QSPR Study of Alkylbenzenes using Principal Component Regression Analysis

Aditya Pegu[1*], Monjit Chamua[2], Sumanta Borah[3] and A Bharali[4]

[1,2,3,4] *Department of Mathematics, Dibrugarh University, Assam, India*

Correspondence Author : **A Pegu**

Abstract: The QSPR, which connects structural features to physicochemical properties, is a useful part of drug design and discovery. The ultimate goal of the QSPR formulation is to develop mathematical models that estimate the physicochemical properties of molecular structures.There are over 3000 Topological indices (TIs) in the literature, therefore one must decide how to pick those that best describe the physicochemical property being studied. And in a regression equation inclusion of large number of TIs may increases the fit but the predictive ability of the developed model will face a substantial decrease due to their multicollinearity. Applying principal component analysis is the best method as they will reduce the dimension without losing the original data of the indices. Also, it eliminates the problems of multicollinearity among the indices and hence provides a good predictive model. In this article we have considered 37 degree-based and neighborhood degree-based topological indices to predict the physicochemical properties of 42 alkyl benzenes such as boiling point, critical pressure, critical volume and critical temperature using multilinear regression analysis. Also, we use principal component analysis to reduce the dimension and to overcome the multicollinearity among the indices.

Keywords: Principal component analysis, degree, neighbourhood degree, QSPR, topological index.

1. Introduction

Graph-theorical and topological models have applications in diverse areas of scientific research as theoretical physics, chemistry, pharmacology and pharmaceutical chemistry, toxicology, engineering, computer science, sociology, geography, architecture, and linguistics. In particular, the interactions of graph theory with chemistry have enriched both fields for more than a century. For example, graph theoretical concepts have been used in statistical mechanics, chemical physics, polymer chemistry, enumeration of constitutional isomers associated with an empirical formula, rationalization of the additivity principle of physical chemistry, determination of structural similarity and dissimilarity of congeners, calculation of quantum mechanical parameters of chemical structure, analysis of three-dimensional aromaticity of polyhedral cage compounds, chemical documentation, isomer discrimination and quantitation of molecular branching, representation of chemical changes through "reaction graphs", and computer-assisted prediction of chemical syntheses pathways. Graph-theoretic methods have a primary role in explaining physical and chemical phenomena where the topology and combinatorial nature of the problem is of paramount importance. Planar graph models of chemical species retain their full topology which largely determines such important of characteristics of molecules as energy, bond order, charge density, stability, and dipole moment. Hence it is not surprising that many important aspects of molecular constitution, reaction, and properties are explainable in terms of graph theoretical formalism.

In particular, during the past few years, numerous graph-theoretic methods have been developed for the analysis and prediction of physicochemical, environmental, and biomedical properties of molecules. The constitutional formula of a molecule is, in essence, a planar graph where vertices represent the atoms and edges are covalent bonds. Since such a graph adequately depicts the topology of the molecule, it is not surprising to see the success of graph-theoretic approaches in explaining physical and biological properties of diverse group of chemicals.

It is essential to choose the proper molecular descriptors in order to develop quantitative structure property (QSPR) [1] models with strong predictive capability. A molecular descriptor attempts to fully mathematically describe a molecular structure or a particular property of the structure. Topological indices are most widely used molecular descriptor. TIs are referred to as graph invariants as they are formulated using graph theory ideas.

The QSPR, which connects structural features to physicochemical properties, is a useful part of drug design and discovery. The ultimate goal of the QSPR formulation is to develop mathematical models that estimate the physicochemical properties of molecular structures.There are over 3000 TIs in the literature, therefore one must decide how to pick those that best describe the physicochemical property being studied.

Principal component analysis (PCA) [2,3] is a popular method for multivariate data analysis and data reduction. It is a practical algebraic tool for drug design and discovery. A QSPR study often uses PCA to analyse an initial data matrix in which molecules are defined by a number of interconnected quantitative dependent variables. The researchers confront problems such as the elimination of irrelevant information in the original descriptors matrix involved, the unfavourable ratio of the number of descriptors to that of molecules of interest, and collinearity among descriptors used. PCA is a method that can be useful in dealing with the problem of the unfavourable more descriptor ratio and collinearity.

PCA is generally employed when there are numerous data sets and difficult to interpret and when it is difficult to discover and visualize complex correlations between independent variables. Such type of conditions occurs in QSPR study when large number of topological indices are calculated. After applying PCA, new variables are generated and are called principal components (PCs). PCs are linear combinations of the original descriptors that have been sorted to best account for the variance of the original data. Every PC is independent of and orthogonal to every other PC. One of the most significant characteristics of PCs is that they individually include unique information about the variability of the original descriptors. Usually, after the first few PCs, the information on the other PCs is relatively useless as they don't have much information left. Thus, we often take into account PCs with eigen values greater than 1. As a result, each PC is an independent source of variability from the others in the original descriptor matrix.

2. *Methodology*

In this paper we consider 42 molecules of Alkylbenzenes with their physicochemical properties viz boiling point (BP), critical temperature (T_c), critical pressure (P_c) and critical volume (V_c), which are taken from [4]. To develop the QSPR model of the properties of the molecule we take 37 TIs based on degree and neighborhood degree of the vertices. The degree-based topological indices are as follows: Redefined first Zagreb index ($ReZG_1$) [5], redefined second Zagreb index ($ReZG_2$) [5], General Sum-Connectivity index for $\alpha = -1$ ($^{n}M_1$) [6], Reduced Reciprocal Randić index (RRR) [6] and Sombor index (SO) [6], Geometric Arithmetic index (GA) [6], Sum-Connectivity index (SCI) [6], Randić index (R) [6], Atom-Bond connectivity index (ABC) [6], Reciprocal Randić index (RR) [6], First Gourava index (GO_1) [7], General Randić index for $\alpha = -1, -2$ (R_1, R_2) [8], first Zagreb index (M_1)[8], second Zagreb index (M_2) [8], Forgotten index (F) [8], Harmonic index (H) [8], Symmetric Division Deg index (SDD) [8], Redefined third Zagreb index ($ReZG_3$) [8], Hyper Zagreb index (HM) [9], Third Zagreb index (ZG_3) [9], The neighborhood degree-based indices are neighborhood second Zagreb index (NM_2) [10], neighborhood forgotten index (F_N^*) [10], neighborhood fifth hyper first Zagreb index (HM_1G_5) [11], neighborhood fifth geometric arithmetic index (GA_5) [11], neighborhood fourth atom bond connectivity index ABC_4) [11], neighborhood second modified Zagreb index (M_2^{nm}) [11], neighborhood fifth arithmetic geometric index (AG_5) [11], neighborhood fifth

second Hyper Zagreb index (HM_2G_5) [11], ND_1 index [12], ND_2 index [12], ND_3 index [12], ND_4 index [12], ND_5 index [12], neighborhood third version of Zagreb index (NM_1) [13], neighborhood harmonic index (NH) [13], neighborhood ISI index (NI) [13].

Firstly, we obtain the 37 TIs for the 42 alkylbenzenes. The structures of the alkylbenzenes (figure 1) are taken from https://pubchem.ncbi.nlm.nih.gov/. Since the number of independent variables i.e., the TIs are very large there may be chance of high correlation among the indices. And in a regression equation inclusion of large number of TIs may increases the fit but the predictive ability of the developed model will face a substantial decrease.

It is better to select a set of TIs that leads the model with good predictability instead of resorting into indiscriminate use of all TIs. Therefore, we use PCA for data reduction. It is known that PCA is affected by scaling factors. Before applying PCA a normalization or a scaling procedure is applied. So, we transform the TIs by natural logarithm of the index (TI) plus 1($\log_e(TI + 1)$) as the magnitude of the TIs are much larger than those of others. This scaling procedure was introduced by Basak et al. [14]. We use software SPSS to perform PCA and all other analyses.

3. *Results and Discussion*

In PCA the first PC describes the majority of the variability in the data, the second factor, which is orthogonal to the first, describes the majority of the variability left behind by the first factor, and so on. We extract first two principal components (PCs) for further analysis because they have eigen values greater than or equal to 1. The first PC account for 82.121% of the variance and the second PC account for 16.162% of the variance. Which is a total of 98.284% of the total variance. See table 1. The PC scores are obtained by regression factor method. We have also used Anderson-Rubin method to obtain the PC scores but PC scores obtain from regression factor method have better correlation with the properties than the PC scores obtained from Anderson-Rubin. The PC scores are given in table 2.

The PC scores are well correlated with the individual TIs. The 10 most highly correlated topological indices with the PC scores are given in table 3. The PC_1 is highly correlated with the degree-based indices other than ND_5, NI and ND_1. The PC_2 is highly correlated with the neighborhood degree-based indices other than ZG_3, H and R_2 index. So, the data in this study we have taken, its properties are largely described by the degree-based indices than the neighborhood degree-based indices. Also, in table 4 we show the correlation of the PC scores with the properties of alkylbenzenes.

The relationship between each of the 37 TIs is shown in figure 2 simultaneously. Indices that contribute similar information are grouped together i.e., they are correlated. All of the indices in this instance have a positive correlation because of their placement on the same side of the plot origin. The other indices frequently alter in the same way when one index rises or

falls. Moreover, the distance from the origin reveals information. An index's impact on the model increases with its distance from the plot origin.

3.1 Regression model for alkylbenzenes

The multiple linear regression equation is as follows:

$$P = m_0 + m_1X_1 + m_2X_2 + \cdots + m_nX_n,$$

where P is the predicted value, $X_1, \ldots, X_n$ are n distinct predictor variables, $m_1, \ldots, m_n$ are regression coefficient and m_0 is constant.

We use multilinear regression model using the two PC scores as the independent parameters and properties as dependent parameters to predict the properties of alkylbenzenes. The two-parameter model contains PC_1 and PC_2. The regression equation for boiling point, P_c, V_c and T_c are as follows:

$$BP = 31.663PC_1 + 3.235PC_2 + 460.331$$

$$P_c = -0.459PC_1 - 0.050PC_2 + 2.911$$

$$V_c = 73.7PC_1 + 22.913PC_2 + 489.598$$

$$T_c = 29.235PC_1 + 3.394PC_2 + 663.123$$

The predicted properties of alkylbenzenes are given in table 5-8.

4. Degeneracy

A TI seeks to decode the structural property as completely as possible. A good TI should be able to distinguish between various structural formulae. Most TIs are degenerate, meaning that two or more structures share the same TIs. High-discriminating topological indices are able to collect more structural data. We use the sensitivity, measure of degeneracy developed by Konstantinova [15], which is defined as follows.

$$S_I = \frac{N - N_1}{N},$$

Where N is the total number of structures considered and N_1 is the number of them that cannot be distinguish by the TI I. As, $S_I$increases, the discrimination power of topological power increases. Here we have checked the sensitivity of the PC scores. As we have used the PC scores extracted from the TIs for the regression analysis to predict the properties of alkylbenzenes. The sensitivity of the PC scores is 0.952. See table 9. So, the PC scores are high discriminating and collects very good amount of structural data.

5. *Application*

Using QSPR analysis a molecule can be decomposed into a set of numerical values that describe its relevant physicochemical properties and biological activities. The strongest correlated descriptors in this study provide information on the compounds under consideration. Wet lab testing of a molecule is often quite expensive, but the QSPR study allows for a reduction in cost with no risk. There are over 3000 TIs in graph theory. It is not easy to apply QSPR analysis to model the regression equation for prediction of properties of some chemical structure with all the TIs individually. When there is a problem in multiple linear regression analysis where the results don't fit the facts, multicollinearities among the independent variable are generally expected. We can then employ the PCA at that point. SPSS's principal component regression analysis is a useful technique. It not only identifies collinearity for each independent variable, but also deals with the problem of collinearity. So, applying PCA is the best method as they will reduce the dimension without losing the original data of the variables. Also, it eliminates the problems of multicollinearity among the variables and hence provides a good predictive model. Due to its applicability, we can take large number of TIs for prediction of chemical structures without losing the original data and can get a good predictive model for the physicochemical properties of chemical structures.

6. *Conclusion*

In this article we have considered 37 topological TIs for the QSPR analysis of 42 Alkylbenzenes. Due to the size of the TIs, there is a possibility of high correlation. Including several topological indices in a regression equation may improve fit, the generated model's ability for prediction will suffer significantly. So, PCA is applied to eliminate the multicollinearity issue. Initially we have obtained the topological indices and then normalize the values using the method introduced by Basak before applying PCA. It is necessary as PCA is affected by scaling factors. We also tried the standardize method to normalize the indices but Basaks methods is found to be more effective. We use PC_1 and PC_2 scores to perform the regression model for the prediction. Both the PC accounts for 98.284% of the total variance. We could have chosen more PCs but there isn't much information left. In this analysis we get 10 PCs that accounts for 100% of the total variance. There will be no data loss if we consider all those 10 PCs. The PC scores are very well correlated topological indices. Among them we have chosen top 10 topological indices which are highly correlated with the PC scores. Those indices are RR, $ReZG_2$, ND_5, M_1, ABC, SO, GO_1, NI, $ReZG_1$, ND_1. Using the regression models, we predict the properties of the drugs. We also check the sensitivity of the PC scores and they are high discriminating (0.952) i.e., they can collect very good amount of structural data. The predicted data are found to be reasonably well and hence PCA is an effective analysis to be considered for prediction when there is a large number of TIs are considered for multilinear regression model. At the same time, most computing tasks are carried out with a computer, considerably

reducing the need for a labour-intensive manual and finally, we get a simpler, accelerated, and accurate statistical effect results. In future researcher can choose the above 10 indices to perform QSPR/QSAR analysis for any chemical compounds over the considered TIs.

7. *Conflict of Interest :* The authors declare no conflict of interest.

References

[1] A. Bharali, A.Pegu, P.Agarwal, and M.Chamua, *On M-Polynomial and some topological indices of Favipiravir (T-705) and Ribavirin*,J. Discrete Math. Sci. Cryptogr , 24(2021), pp. 2121-2135.

[2] R. X.Liu, J.Kuang, Q.Gong, and X. L.Hou, *Principal component regression analysis with spss*, Comput. Methods and Programs Biomed. 71 (2003), pp. 141-147.

[3] R.Nataranjan, P.Kamalakanan, and I.Nirdosh, *Application of topological indices to structure-activity relationship modelling and selection of mineral collectors*, IJC, 42A (2003),pp. 1330-1346.

[4] S.Yang, W.Lu, N.Chen, and Q.Hu, *Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes,J. Mol. Struct: THE CHEM*. 719 (2005),pp. 119-127,

[5] W.Gao, M. R.Farahani, M. K.Jamil, and M. K.Siddiqui, *The Redefined First, Second and Third Zagreb Indices of Titania Nanotubes [206],*Open BiotechnolJ, 10 (2016),pp. 272-277,

[6] M. Y. H.Malik, M. A.Binyamin, and S.Hayat,*Correlation Ability of Degree-Based Topological Indices for Physicochemical Properties of Polycyclic Aromatic hydrocarbon with Applications*, Polycycl Aromat Compd. (2021),

[7] VR Kulli, *The Gourava Indices and Coindices of Graphs*, Annal of Pure and Applied Mathematics, (2017),

[8] M.Chamua, R.Moran, A.Pegu, and A.Bharali,M-polynomial and neighborhood M-polynomial of some concise drug structures: Azacitidine, Decitabine and Guadecitabine, J. Mol. Struct. 1263 (2022), pp. 1333197,

[9] M.Adnan, S. A. U. H. Bokhary, G.Abbas, and T.Iqbal, *Degree-Based Topological Indices and QSPR Analysis of Antituberculosis Drugs,*J. Chem.,(2022),

[10] S.Mondal, N. De, and A. Pal, *On Some New Neighbourhood Degree Based Indices*, Acta Chemica lasi. 27(2019), pp. 31-46,

[11] A. Ahzal, S. Hameed, U. Ashraf, A Mehmood, F Chaudhry, and DK Thapa, *Study of Neighborhood Degree-Based Topological Indices via Direct and NM-Polynomial of Starphene Graph,*J. Funct. Spaces, (2022)

[12]S.Mondal, A. Dey, N.De, and A. Pal,*QSPR analysis of some novel neighborhood degree-based topological descriptors*, Complex & Intell Syst. 7 (2021), pp. 977-996,

[13]S. A. K.Kirmani, P.Ali, and F Azam, *Topological indices and QSPR/QSAR analysis of some antiviral drugs being investigated for the treatment of Covid-19 patients*, Int J Quantum Chem. (2020),

[14] S. C.Basak, V. R. Magnuson, G J Niemi, R R Regal, and G D Veith, *Topological Indices: Their Mutual Relatedness and Applications*, Mathematical Modelling, 8 (1987), pp. 300-305.

[15]E. V.Konstantinova, *The discrimination ability of some topological and information distance indices for graphs of unbranched hexagonal systems*, J Chem Inf Comput Sci. 36 (1996), pp. 54-57.

Tables

Table 1. PCA of Alkylbenzenes

PCA for Alkylbenzenes			
PC	Eigen values	Percent of variance	Cumulative variance
1	30.385	82.121	82.121
2	5.980	16.162	98.284

Table 2. PC Scores

Compound	PC_1	PC_2	Compound	PC_1	PC_2
Benzene	-3.71462	1.05698	n-Octyl benzene	1.25426	2.42579
n-Propyl benzene	-0.90753	0.90399	1-methyl-4-ethyl benzene	-0.73845	-0.07023
1,3,5-Trimethylbenzene	-0.66305	-0.91475	1,3-Diethylbenzene	-0.02741	0.41456
1-Methyl-3-propylbenzene	-0.11167	0.45472	1,2-Dimethyl-4-ethylbenzene	0.07116	-0.44223
1,3 Dimethyl-5-ethylbenzene	0.03474	-0.36408	1,2,4,5-Tetramethylbenzene	0.15852	-1.27561

1,4-Diisopropylbenzene	1.17955	-0.23039	n-Pentyl benzene	0.25914	1.71179
1-Methyl-2-isopropylbenzene	0.04976	-0.54133	sec-Butylbenzene	-0.07136	0.33695
Toluene	-2.48913	-0.0309	n-Heptyl benzene	1.25426	2.42579
1-Methyl-2-ethylbenzene	-0.66253	-0.14129	1,2,3-Trimethylbenzene	-0.55947	-1.08025
n-Butylbenzene	-0.29766	1.32122	1,4-Diethylbenzene	-0.04127	0.42011
1-Methyl-4-propylbenzene	-0.12227	0.45511	1,3-Dimethyl-2-ethylbenzene	0.14316	-0.50059
1,4-Dimethyl-2-ethylbenzene	0.08479	-0.44349	1,2-Diisopropylbenzene	1.27299	-0.29635
1,2,4-Trimethylbenzene	-0.61701	-1.01151	n-Hexylbenzene	0.7745	2.07905
1-Methyl-3-isopropylbenzene	-0.01467	-0.47982	tert-Butylbenzene	0.09414	-1.12135
Ethylbenzene	-1.59969	0.41294	Hexamethylbenzene	1.54729	-1.60398
1-Methyl-3-ethylbenzene	-0.72688	-0.07086	1,2,4-Triethylbenzene	1.29421	0.57009
1,2-Diethylbenzene	0.04214	0.3547	1-Methyl-2-propylbenzene	-0.0497	0.38898
1,2-Dimethyl-3-ethylbenzene	0.13466	-0.49967	1,3-Dimethyl-4-ethylbenzene	0.08149	-0.43699
1,2,3,5-	0.16583	-1.26727	1,3-Diisopropylbenzene	1.19561	-0.2374

Tetramethylbenzene					
1,3,5-Triethyl benzene	1.24902	0.62122	1-Methyl-4-isopropylbenzene	-0.02811	-0.47634
Pentamethyl benzene	0.8938	-1.48273	1,2,3,4-Tetramethylbenzene	0.20747	-1.33457

Table 3. Correlation between the PC scores and TIs

PC_1		PC_2	
TI	R	TI	R
RR	0.987	M_2^{nm}	0.867
$ReZG_2$	0.984	ZG_3	-0.745
ND_5	0.983	NH	0.716
M_1	0.980	ND_4	0.684
ABC	0.973	HM_2G_5	-0.584
SO	0.967	ND_3	-0.550
GO_1	0.967	NM_1	0.539
NI	0.966	H	0.539
$ReZG_1$	0.961	ND_2	0.515
ND_1	0.958	R_2	0.509

Table 4. Regression analysis of the PCs

Regression analysis of PC_1 and PC_2 with the properties of Alkylbenzenes				
Properties	R	R^2	Adjusted R^2	SEE
BP	0.939	0.881	0.875	11.984
P_c	0.976	0.953	0.950	0.105
V_c	0.977	0.955	0.953	17.146
T_c	0.889	0.790	0.779	15.577

Table 5. Experimented and Predicted BP/K

Compound	Expt BP	Predict BP	Compound	Expt BP	Predict BP
Benzene	353.25	346.18	n-Octyl benzene	537.6	507.94
n-Propyl benzene	432.37	434.48	1-methyl-4-ethyl benzene	435.16	436.71
1,3,5-Trimethylbenzene	437.87	436.38	1,3-Diethylbenzene	454.29	460.76
1-Methyl-3-propylbenzene	455.16	458.24	1,2-Dimethyl-4-ethylbenzene	463.15	461.15
1,3 Dimethyl-5-ethylbenzene	456.77	460.23	1,2,4,5-Tetramethylbenzene	469.99	461.24
1,4-Diisopropylbenzene	483.5	496.96	n-Pentyl benzene	478.17	474.07

1-Methyl-2-isopropylbenzene	451.33	460.15	sec-Butylbenzene	446.45	459.15
Toluene	383.77	381.41	n-Heptyl benzene	519.2	507.94
1-Methyl-2-ethylbenzene	438.33	438.88	1,2,3-Trimethylbenzene	449.23	439.15
n-Butylbenzene	456.42	455.16	1,4-Diethylbenzene	456.94	460.36
1-Methyl-4-propylbenzene	456.57	457.92	1,3-Dimethyl-2-ethylbenzene	463.2	463.23
1,4-Dimethyl-2-ethylbenzene	460.02	461.57	1,2-Diisopropylbenzene	477	499.65
1,2,4-Trimethylbenzene	442.5	437.54	n-Hexylbenzene	499.2	491.60
1-Methyl-3-isopropylbenzene	448.23	458.31	tert-Butylbenzene	442.27	459.91
Ethylbenzene	409.34	410.97	Hexamethylbenzene	538	504.13
1-Methyl-3-ethylbenzene	434.48	437.06	1,2,4-Triethylbenzene	490.7	503.13
1,2-Diethylbenzene	456.61	462.77	1-Methyl-2-propylbenzene	458.12	460

1,2-Dimethyl-3-ethylbenzene	467.1	462.98	1,3-Dimethyl-4-ethylbenzene	461.39	461.49
1,2,3,5-Tetramethylbenzene	471.19	461.49	1,3-Diisopropylbenzene	476.4	497.42
1,3,5-Triethyl benzene	489.1	501.85	1-Methyl-4-isopropylbenzene	450.28	457.91
Pentamethyl benzene	505	483.84	1,2,3,4-Tetramethylbenzene	478.24	462.60

Table 6. Experimented and Predicted Pc/MPa

Compound	Expt P_c	Predict P_c	Compound	Expt P_c	Predict P_c
Benzene	4.9	4.56	n-Octyl benzene	2.26	2.21
n-Propyl benzene	3.2	3.28	1-methyl-4-ethyl benzene	3.15	3.25
1,3,5-Trimethylbenzene	3.13	3.26	1,3-Diethylbenzene	2.8	2.90
1-Methyl-3-propylbenzene	2.89	2.94	1,2-Dimethyl-4-ethylbenzene	2.86	2.90
1,3 Dimethyl-5-ethylbenzene	2.85	2.91	1,2,4,5-Tetramethylbenzene	2.91	2.90
1,4-Diisopropylbenzene	2.45	2.38	n-Pentyl benzene	2.72	2.71

1-Methyl-2-isopropylbenzene	2.82	2.92	sec-Butylbenzene	2.95	2.93
Toluene	4.11	4.05	n-Heptyl benzene	2.39	2.21
1-Methyl-2-ethylbenzene	3.15	3.22	1,2,3-Trimethylbenzene	3.45	3.22
n-Butylbenzene	2.89	2.98	1,4-Diethylbenzene	2.8	2.91
1-Methyl-4-propylbenzene	2.89	2.94	1,3-Dimethyl-2-ethylbenzene	2.86	2.87
1,4-Dimethyl-2-ethylbenzene	2.85	2.89	1,2-Diisopropylbenzene	2.44	2.34
1,2,4-Trimethylbenzene	3.23	3.24	n-Hexylbenzene	2.39	2.45
1-Methyl-3-isopropylbenzene	2.82	2.94	tert-Butylbenzene	2.97	2.92
Ethylbenzene	3.61	3.62	Hexamethylbenzene	2.49	2.28
1-Methyl-3-ethylbenzene	3.15	3.25	1,2,4-Triethylbenzene	2.36	2.29

1,2-Diethylbenzene	2.81	2.87	1-Methyl-2-propylbenzene	2.89	2.91
1,2-Dimethyl-3-ethylbenzene	2.86	2.87	1,3-Dimethyl-4-ethylbenzene	2.85	2.90
1,2,3,5-Tetramethylbenzene	2.91	2.90	1,3-Diisopropylbenzene	2.45	2.38
1,3,5-Triethyl benzene	2.36	2.31	1-Methyl-4-isopropylbenzene	2.82	2.95
Pentamethyl benzene	2.67	2.56	1,2,3,4-Tetramethylbenzene	2.92	2.88

Table 7. Experimented and Predicted V_c/cm^3mol^{-1}

Compound	Expt V_c	Predict V_c	Compound	ExptV_c	Predict V_c
Benzene	258.66	240.04	n-Octyl benzene	697.5	637.63
n-Propyl benzene	440.27	443.38	1-methyl-4-ethyl benzene	429.8	433.55
1,3,5-Trimethylbenzene	429.35	419.79	1,3-Diethylbenzene	481.9	497.03
1-Methyl-3-propylbenzene	482.3	491.77	1,2-Dimethyl-4-ethylbenzene	477.5	484.71
1,3 Dimethyl-5-ethylbenzene	477.5	483.81	1,2,4,5-Tetramethylbenzene	473.1	472.08
1,4-	569.8	571.28	n-Pentyl benzene	539.8	547.90

Diisopropylbenzene					
1-Methyl-2-isopropylbenzene	495.7	480.86	sec-Butylbenzene	477.65	492.04
Toluene	315.55	305.42	n-Heptyl benzene	644.9	637.63
1-Methyl-2-ethylbenzene	429.8	437.51	1,2,3-Trimethylbenzene	429.26	423.64
n-Butylbenzene	497.12	497.90	1,4-Diethylbenzene	481.9	496.16
1-Methyl-4-propylbenzene	482.3	491	1,3-Dimethyl-2-ethylbenzene	477.5	488.67
1,4-Dimethyl-2-ethylbenzene	477.2	485.68	1,2-Diisopropylbenzene	569.8	576.61
1,2,4-Trimethylbenzene	429.26	420.97	n-Hexylbenzene	592.3	594.31
1-Methyl-3-isopropylbenzene	495.7	477.53	tert-Butylbenzene	461.24	471.06
Ethylbenzene	373.84	381.11	Hexamethylbenzene	568.5	566.90
1-Methyl-3-ethylbenzene	429.8	434.38	1,2,4-Triethylbenzene	581.8	598.02

1,2-Diethylbenzene	481.9	500.79	1-Methyl-2-propylbenzene	482.3	494.83
1,2-Dimethyl-3-ethylbenzene	477.5	488.08	1,3-Dimethyl-4-ethylbenzene	477.5	485.59
1,2,3,5-Tetramethylbenzene	473.1	472.81	1,3-Diisopropylbenzene	569.8	572.29
1,3,5-Triethyl benzene	581.8	595.85	1-Methyl-4-isopropylbenzene	495.7	476.63
Pentamethyl benzene	581.8	521.12	1,2,3,4-Tetramethylbenzene	473.1	474.34

Table 8. Experimented and Predicted T_c/K

Compound	Expt T_c	Predict T_c	Compound	Expt T_c	Predict T_c
Benzene	562.09	558.12	n-Octyl benzene	742.95	708.04
n-Propyl benzene	638.3	639.61	1-methyl-4-ethyl benzene	639.28	641.28
1,3,5-Trimethylbenzene	637.28	640.64	1,3-Diethylbenzene	653.28	663.68
1-Methyl-3-propylbenzene	657.22	661.37	1,2-Dimethyl-4-ethylbenzene	664.31	663.71
1,3 Dimethyl-5-ethylbenzene	655.84	662.89	1,2,4,5-Tetramethylbenzene	675.7	663.46

1,4-Diisopropylbenzene	678.81	696.86	n-Pentyl benzene	683.7	676.49
1-Methyl-2-isopropylbenzene	652.54	662.74	sec-Butylbenzene	664.15	662.17
Toluene	591.72	590.23	n-Heptyl benzene	724.17	708.04
1-Methyl-2-ethylbenzene	644.09	643.26	1,2,3-Trimethylbenzene	664.45	643.14
n-Butylbenzene	660.45	658.87	1,4-Diethylbenzene	657.88	663.31
1-Methyl-4-propylbenzene	659.26	661.08	1,3-Dimethyl-2-ethylbenzene	665.07	665.61
1,4-Dimethyl-2-ethylbenzene	660.51	664.90	1,2-Diisopropylbenzene	669.96	699.31
1,2,4-Trimethylbenzene	649.05	641.68	n-Hexylbenzene	704.21	692.82
1-Methyl-3-isopropylbenzene	648.05	661.07	tert-Butylbenzene	660.15	662.29
Ethylbenzene	617.09	617.70	Hexamethylbenzene	758	702.94
1-Methyl-3-ethylbenzene	638.28	641.60	1,2,4-Triethylbenzene	682.78	702.88

1,2-Diethylbenzene	656.62	665.52	1-Methyl-2-propylbenzene	661.5	662.98
1,2-Dimethyl-3-ethylbenzene	670.67	665.37	1,3-Dimethyl-4-ethylbenzene	662.47	664.02
1,2,3,5-Tetramethylbenzene	662.2	663.69	1,3-Diisopropylbenzene	668.84	697.28
1,3,5-Triethyl benzene	680.41	701.71	1-Methyl-4-isopropylbenzene	651.02	660.70
Pentamethyl benzene	691.2	684.28	1,2,3,4-Tetramethylbenzene	685.63	664.69

Table 9. Measure of sensitivity of PC scores for Alkylbenzenes

PC Scores	Sensitivity (S_l)
	Alkylbenzenes
PC_1	0.952
PC_2	0.952

Figures

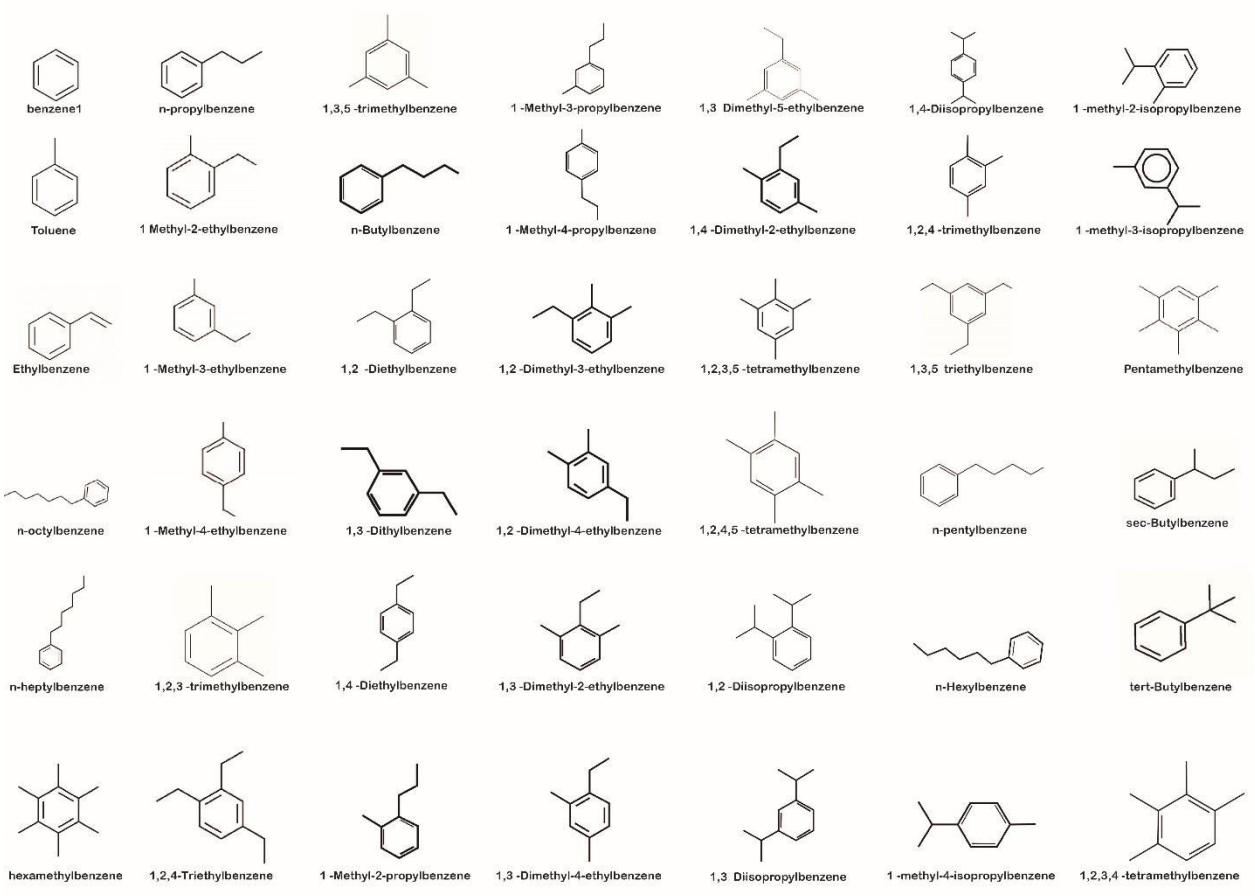

Figure 1. Alkylbenzenes

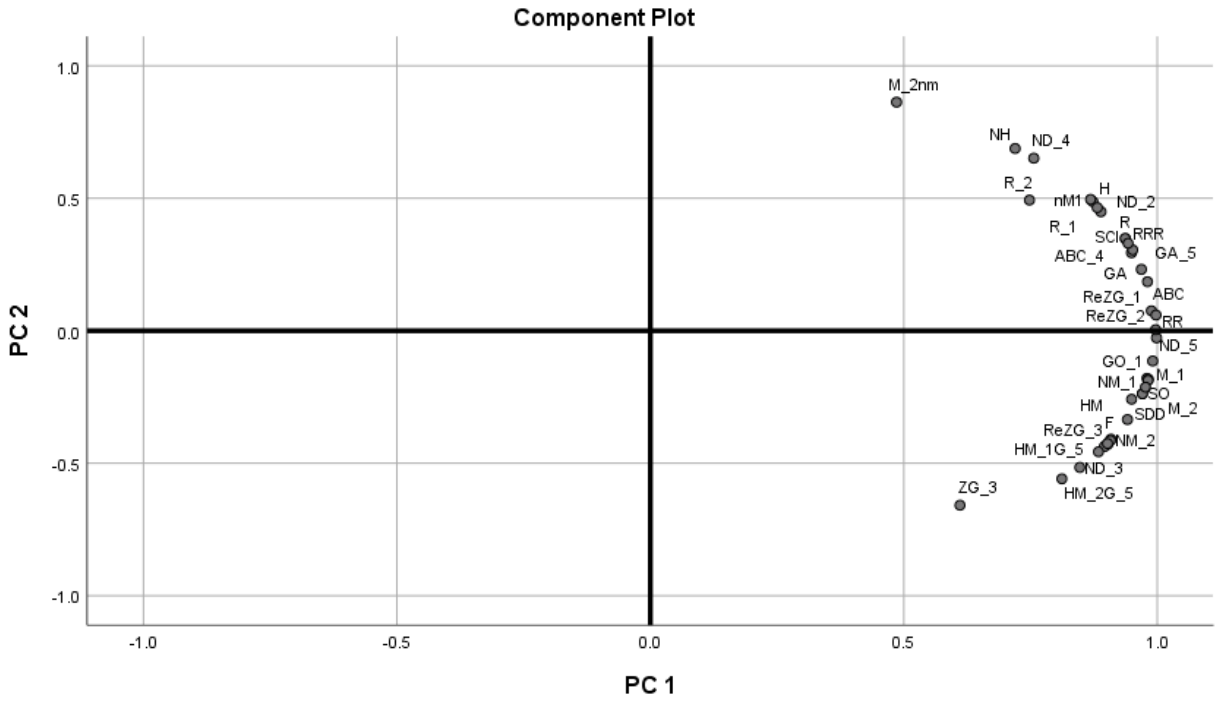

Figure 2. Component plot of the two PCs