alkenes [2][12].

Table 2: Physical Properties of Alkenes

IUPAC Name	Molecular Formula (MF)	Structural Formula (SF)	MV_1	VL	MP (°C)	BP (°C)
Ethene	C_2H_4	$CH_2{=}CH_2$	41.95	84	−169.15	−90.28
Propene	C_3H_6	$CH_2{=}CHCH_3$	63.00	126	−185.20	−49.28
1-Butene	C_4H_8	$CH_2{=}CHCH_2CH_3$	84.00	168	−185.30	−10.13
1-Pentene	C_5H_{10}	$CH_2{=}CH(CH_2)_2CH_3$	105.00	210	−165.00	26.68
1-Hexene	C_6H_{12}	$CH_2{=}CH(CH_2)_3CH_3$	126.00	252	−140.00	62.95
1-Heptene	C_7H_{14}	$CH_2{=}CH(CH_2)_4CH_3$	147.00	294	−119.00	99.68
1-Octene	C_8H_{16}	$CH_2{=}CH(CH_2)_5CH_3$	168.00	336	−101.70	136.83

This section constructs a linear regression model using the VL-index and melting point as predictors, with boiling point as the response. Results show a negative correlation between melting point and boiling point, indicating an inverse relationship.

Residuals Summary

Statistic	Min	1Q	Median	3Q	Max
Values	−15.87243	−6.83091	2.30070	8.45874	10.31607

Regression Output Summary

Predictor	Coefficient Estimate	SE	t-Ratio	Sig. (p)
Baseline Term	−188.0828	86.9205	−2.164	0.0965
VL-metric	0.9321	0.1349	6.909	0.0023
Melting Point (MP)	−0.1153	0.4304	−0.268	0.8020

However, the resulting regression formula to approximate the boiling point (BP) of alkenes is:

$$\mathbf{BP = (-188.083) + 0.9321 \times VLi - 0.1153 \times MP}$$

Boiling Point Regression Formula for Alkynes

This section extends the analysis to alkynes, a class of hydrocarbons characterized by the presence of at least one carbon-carbon triple bond. The MV_1 and VL-indices are computed for selected lower members, following the procedure outlined in Table 3.

These computations provide a quantitative basis for examining the relationship between the topological indices and the structural features of alkynes[2][12].

Table 3: Alkyne Data with Molecular Formula, MV_1-index, VL-index, MP, and BP

IUPAC Name	Molecular Formula (MF)	MV_1-index	VL-index	MP (°C)	BP (°C)
Ethyne	C_2H_2	36.33	90	−80.8	−74.74
Propyne	C_3H_4	49.20	123	−102.7	−33.48
1-Butyne	C_4H_6	62.40	156	−125.7	8.14
1-Pentyne	C_5H_8	75.60	189	−105.5	35.49
1-Hexyne	C_6H_{10}	88.80	222	−132.0	78.27
1-Heptyne	C_7H_{12}	102.00	255	−81.0	95.45
1-Octyne	C_8H_{14}	115.20	288	−80.0	129.14

In this section, we formulate a linear regression model where the VL-index and melting point serve as independent variables, and the boiling point is the dependent variable.

The analysis reveals that the melting point significantly contributes to explaining variations in the boiling point.

Residuals Summary

Statistic	Min	1Q	Median	3Q	Max
Values	−9.25780	−4.64298	0.00358	4.12930	10.28157

Parameter Estimates

Term	Value	SE	t-stat	p-value
Constant	−194.23364	19.71320	−9.853	0.000595
VL-descriptor	1.03106	0.04831	21.341	2.8505
MP	−0.33040	0.15731	−2.100	0.103610

The resulting regression formula to approximate the boiling point (BP) of alkynes is:

$$\mathbf{BP = (–194.23364) + 1.03106 \times VL\text{-}index – 0.33040 \times MP}$$

CONCLUSION

This study examined the variation of physical properties in aliphatic hydrocarbons using the Zagreb vector index and VL-index, applying multiple linear regression to predict melting and boiling points.The novel vector indices, incorporating vertex degrees and connectivity patterns, demonstrated suitability for chemical graph analysis and outperformed traditional degree or eccentricity-based topological indices in regression modeling. While optimal regression curves remain challenging due to multifaceted molecular influences, hydrogenation and polymerization processes revealed linear property trends, addressing key chemical, mathematical, and statistical challenges in quantitative structure-property relationships.

REFERENCES

[1]. Anand G. Puranik, Narendra V. H., and Mahalakshmi P., On the Correlation Analysis Between Vector Index of Alkanes, “International Journal of Creative Research Thoughts”, vol. 11, no. 10, pp. 103--113, 2023.

[2]. A. Bahl and B. S. Bahl, “Advanced Organic Chemistry”, S. Chand Publishing, New Delhi, India, 2000.

[3]. J.S. Chickos, in “NIST Chemistry WebBook”, ed. P. J. Linstrom and W. G. Mallard, National Institute of Standards and Technology, Gaithersburg, MD, NIST Standard Reference Database Number 69.

[4]. Darlami, J., & Sharma, S. (2024). “The role of physicochemical and topological parameters in drug design”. Frontiers in Drug Discovery, 4, 1424402. https://doi.org/10.3389/fddsv.2024.1424402

[5]. Draper, N. R., & Smith, H. (1998). ”Applied Regression Analysis (3rd ed.)”. Wiley-Interscience.

[6]. Gutman, I., & Trinajstić, N. (1972). Graph theory and molecular orbitals. “Total φ-electron energy of alternant hydrocarbons”. Chemical Physics Letters, 17(4), 535–538.

[7]. Gutman, I., & Trinajstić, N. (1972). Graph theory and molecular orbitals. ”Chemical Physics Letters”, 17(4), 535–538.

[8]. W. M. Haynes, D. R. Lide, and T. J. Bruno, “CRC Handbook of Chemistry and Physics”, 97th ed., CRC Press, Boca Raton, FL, 2016.

[9]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). ”An Introduction to Statistical Learning with Applications in R (2nd ed.)”. Springer.

[10]. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). ”Introduction to Linear Regression Analysis (5th ed.)”. Wiley

[11]. Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. ”Malawi Medical Journal”, 24(3), 69–71.

[12]. Puranik, A. G., & Narendra, V. H. (2024). Prediction of physical properties of hydrocarbons using Zagreb vector index and multiple linear regression. “International Journal of All Research Education and Scientific Methods (IJARESM)”, 12(7), July 2024. ISSN 2455-6211.

[13]. Randić, M. (1975). Characterization of molecular branching. ”Journal of the American Chemical Society”, 97(23), 6609–6615.

[14]. Sharma, V., Goswami, R., & Madan, A. K. (1997). Eccentric connectivity index: A novel highly discriminating topological descriptor for structure–property and structure–activity studies. ”Journal of Chemical Information and Computer Sciences”, 37(2), 273–282.

[15]. Todeschini, R., & Consonni, V. (2000). ”Handbook of molecular descriptors”. Wiley-VCH.

[16]. Todeschini, R., & Consonni, V. (2009). Molecular Descriptors for Chemoinformatics. Wiley-VCH.