Journal of the Serbian Chemical Society

JSCS-info@shd.org.rs • www.shd.org.rs/JSCS

J. Serb. Chem. Soc. *89* (4) 485–503 (2024)
JSCS–5735

Original scientific paper
Published 15 April 2024

Multivariate statistical analysis approach to investigate the thermodynamic quantities of the benign alternative fuel

KASSIO FELIPE DA COSTA SERRA[1], ALAMZEB KHAN[2], RAQUEL MARIA TRINDADE FERNANDES[3], PEDRO ANTONIO MUNIZ VAZQUEZ[4] and ALAMGIR KHAN[3]*

[1]Programa de Pós-Graduação Engenharia Aeroespacial, Universidade Estadual do Maranhão – UEMA, Cidade Universitária Paulo VI, Campus São Luís/MA, CEP 65000-00, Brasil, [2]Laboratório de Ressonâ ncia Paramagnética Eletrônica (LARPE), Departamento de Física, Universidade Estadual de Londrina (UEL), Londrina, PR, Brasil, [3]Departamento de Química, Centro de Educação, Ciências Exatas e Naturais – CECEN, Universidade Estadual do Maranhão – UEMA, Cidade Universitária Paulo VI, Campus São Luís/MA, CEP 65000-00, Brasil and [4]Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas – UNICAMP, Caixa Postal 6154, Campinas, SP, 13083-970, Brasil

(Received 30 May, revised 29 August, accepted 20 November 2023)

Abstract: In order to extract meaningful interpretation from the large data and provide their value to the application areas, chemical data analysis has become a serious challenge in the development and applications of new protocols, technique and methodologies for the mathematical modelling communities and other data science societies. Therefore, in the present work a rapid and robust box-and-whisker plot and multivariate principal component statistical techniques (PCA) are being proposed for the evaluations of the thermodynamic molecular properties data of the benign fuel structures. We observed that, the box-and-whisker plot technique successfully explored all of the thermochemical molecular properties precisely, and described symmetrical distribution of the data along the median values with respect to the rise in temperature. Moreover, applying the PCA technique, the score-plots of PCs diagnosed the peculiar molecular properties variations after a certain peak of temperature with descendant variation in the statistical parameters. Furthermore, PCA parameters not only segregated the thermodynamic properties of propanol and butanol but also, their variations with the temperature. Thus, we concluded that, Box-whisker and PCA statistical techniques are robust and rapid method for the assessment and evaluation of the large molecular thermodynamic quantities data.

Keywords: computational study; benign fuels; statistical analysis; DFT; thermodynamic quantities.

* Corresponding author. E-mail: alamgir@cecen.uema.br

https://doi.org/10.2298/JSC230530090S

INTRODUCTION

Over the last few decades, Researchers have been testing some novel fuels by integrating biofuels and conventional fuels, taking into consideration the greenhouse gas emissions as well as calorific capacity (improved octane number and cetane number).[1] In this regard, efforts are being taken to develop alternative and renewable fuels that can produce useful energy while reducing global warming and, as a corollary, environmental pollution.[2] The blending of ethanol–gasoline, butanol–gasoline, butanol–diesel, butanol–biodiesel, diesel–biodiesel, kerosene–biokerosene have been tested.[3–5] Butanol is considered to be one of the most promising biofuels with several advantages over bioethanol and has been utilized in large-scale processes in recent decades.[6] In addition, studies demonstrate that butanol is far more useful than ethanol, with a higher calorific value (29.2 MJ/dm^3), lower heat of vaporization (0.43 MJ/kg), and less corrosive properties. Furthermore, it has been discovered that the energy content per volume unit of butanol is comparable to that of gasoline and higher than that of ethanol.[7] Because of butanol's specific mass and viscosity similarity to those of diesel, it has good solubility in heavier hydrocarbons as well, so it can be added to diesel in higher proportions.[8]

Since conventional investigations involved expensive and time-consuming experimental trial and error technique to determine the energy storage or transformation of innovative materials.[9] As a result, the scientific community is increasingly supporting the development of atomistic modelling methodologies for use in energetic materials (EM) research and development programs. Where, the change of heat of formation ($\Delta_f H^\circ$) is regarded to be a significant property in predicting the performance during designing and manufacturing of the new energetic materials in technological applications.[10] Moreover, an energetic molecule's $\Delta_f H^\circ$ value should be as high as possible, as this ensures that the chemical material will be unstable when disintegrated into its constituent parts in their standard forms.

Therefore, it appears that a high-density material with such a high heat of formation would be a suitable candidate to be used as a fuel. In this regard, a range of computational approaches based on semi-empirical, ab initio, and Density functional theory (DFT methodologies) have become powerful tools for predicting the structures and thermochemical molecular characteristics of materials in recent decades.[10,11] Aside from measuring molecular characteristics, a trustworthy and efficient analysis of chemical data is also believed essential. Whenever a massive quantity of data is involved in a short period of time, evaluation is becoming a challenge for the researcher. To address this problem, certain systematic approaches for big data analysis projects are being developed. Furthermore, several analysis features such as data analysis, data preprocessing strategy development, visualization of data, and validation of the model are all taken into account. To extract useful analytical information from spectral data, several multivariate statistical analysis (MVA) techniques are proposed.[12,13] Typically, these methods use two- or three-dimensional graphs to facilitate multivariate understanding of complex data sets. Pattern recognition, classification and multivariate calibration may all be done with a simple MVA of PCA approach. This statistical method is frequently used to reduce the size of a data set, detect sample similarity, display data structure and find outliers (abnormal samples).[12] Another MVA calibration method is partial least square (PLS), which uses the principal component analysis technique to reduce the dimensionality of the data set in preparation for subsequent spatial correlations.[12]

Therefore, in the present work a robust statistical multivariate PCA technique is presented to explored and diagnosed the thermochemical properties differences with respect to temperature (5 to 1500 K), in between the isomers of butanol and propanol fuels additives, based on the small structural effects. The thermodynamic data (*i.e.*, enthalpy change of formation, specific isobaric heat capacity, entropy change of formation, and Gibbs energy change of formation) were computationally predicted for the gaseous state of different chemical structures (isomers of propanol and butanol) used as benign fuel (oxygenated and environmental friendly), using DFT and semi-empirical method. The quality of the thermodynamic quantities is being verified by comparing with literature values using root-mean square error (*RMSE*) and Box-and-whisker technique.

EXPERIMENTAL

In this study, Becke 3-parameter hybrid functional combined with the gradient-correlation functional of Lee–Yang–Parr (B3LYP)[14] and pure functional of Perdew, Burke and Ernzerhof as made into a hybrid by Adamo (PBE0)[15] were tested in the computations using DFTs. All electron Dunnings correlation-consistent (triple zeta) polarized basis set (cc-pVTZ)[16] was employed. A semi-empirical method at the PM6 Hamiltonian was also employed, as implemented in Gaussian.[17] Ground state equilibrium geometry, hessian matrix and normal modes of the vibrations were computed for the gas state 1-propanol, 2-propanol, 1-butanol, 2-butanol and *tert*-butanol molecules, using the afore mentioned methods. The thermodynamic properties like enthalpy, eentropy, Gibbs energies and specific isobaric heat values were extracted at different temperatures from the normal-modes computational Gaussian-output-files. All the computations were made at different temperatures of 5, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400 and 1500 K. All the computations were performed using the electronic structure program Gaussian09[18] and Statsoft Statistica.[18]

Statistical analysis of data

The acquired data have undergone a thorough descriptive statistical analysis,[19,20] including the Shapiro–Wilk normality test, mean, standard deviation, variance and coefficient of variation. Root mean-square error (*RMSE*) of the data has also been calculated between the thermochemical properties (*i.e.*, enthalpy change of formation, specific heat at constant pressure, entropy change of formation and Gibbs energy change of formation) for the target molecules in order to investigate a comparison in between the expected results obtained with our proposed computational methods compared to those of the literature.[21-24] The *RMSE* were calculated for the data set as:[25]

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n}} \tag{1}$$

Where, e_i represents the error in between the computed obtained values and the literature values,[21-24] and *n* represents the observation of each parameter at the given temperature.

Before doing the Principal PCA on the data, the data distribution of the computational values were also examined using a Box-plot[26] (skewness) of the entire data set and as well as for the individual property. The Box-whisker-plot also helped to understand the symmetry distribution in the thermodynamic quantities data-set. Multivariate PCA was performed to reduce the dimensionality problem and possible exploratory analysis of the molecular structures, temperature dependence and proposed theoretical methods used in the computation of thermodynamic properties. PCA was performed for the whole data set, including the enthalpies, entropies, Gibbs energies and specific heat values at constant pressure, considering the proposed theoretical protocols by using cc-pVTZ basis functions at DFT level of theory (*i.e.*, B3LYP & PBE0 functionals) and PM6 semi-empirical method. The "Statistica" a Data Analysis Software System[18], version 8, was used for performing PCA and other basic statistical analysis.

RESULT AND DISCUSSION

In this section, the results of the computations for the thermodynamic properties (*i.e.*, specific isobaric heat capacity (*Cp*), enthalpy (*H*), entropy (*S*) and Gibbs energy), using cc-pVTZ basis set at DFT functionals (*i.e.*, PBE0 and B3LYP functional) level along with semi-empirical method, for a set of 1-propanol, 2-propanol, 1-butanol, 2-butanol and *tert*-butanol molecules is presented. The computed thermodynamic molecular properties were validated by comparing theoretically predicted values against experimental values collected from the literature.[21–24]

Descriptive statistical analysis

The dispersion or variability of a data set inside a statistical function provides an explanation for the numerical values. Researchers utilize variability to calculate the separation between data points and the distribution's center.[19,20,27] This type of study allows the researcher to investigate the heterogeneity or homogeneity of each data set, while understanding the variability of the different data--sets.[19,20,27] In order to understand the similarities and differences between the results obtained, a thorough descriptive statistical analysis for each parameter has been calculated. Firstly, we tested the statistical dispersion techniques to study the theoretical models. In this regard, to evaluate how well our proposed computationally modeled data fit to a normal distribution, the Shapiro–Wilk normality test[27] approach for the standardized data-set was being performed. The results are presented in Table I and Fig. 1. On the basis of our data distribution results applying the Shapiro–Wilk normality test null hypothesis for the data is to be normally distributed, we can say that our data is not normally distributed because the p value at the 5 % level of significance compared to 0.05 is smaller than the obtained value, as can be seen in the Table I.

TABLE I. Shapiro–Wilk normality variable test utilizing standardized full thermodynamic data set, mean and standard deviation (St. Dev.) calculated for each thermodynamic quantity absolute error in between the obtained the literature values.[21-24] In the acronym Pxy and Bxy, P and B stand for propanol and butanol, respectively. Subscripts such as "*x* = 1, 2 or 3" are used to number the isomers 1-propanol, 2-propanol, 1-butanol, 2-butanol and *tert*-butanol, while "*y* = 1, 2 or 3" stand for the PBE0, B3LYP and PM6 methods, respectively

Sample	Statistics	*P* value	Enthaply		Entrophy		Specific heat		Gibbs energy	
			Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
P11	0.943	0.007	2.60	2.10	2.43	1.19	5.03	3.28	25.70	2.64
P12	0.941	0.006	1.84	1.72	1.50	0.83	2.98	1.62	18.90	1.25
P13	0.941	0.006	2.60	2.10	2.43	1.19	5.03	3.28	25.70	2.64
P21	0.942	0.007	3.12	1.89	3.41	1.90	5.12	2.64	7.42	1.27
P22	0.943	0.007	2.14	1.42	1.96	1.22	5.18	2.49	9.33	0.99
P23	0.940	0.006	3.12	1.89	3.41	1.90	5.12	2.64	7.42	1.27
B11	0.941	0.006	2.49	1.47	2.72	1.50	3.98	1.84	23.49	1.21
B12	0.941	0.006	1.27	0.88	1.08	0.77	3.92	1.96	26.02	0.85
B13	0.941	0.006	2.49	1.47	2.72	1.50	3.98	1.84	23.49	1.21
B21	0.943	0.007	2.95	1.69	3.08	1.84	5.00	2.82	23.21	2.58
B22	0.946	0.010	1.70	1.17	1.87	1.15	4.67	2.39	30.17	8.58
B23	0.940	0.006	2.95	1.69	3.08	1.84	5.00	2.82	23.21	2.58
B31	0.945	0.009	2.94	1.76	3.60	2.09	3.45	1.99	4.22	6.43
B32	0.944	0.009	1.66	1.06	1.61	0.94	3.60	1.84	2.78	6.98
B33	0.942	0.007	2.94	1.76	3.60	2.09	3.45	1.99	4.22	6.43

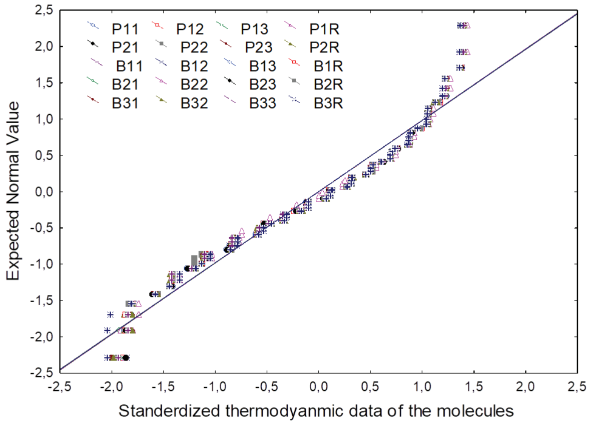

Fig. 1. Shapiro–Wilk normality variable test for the standardized full thermodynamic data set. Where, in the acronyms P*xy* and B*xy*, P and B stand for propanol and butanol, respectively. Subscripts such as "*x* = 1, 2 or 3" are used to number the isomers 1-propanol, 2-propanol, 1-butanol, 2-butanol and *tert*-butanol, while "*y* = 1, 2 or 3" stand for the PBE0, B3LYP and PM6 methods, respectively.

Since we are using the temperature dependence thermodynamic molecular properties data for the different isomers of propanol and butanol molecules, the thermodynamic properties have been steadily increasing and decreasing with the change in temperature. This made our data-set redundant toward normality, as can be seen in Fig. 1. When the Shapiro–Wilk normality test diagram was analyzed, a significant positive skew in the data was shown to be caused by the increase in thermodynamic properties brought on by raising the temperature. Consequently, it appears towards the end of the curve as a tail.

The standard deviation, variance and coefficient variance values were also used to test for statistical dispersion, however we found that the measurement of these statistical distribution techniques, due to the lack of normal distribution, are not appropriate for the current thermodynamic features. Thus, the precision of the results is being estimated, using the absolute error data in between the literature[21–24] and the obtained values, by calculating the mean, standard deviation, variance and coefficient of variation. The results are illustrated in Tables I and II.

TABLE II. Root-mean square error (*RMSE*), Variance (Var.) and percent coefficient of the variance (%*CV*) calculated in between the thermodynamic quantities calculated by using the theoretical data set at computational level to the reference data of NIST web book for the molecules of interests. In the acronym P*xy* and B*xy*, P and B stand for propanol and butanol, respectively. Subscripts such as "*x* = 1, 2 or 3" are used to number the isomers 1-propanol, 2-propanol, 1-butanol, 2-butanol and *tert*-butanol, while "*y* = 1, 2 or 3" stand for the PBE0, B3LYP and PM6 methods, respectively

Sample	Enthalpy			Specific heat			Entropy			Gibbs energy		
	RMSE	Var.	%*CV*	*RMSE*	Var.	%*CV*	*RMSE*	Var.	%*CV*	*RMSE*	Var.	%*CV*
P11	0.84	4.43	80.94	1.81	1.43	49.16	6.34	10.77	65.30	6.17	6.99	10.29
P12	0.66	2.97	93.54	1.17	0.68	55.22	4.53	2.61	54.13	4.53	1.56	6.62
P13	2.15	4.43	80.94	2.98	1.43	49.16	1.84	10.77	65.30	1.17	6.99	10.29
P21	1.18	3.59	60.70	2.35	3.59	55.67	1.45	6.94	51.46	1.80	1.61	17.07
P22	0.70	2.01	66.23	1.90	1.49	62.16	2.18	6.20	48.03	2.24	0.98	10.61
P23	2.07	3.59	60.70	3.60	3.59	55.67	2.12	6.94	51.46	0.74	1.61	17.07
B11	0.97	2.17	59.09	1.83	2.26	55.15	5.05	3.38	46.16	5.62	1.47	5.15
B12	0.38	0.78	69.66	1.29	0.59	71.32	6.27	3.84	49.96	6.22	0.73	3.27
B13	2.10	2.17	59.09	3.34	2.26	55.15	1.56	3.38	46.16	2.06	1.47	5.15
B21	1.01	2.86	57.22	2.11	3.38	59.67	5.37	7.95	56.33	5.58	6.67	11.13
B22	0.49	1.38	69.16	1.55	1.32	61.62	6.58	5.69	51.09	7.48	73.54	28.42
B23	2.13	2.86	57.22	2.15	3.38	59.67	1.94	7.95	56.33	2.52	6.67	11.13
B31	1.22	3.10	59.88	2.01	4.38	58.15	1.74	3.98	57.77	1.79	41.36	152.39
B32	0.55	1.13	63.79	1.37	0.88	58.09	0.45	3.40	51.20	1.74	48.68	251.14
B33	2.42	3.10	59.88	3.62	4.38	58.15	6.01	3.98	57.77	4.31	41.36	152.39

To understand the deviations and spreading in the molecular properties data-set, root mean square error (*RMSE*) a regression analysis bi-variate technique, variance and coefficient of variations of the absolute errors were calculated.

Table II provides an illustration of the calculated values for the *RMSE*, variance (Var.) and percent coefficient of variations (%*CV*).

The *RMSE* identifies the degree to which the data deviates from actual or published values.[21–24] Root mean-square error (*RMSE* or *RMS* deviation) between each computed thermodynamic quantity at DFT levels of theory and semi-empirical technique to that of the reference values for the tested benign energy fuels has been calculated for the sake of comparison and evaluation. The results show that, in comparison to other computational approaches, the semi--empirical method PM6 exhibits some higher variations in accuracy for the Enthalpy change of formation and the specific heat at constant pressure. While, an uncommon case for the *tert*-butanol molecule has also been noted, with some greater deviations for entropy and Gibbs energy. Additionally, after analysis, the largest deviations were found for the enthalpy, specific heat at constant pressure, entropy and Gibbs energy to be 2.42, 3.62, 6.01, and 4.31 kJ/mol, respectively. While the global *RMS* deviation was also being calculated, the thermodynamic characteristics of 2-butanol at the B3LYP level and 1-butanol using PBE0, B3LYP and PM6 approaches exhibit larger deviations when compared to the other molecular systems. The global RMSE for the theoretical approaches shows an overall variance for the PBE0, B3LYP and PM6 of 9.65, 9.80 and 9.65 kJ/mol, respectively. Therefore, it can be inferred that all theoretical methods, precisely and in good accordance with each other, determined thermodynamic properties.

Additionally, Var. and *%CV* of the absolute error between the obtained data and that of the reference data were calculated; the results of these calculations can be seen in Table II. The statistical concept of *%CV* is a statistical technique to measure the dispersion of data points, by displaying the size of standard deviation to its mean values in a data series around the mean.[28]

Due to the consistent increase or decrease of the thermodynamic properties, our standardized data set confirms a rejection of the null hypothesis, which led to certain greater variability and the standard deviation numbers. Therefore, the variance and percent coefficient of variance of the data set to that of the reference data set (*i.e.*, absolute error) were determined in order to evaluate the efficiency of the theoretical approaches and the nature of the acquired data set (as shown in Table II). On analyzing the calculated values, the lowest percent dispersion of the absolute error data observed for the 1-butanol using the B3LYP, PBE0 and PM6 techniques, respectively, was 3.27, 5.15 and 5.15 %. While, using the PBE0, PM6 and B3LYP techniques, the highest *%CV* observed for the *tert*-butanol was 152.39, 152.39 and 251.14 %, respectively. As therefore, the descriptive statistical evaluation of the current investigation has demonstrated that our theoretical methodologies were efficient at acquiring and estimating the molecular characteristics with experimental precision.

Box-Whisker-plot

The distribution of the computational thermodynamic characteristics results acquired, at the DFT levels (*i.e.*, PB3LYP and PBE0) using cc-pVTZ, as well as the reference data was examined by the box-and-whisker plot.[26] The plots of all the distribution data is depicted in the Fig. 2. The analysis of the box-and-whisker diagram not only assisted in detecting the outliers in the data-set, but also provide an overall picture of the data´s typical distribution.

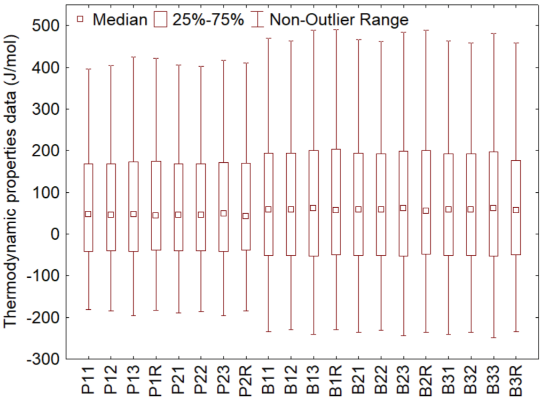

Fig. 2. Box-plot for the different molecular structures used for the computation of complete thermodynamic properties data using DFT and semi-empirical methods. Where, P*xy* and B*xy* stands for P = propanol, B = butanol; *x* = 1 or 2 propanol and 1, 2 or *tert*-butanol; *y* = 1, 2,3 for PBE0, B3LYP and PM6 DFT methods, respectively, while P2R and B3R means literature values[21-24] for 2-propanol and *tert*-butanol.

Since, the thermodynamic data set is constituted of specific heat, Gibbs energy, enthalpy and entropy data set, which described the characteristics of energetic materials and, are highly reliant on the temperature conditions.Thus, a very similar behaviour in variation with respect to temperature has been observed for each thermodynamic property. Therefore, these small increments values influence the overall computational thermochemical data distribution for the chemical substance of interest. The box plots are made up of two parts; 1) a box which depicts the data spread at 25 to 75 %, and 2) whiskers which depict the data spread from underneath the 25 % to the minimum point and over 75 % to the maximum point.[12] While examining the graph (as shown in Fig. 2), it has been observed that, the lower quartile (lower end of the box, Q_1) for the propanol and butanol isomers spanned from –195.11 to –35.31 J/mol and –248.77 to –45.50 J/mol, respectively.

Furthermore, the highest quartile (upper end of the box, Q_3) that constitute almost 25 % of the data set, have values in the range from 153.37 to 425.51 and 175.49 to 490.36 J/mol for the propanol and butanol isomers, respectively. Surprisingly, all of the molecules' interquartile ranges or mid-spreads are quite comparable, suggesting that our theoretical approaches computed the thermochemical molecular properties in accordance with each other and with the reference data. Using the B3LYP DFT approach, a box-plot demonstrated a small interquartile range for propanol isomers and a higher interquartile range for butanol isomers. Moreover, the propanol and butanol isomers' median value (Q_2) appears to be pushed more towards the lower quartile. Additionally, the thermochemical data was spread more than 30 % towards the upper quartile, demonstrating that the range between (Q_1–Q_2) < (Q_2–Q_3) in the computed data seems to have a positive symmetry or positive skew.

The individual thermodynamic property (*i.e.*, specific isobaric heat capacity (Cp), enthalpy change of formation ($\Delta_f H°$), entropy change of formation ($\Delta_f S^0$) and Gibbs energy change of formation ($\Delta_f G^0$)) was also analysed by the Box–Whisker plot, in order to determine the discrepancies and departures in the computed data-set, respectively. The Box-plot statistical data distribution analysis is depicted in Fig. 3. After the statistical analysis of computed enthalpy molecular property of the aforementioned isomers of propanol and butanol molecules, it was observed that, Q_1 varied from –126.44 to –62.83 J/mol and Q_3 varied from 55.34 to 179.07 J/mol, respectively. Moreover, the Fig 3b showed that the molecular property data was spread more than 59 % towards the upper quartile, which means that there exist a positive skew in the obtained computed characteristic.

Analysing the data-set for the specific heat, entropy and Gibbs energy (as depicted in Fig. 3b–d), a variation has been observed from –248.77 to –71.72 J/mol, –156.60 to –34.64 J/mol, 174.89 to 324.68 J/mol for the lower quartile, and 80.64 to 189.02 J/mol, 44.59 to 92.73 J/mol and 357.90 to 490.36 J/mol for the upper quartile, respectively. Unexpectedly, the data set for all three parameters were spread at least 43.04 % towards the lower quartile, suggesting that the thermochemical data set for each property exhibit some negative symmetry.

Principal component analysis

The multivariate PCA was performed to reduce the dimensionality problem and possible exploratory analysis of the proposed computed thermodynamic physicochemical properties of the molecules. The summarized eigenvalues and the variance variabilities on each principal component for each data set constituting of 15 cases (*i.e.*, temperature variations) for each thermodynamic property using the 22 variables (*i.e.*, molecular structures at the given model) are being summarized and illustrated in Table III.

Normally, PCA data processing comprises of two distinct steps,[29] the first step is the PCA "data compression" step, which removes the redundancy in the original data, where the first few main components (PC) generally describe most of the information contained in the original data. The second stage is the "cluster visualization” formed in the datasets, making a two-dimensional graph of the principal components. Information regarding data analysis is being obtained by interpreting PCA parameters such as loadings, scores, eigenvectors, eigenvalues, etc., of the principal components (PCs).

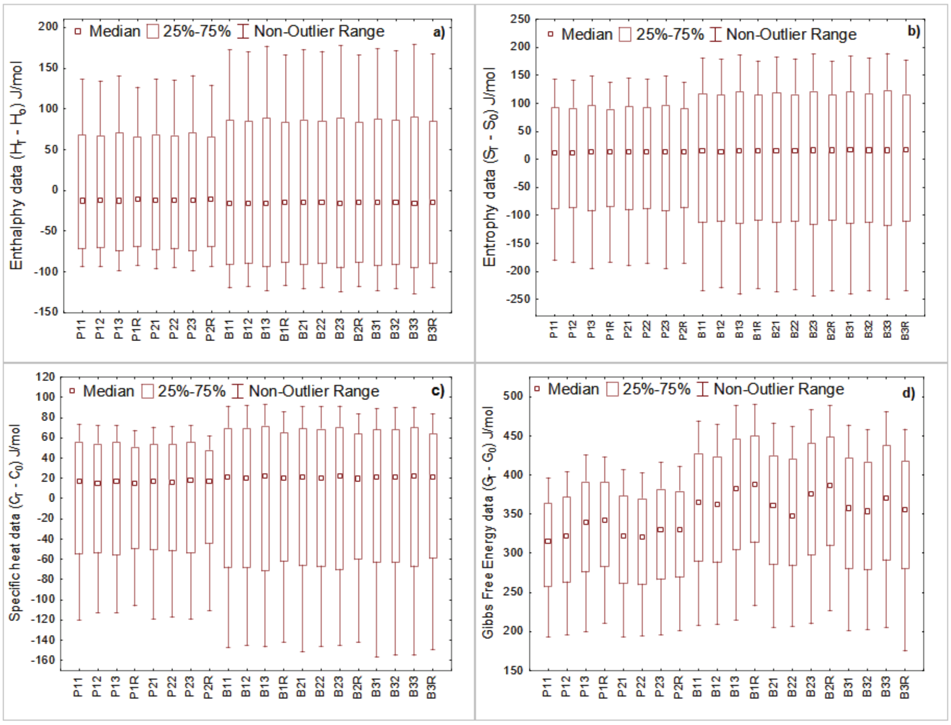

Fig. 3. Box-plot for the different molecular structures used for the computation of thermodynamic properties using DFT and semi-empirical methods, where, the individual data is depicted in sub-figure; a) enthalpy data, b) entropy data, c) epecific isobaric heat coefficient data and d) Gibbs energy data. P*xy* and B*xy* stands for P = propanol, B = butanol; *x* = 1 or 2 propanol and 1, 2 or *tert*-butanol; *y* = 1, 2, 3 for PBE0, B3LYP and PM6 DFT methods, respectively, while P2R and B3R means literature values[21-24] for 2-propanol and *tert*-butanol.

The molecules of interests were similar in chemical nature and almost of the same size having same number of basis functions, which make a difficulty in the understanding the small variance in the molecular properties. Thus, the principal components (PCs) and their eigenvalues is used to identify the most important minuscular changes and inclinations in the data sets.[30] PCA for the molecules of interest was performed in two steps: for the entire data set, including the specific isobaric heat capacity, Gibbs energy, enthalpy and entropy data set, and individually for each thermochemical property, taking into account computational methodologies, temperature variation and molecular structure effect.

TABLE III. Eigenvalues and proportion of total variability of thermodynamic properties computed at DFT/cc-pVTZ methodologies and PM6 semi-empirical method for the molecules of interests

Property	PCA parameter	PC1	PC 2	PC3
Complete data	Eigenvalues	21.9498	0.0473	0.0013
	% variability	99.7718	0.2151	0.0058
Specific heat	Eigenvalues	21.9989	0.0009	0.0002
	% variability	99.9950	0.0041	0.0008
Enthalpy	Eigenvalues	21.9993	0.0006	0.0001
	% variability	99.9969	0.0026	0.0005
Gibbs energy	Eigenvalues	21.9853	0.0104	0.0034
	% variability	99.9334	0.0474	0.0155
Entropy	Eigenvalues	21.9989	0.0009	0.0002
	% variability	99.9950	0.0041	0.0008

The first three principal components (PCs) were the most important, accounting for 99.992 % of the total variance in molecular properties. PCs are fitted to a data set in PCA so that the first PC may explain as much of the original variation between the cases as possible. The second PC, on the other hand, is adjusted orthogonally to the first PC and is intended to characterize the majority of the remaining variances and so on.[29] According to the Table III, the 1st and 2nd principal components in this analysis demonstrate an overall variance of at least 99.980 % for characteristics of interest**.**

Temperature dependence on the thermodynamic quantities (Score-plots)

The scores are the projections of molecular property at each temperature using the computational methods along the principal component line, and plotting them in the form of a bi-plot can explore the similarities and differences in between them, resulting in the development of groups.[30] The bi-plots in between the *T*-distributions for the complete thermodynamic data set is depicted in Fig. 4.

It can be observed that, the dataset's score-plot PC1–PC2 (as can be seen Fig. 4b) efficiently showed the distribution and categorization of thermochemical property variation with temperature. Based on the orientation of PC2 over PC1, it was possible to detect temperature variation in two agglomerating groups, indicating a variance of 99.998 % for the entire thermochemical properties data set. Aside from grouping the distinct factors, the bi-plot of scores can also disclose and identify data outliers in the data. While, on PC1 and PC2, for the molecule of interest, we discovered an interesting ascending and descending order in each molecular feature with an increase in temperature, correspondingly. The Gibbs energy and entropy molecular properties were found to follow a similar pattern as temperature rises, with the only difference being such a rapid downward order in the variation of the Gibbs energy as compared to the entropy variation, due to the small intermission of increment of molecular property.

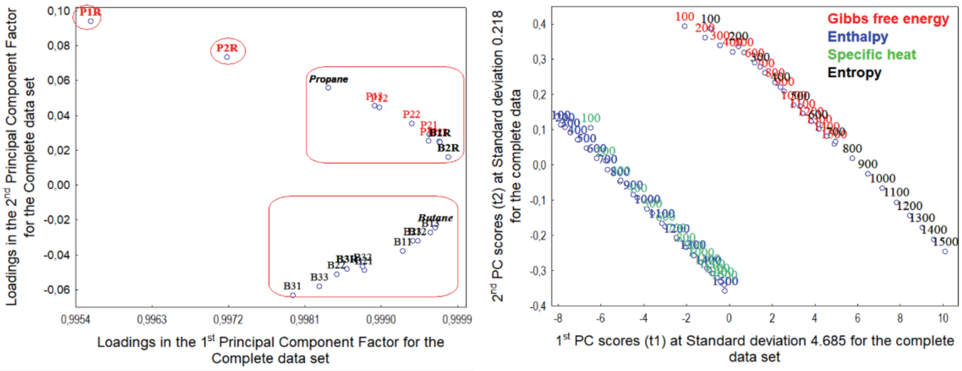

Fig. 4. Plots graphs representing the combination of loadings coefficient (a) and scores (b) for the 1st and 2nd principal components obtained from the thermodynamic quantities data analysis of interest at all the applied computational methods. P*xy* and B*xy* stands for P = propanol, B = butanol; *x* = 1 or 2 propanol and 1, 2 or *tert*-butanol; *y* = 1, 2, 3 for PBE0, B3LYP and PM6 DFT methods, respectively, while P2R and B3R means literature values[21-24] for 2-propanol and *tert*-butanol.

Since, the entropy and Gibbs energy of a chemical structure is thought to be strongly affected by the physical state of the molecules, particularly intermolecular forces affecting the entropy of the substances and the dependence of the Gibbs energy on the entropy value. Which gives the similar results in the computation process for the present studies gas-phase geometries of interest. The response of the specific isobaric heat capacity and enthalpy variation with temperature, on the other hand was found to be equivalent, where both thermodynamic variables represent the heat content of a chemical substance. To understand in detail the behaviour of each molecular property with respect to the temperature variation, the thermochemical properties were studied separately.

The molecular structures showed an upward and subsequently a declining order on the 2nd PC for the specific isobaric heat capacity values and entropy molecular properties, according to the score-plot (explored in Fig. 5). It's possible that the declining order after 700 K is due to a modest increase in molecular characteristics as temperature rises. Furthermore, the enthalpy of molecular structures first display an ascending order on PC2 from negative to positive scores values, culminating at 700 K, followed by a decline in molecular characteristics.

This suggests that the temperature of 700 K is assumed to be a critical temperature for the specific heat, enthalpy, and entropy of the molecular structures of interest, with a consequent decrease in computed values. In contrast, after examining Fig. 5b, it was observed that the Gibbs energy property appears different from the other three molecular thermochemical properties. In comparison to the other computed values, PCA separated the molecular feature for all the molecules at 100 K. In addition, the score-plots separated the data into three groups, each varying from 200 to 400 K, 500 to 900 K and 1000 to 1500 K. This unexpectedly fluctuating molecular property behaviour may be due to the propanol isomers, that have a smaller number of carbon and hydrogen atoms than the butanol isomers, leading to a lower molecular property increment. Thus, resulted in the segregation of the Gibbs energy in to three different increment order with the variation of temperature.

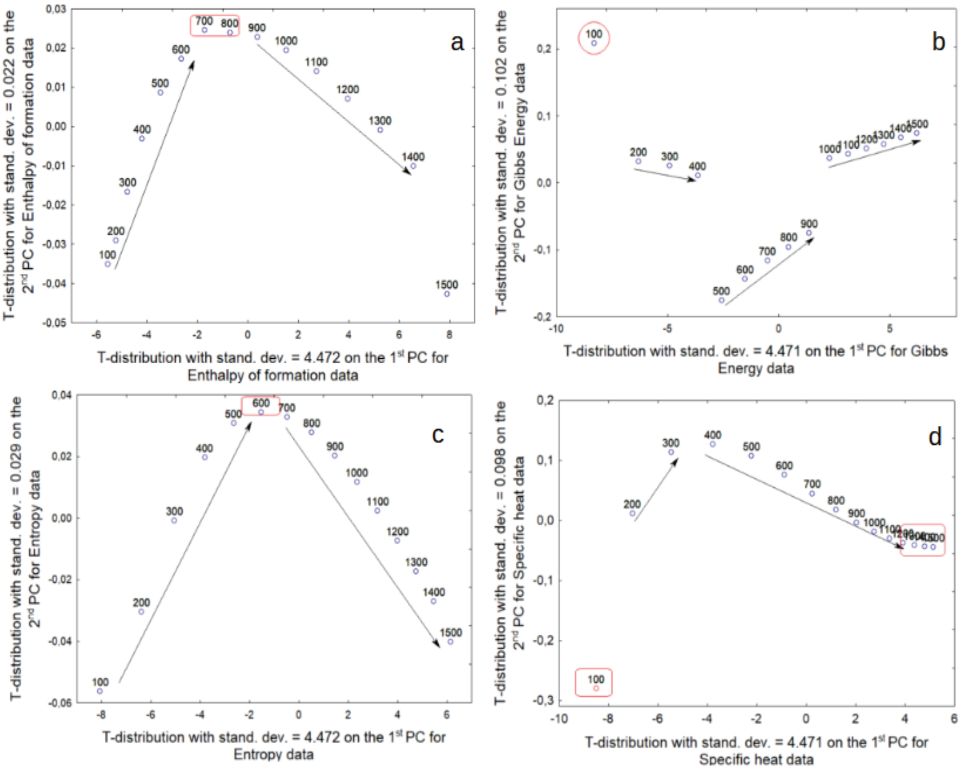

Fig. 5. Score (*t*) plots in between the 1st and 2nd principal components obtained for the thermodynamic characteristics data analysis of interest for all the applied computational methods; a) enthalpy, b) Gibbs energy, c) entropy, d) specific heat at constant pressure data.

Exploration of thermochemical properties (loading-coefficients plots)

The analysis of loading factors is suggested as a very useful tool to explore the importance of the variables based on the correlations in between them,[31] which in the present case is applied to the thermochemical properties of the molecular structures using the theoretical methods. In the current research, the association between thermodynamic quantities for each molecular structure at the given theory level was evaluated for the entire data set, taking into consideration commonalities and contrasts with reference data, and subsequently, the individual molecular property analysis were also performed for each computed thermochemical characteristic. Since the thermochemical property represents the energy content of a chemical substance, it varies and depends on the number and nature of chemical bonds. In this situation, we were studying the propanol and butanol isomers, which have almost the same amount of chemical linkages and are very much similar in nature. As a result, the variance in thermochemical characteristics with temperature is remarkably similar.

During the principal component analysis of the data set, a strong correlation on the first principal component resulted in the aggregation of all the loading factors values on it. Such agglomeration problems can be solved, while examining the latent minuscule information caused by the high associations between the strong correlated variables, by making a possible projection bi-plot in between the loading-coefficients of two principal components.[14] For each variable, which is the molecular property computed at the computational method for the molecular structures, the loading factor values on each component represent how much of the characteristic is common in between the molecular structure on that PC. Khan *et al.*[11] had used arrows projected strategy to assess the vanishingly small differences in between the theoretical methodologies used to compute Raman intensities, where a smaller angle than 45° was considered positive and stronger correlation, 90° was considered insignificant correlation, and above 135° up to 180° was considered negative strong correlation in between the variables of interest. The authors of the present study utilized a bi-plot between the PC2 and PC3 where the ascending shift on 3rd PC from –0.01354 to 0.01808 values to classify and grouping in between the variables of interest. Upon analysing, it was observed that the 1st PC×2nd PC factor loadings bi-plots values successfully separated the structural isomers of propanol and butanol. It means that, the smaller variance on 1st by 2nd component can explore the hidden minuscule variation in between the chemical compounds of similar nature containing almost similar number of atoms. On the other hand, while studying the loadings on the principal components, it also helped us to diagnose the strong similarities in the results in between the utilized different computational methodologies.

Where, we have observed that the semi-empirical methodology successfully computed the thermodynamic quantities of the benign biofuel as precisely to that of DFT methodologies. Furthermore, the loading-coefficients for the specific isobaric heat capacity, enthalpy, entropy and Gibbs energy for the isomeric structures of the propanol and butanol molecules were analysed in detail for the quanti–quali similarities and differences in between molecular properties for the structural isomers to the reference data, as illustrated in the Fig. 5. PCA computed the loading values considerably differently since each molecular thermochemical property of interest varies with temperature differently.

This allows us to investigate the materials' molecular energetic characteristics. In the analyses of the enthalpy data, the loading coefficient bi-plots (as seen in the Fig. 6a) classified the molecular property computed for each molecule of propanol and butanol isomers at the theoretical level in well-distributed correlated way. Where, plotting the loadings coefficients for the 2nd PC upon 1st PC segregated the correlated molecular property for the isomers of propanol and butanol molecular properties much closer to the traditional fuels, *i.e.*, propane and butane molecular property along with their reference data. It is suggested that variation of enthalpy of formation for the molecules varied with same fraction. Interestingly, the *tert*-butanol molecule represents a small dissimilarities in the molecular property compared to the other molecules, which can be visualized in the score-plot analysis. Furthermore, the loading-coefficients values on the 3rd component spread the molecular property against 2nd PC, while examining the PCA loading-coefficient for the specific isobaric heat capacity data set (as shows in the Fig. 6d). The bi-plot distributed the molecular properties equally in all sides, where the *tert*-butanol was being segregated from the rest of the molecules

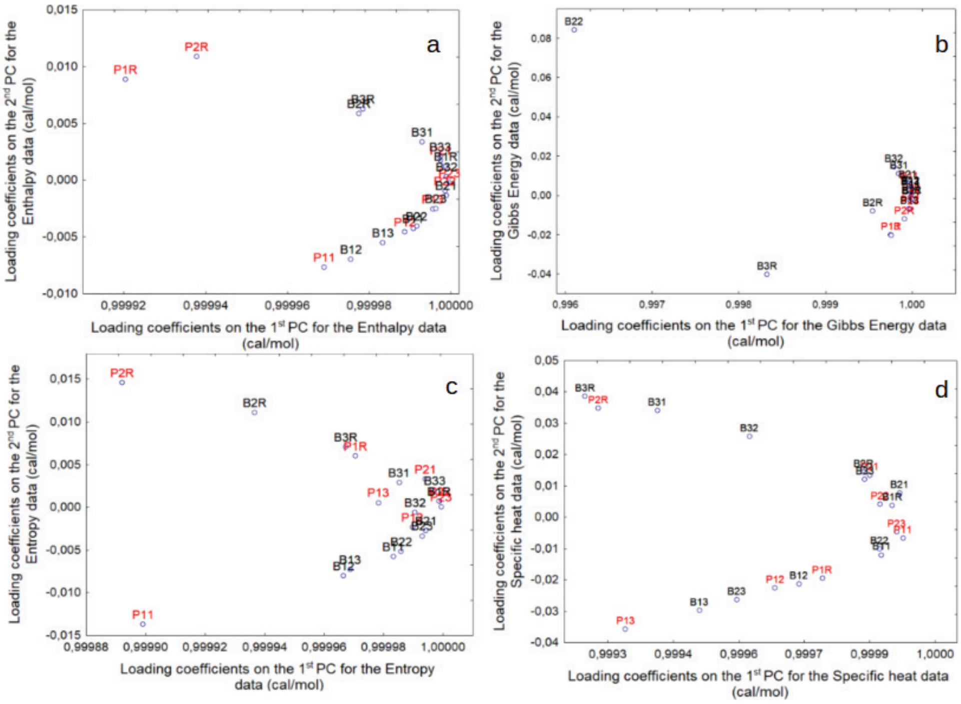

Fig. 6. Plots graphs representing the combination of loadings for the 1st and 2nd principal components obtained of the thermodynamic quantities data analysis of interest at all the applied computational methods: a) enthalpy, b) Gibbs energy, c) entropy and d) specific heat at constant pressure data.

due to high specific heat content by these structures. PCA interestingly separated the molecular property of butane and propane from the rest of structures due to absence of the "O–H" group in them. On investigating the loading bi-plots for the entropy property (illustrated in the Fig. 6c), the loading points for data-set was all agglomerated in between the –0.0199 to 0.0081 on the PC2. A strong correlation has been observed for all the other structures. In addition, the molecular property for the 1-propanol using the DFT methods were found separated from the rest of the molecules.

While analysing the values of the loading coefficients of the Gibbs energies, we discovered that the molecular property behaved very similarly to the entropy of a system, except for the case of 2-butanol using DFT method, which shows a dissimilarity in molecular property compared to the other structures. Since, Gibbs energy change is a derived quantity that blends the enthalpy and entropy change in the computational equations for a chemical substance.[32] Thus, it resulted in very competitive similarities to the entropy behaviour.

CONCLUSIONS

The present investigation concludes that high-level electronic structure calculations combined with multivariate statistical analysis was very much successful investigating the different thermodynamic properties of the various chemical structures of interests. The *RMS* deviations of the molecular properties for the given structures represent a lot of similarities in the computed values and the literature values. The box-and-whisker plot statistical technique demonstrated that all the thermochemical quantities are precise and symmetrically distributed along the median values according to the rise in temperature. The principal component analysis explored the minuscule changes in the thermochemical properties for the molecular structures of interests containing the same number of bonds and types of atoms rapidly. Furthermore, the score-plots diagnosed the molecular properties variations after a certain peak of temperature with descendant variation in the statistical parameters. In addition, the loading-coefficient parameters explored and categorized the propanol from the butanol molecular properties, where a modest difference in the molecular properties is being found.

Finally, we suggest that PCA is a robust statistical method which can be applied not only to evaluate the molecular thermodynamic quantities, but in future it can also be applied to the other computational molecular properties data set.

Acknowledgment. The authors thank the *Instituto de Química* (IQM/UNICAMP) for the use of the availability of Computational facilities in Campinas. Mr. Serra thanks the UEMA postgraduate program for the monthly stipend and the use of universities resources.

ИЗВОД

ПРИСТУП МУЛТИВАРИЈАНТНОМ СТАТИСТИЧКОМ АНАЛИЗОМ ПРИ ИСТРАЖИВАЊУ ТЕРМОДИНАМИЧКИХ ВРЕДНОСТИ БЕЗОПАСНОГ АЛТЕРНАТИВНОГ ГОРИВА

KASSIO FELIPE DA COSTA SERRA,[1] ALAMZEB KHAN,[2] RAQUEL MARIA TRINDADE FERNANDES,[3] PEDRO ANTONIO MUNIZ VAZQUEZ[4] и ALAMGIR KHAN[3]

[1]Programa de Pós-Graduação Engenharia Aeroespacial, Universidade Estadual do Maranhão – UEMA, Cidade Universitária Paulo VI, Campus São Luís/MA, CEP 65000-00, Brasil, [2]Laboratório de Ressonância Paramagnética Eletrônica (LARPE), Departamento de Física, Universidade Estadual de Londrina (UEL), Londrina, PR, Brasil, [3]Departamento de Química, Centro de Educação, Ciências Exatas e Naturais – CECEN, Universidade Estadual do Maranhão – UEMA, Cidade Universitária Paulo VI, Campus São Luís/MA, CEP 65000-00, Brasil, [4]Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas – UNICAMP, Caixa Postal 6154, Campinas, SP, 13083-970, Brasil

За добијање смисленог тумачења, из мноштва података и обезбеђивања њихове ваљаности у подручју примене, анализа хемијских података је постала озбиљан изазов у развоју и примени нових протокола, техника и методологија за заједнице математичког моделовања и друга друштва за науку о подацима. Зато се у овом раду предлажу брзе и робустне технике „box-and-whisker plot" и статистичке технике мултиваријантне анализе главних компоненти (PCA) за оцену података термодинамичких молекулских особина структура молекула безопасних горива. Опазили смо да "box-and whisker plot" техника успешно истражује све термохемијске молекулске особине прецизно, и описује симетричну расподелу података дуж медианских вредности у односу на пораст температуре. Поред тога, применом PCA технике, графици оцена (score-plots) главних компоненти су дијагностиковали необичне варијације у молекулским особинама након одређеног врхунца температуре са опадајућом варијацијом статистичких параметара. Даље, PCA параметри не само да раздвајају термодинамичке особине пропанола и бутанола, већ такође, њихове варијације са температуром. Тако, закључујемо да су „Box-whisker" и PCA статистичке технике робустан и брз метод за процењивање и вредновање мноштва података о молекулским термодинамичким величинама.

(Примљено 30. маја, ревидирано 29. августа, прихваћено 20. новембра 2023)

REFERENCES

1. J. G. Yu, Q. J. Xiang, M. H. Zhou, *Appl. Catal., B* **90** (2009) 595 (https://doi.org/10.1016/j.apcatb.2009.04.021)
2. A. A. Khan, M. Tahir, *J. CO2 Util.* **29** (2019) 205 (https://doi.org/10.1016/j.jcou.2018.12.008)
3. O. Doğan, *Fuel* **90** (2011) 2467 (https://doi.org/10.1016/j.fuel.2011.02.033)
4. H. F. Mustafa, S. Abdullah, M.Z. Abdullah, K. Sopian, A. K. Ismail, *Renew. Energy* **74** (2015) 505 (https://doi.org/10.1016/j.renene.2014.08.061)
5. N. Yilmaz, A. Atmanli, *Energy* **140** (2017) 1378 (https://doi.org/10.1016/j.energy.2017.07.077)
6. P. Durre, *Curr. Opin. Biotechnol.* **22** (2011) 331 (https://doi.org/10.1016/j.copbio.2011.04.010)
7. Y. Dahman, C. Dignan, A. Fiayaz, A. Chaudhry, in *Biomass, Biopolymer-Based Materials, and Bioenergy*, D. Verma, E. Forunati, S. Jain, X. Zhang, Eds., Woodhead

Publishing Series in Composites Science and Engineering, Woodhead Publishing, Elsevier, New York, 2019, p. 241 (https://doi.org/10.1016/B978-0-08-102426-3.00013-8)

8. A. T. Balaban, *HYLE* **19** (2013) 107 (https://www.hyle.org/journal/issues/19-1/balaban.htm)
9. X. Wu, F. Kang, W. Duan, J. Li, *Prog. Nat. Sci.: Mater. Int.* **29** (2019) 247 (https://doi.org/10.1016/j.pnsc.2019.04.003)
10. K. E. Gutowski, R. D. Rogers, D. A. Dixon, *J. Phys. Chem., B* **111** (2007) 4788 (https://doi.org/10.1021/jp066420d)
11. A. Khan, P. A. M. Vazquez, R. M. T. Fernandes, *Spectrochim. Acta, A* **245** (2021) 118891 (https://doi.org/10.1016/j.saa.2020.118891)
12. E. Szymanska, *Anal. Chim. Acta* **1028** (2018) 1 (https://doi.org/10.1016/j.aca.2018.05.038)
13. A. M. Souza, R. Poppi, *J. Quim. Nova* **35** (2012) 223 (http://dx.doi.org/10.21577/0100-4042.20170480)
14. A. D. Becke, *J. Chem. Phys*. **98** (1993) 5648 (https://doi.org/10.1063/1.464913)
15. C. Adamo, B. V. Toward, *J. Chem. Phys.* **110** (1999) 6158 (https://doi.org/10.1063/1.478522)
16. T. H. Dunning, K. A. Peterson, D. W. Woon, P. V. R. Schleyer, "*Encyclopedia of Computational Chemistry*", Wiley, New York. 1998, pp. 88–115
17. *Gaussian 09, Revision B.01*, Gaussian, Inc., Wallingford CT, 2009
18. StatSoft Inc. (2004). Statistica (data analysis software system), version 7. Available from www.statsoft.com
19. W. J. Dixon, F. J. Massey, Jr., *Introduction to statistical analysis*, McGraw-Hill, New York, 1951
20. R. McGill, J. W. Tukey, W. A. Larsen, *Am. Statistician* **32** (1978) 12 (https://doi.org/10.2307/2683468)
21. E. Stromsoe, H. G. Ronne, A. L. Lydersen, *J. Chem. Eng. Data* **15** (1970) 286 (https://doi.org/10.1021/je60045a040)
22. J. Chao, K. R. Hall, K. N. Marsh, R. C. Wilhoit *J. Phys. Chem. Ref. Data* **15** (1986) 1369 (https://doi.org/10.1063/1.555769)
23. J. Chao, F. D. Rossini, *J. Chem. Eng. Data* **10** (1965) 374 (https://doi.org/10.1021/je60027a022)
24. Daniel Siderius, NIST Standard Reference Simulation Website - SRD 173 (2017) National Institute of Standards and Technology, (https://doi.org/10.18434/mds2-232) (Accessed 2023-10-04)
25. T. Chai1, R. R. Draxler, *Geosci. Model Dev. Discuss*. **7** (2014) 1525 (http://doi.org/10.5194/gmdd-7-1525-2014)
26. E. Nikolic-Doric, K. Cobanovic, Z. Lozanov-Crvenkovic, in *Proceedings of International Conference on Teaching Statistics* (2006) Slavador, Brasil, ICOT 7 Published By IASE, Belgium, 2006, C137 (ISBN: 978-90-73592-24-7)
27. A. R. Henderson, *Clin. Chim. Acta* **366** (2006) 112 (https://doi.org/10.1016/j.cca.2005.11.007)
28. C. E. Brown, in *Applied Multivariate Statistics in Geohydrology and Related Sciences*, C. E. Brown, Ed., Springer, Berlin, 1998, p. 155 (http://doi.org/10.1007/978-3-642-80328-4_13)

29. T. Jolliffe, in *Principal Component Analysis,* T. Jolliffe, Ed., Springer Series in Statistics, Springer, New York, 1986, p. 115 (https://doi.org/10.1007/978-1-4757-1904-8_7)

30. F. Mabood, G. Abbas, F. Jabeen, Z. Naureen, A. Al-Harrasi, A. M. Hamaed, J. Hussain, M. Al-Nabhani, M. S. Al-Shukaili, A. Khan, S. Manzoor, *Food Addit. Contam., A* **35** (2018) 404 (https://doi.org/10.1080/19440049.2017.1418090)

31. F. Mabood, S. A. Gilani, M. Albroumi, S. Alameri, M. M. O. Al-Nabhani, F. Jabeen, A. Alharrasi, R. Boqué, S. Farooq, A. Hamaed, A. Naureen, A. Khan, Z. Hussain, *Fuel* **197** (2017) 388 (https://doi.org/10.1016/j.fuel.2017.02.041)

32. A. D. McQuarrie, J. D. Simon, *Physical Chemistry: A Molecular Approach,* University Science Books, Melville, NY, 1997, p. 1396 (ISBN 978-0935702996).