

Predicting gas phase entropy of select hydrocarbon classes through specific information-theoretical molecular descriptors

C. Raychaudhury, I.H. Rizvi & D. Pal

To cite this article: C. Raychaudhury, I.H. Rizvi & D. Pal (2019): Predicting gas phase entropy of select hydrocarbon classes through specific information-theoretical molecular descriptors, SAR and QSAR in Environmental Research, DOI: [10.1080/1062936X.2019.1624613](https://doi.org/10.1080/1062936X.2019.1624613)

To link to this article: <https://doi.org/10.1080/1062936X.2019.1624613>



Published online: 20 Jun 2019.




Submit your article to this journal [↗](#)



View Crossmark data [↗](#)



Predicting gas phase entropy of select hydrocarbon classes through specific information-theoretical molecular descriptors

C. Raychaudhury, I.H. Rizvi and D. Pal 

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

ABSTRACT

The usefulness of five specific information-theoretical molecular descriptors was investigated for predicting the gas phase entropy of selected classes of acyclic and cyclic compounds. Among them, total information on atomic number (TI^Z), graph vertex complexity (H^V) and total information on bonds (TIB^{AT}), considered together showed the best correlation along with a low standard deviation ($r^2 = 0.97$, $s = 21.14$) with gas phase entropy values of 130 compounds. The multiple regression equation treating these three indices as independent variables was statistically highly significant which was evident from the F -statistics. In particular, very small difference between r^2 and r^2 -pred values indicates that the regression model is not overfitted and is, therefore, suitable for prediction purposes. When truly used as a training set to predict (from regression equation) 40 additional compounds we get a very high correlation ($r^2 = 0.975$), which remains almost identical ($r^2 = 0.97$) for the combined data set of 170 compounds. The three indices appear to be useful descriptors producing correlation that remains stable with the change in the size of the data set. Also, the information-theoretical measures appear to capture an additive-cum-constitutive nature of gas phase entropy yielding an acceptable statistical fit.

ARTICLE HISTORY

Received 15 March 2019

Accepted 24 May 2019

KEYWORDS

Gas phase entropy; hydrocarbons; information theoretical molecular descriptor; regression model; QSPR

Introduction

It has been a continuous endeavour by the researchers to develop methods [1–6] for the prediction of thermodynamic entropy of chemical compounds since it has various important practical applications such as entropy of boiling in chemical engineering and environmental sciences [6]. The idea is to quantify different properties of molecular structures which can be used to develop theoretical methods of predicting thermodynamic entropy. One of the ways of doing that is to use quantitative molecular descriptors to predict entropy from statistically significant predictive regression models [3,4]. In this regard, information-theoretical molecular descriptors, developed using Shannon's measure of information [7], have found significant applications for different series of organic compounds [4,5,8]. For example, an information-theoretical index reflecting molecular size has been found to be useful in explaining gas phase thermal entropy

of a series of diverse chemical compounds in one of our studies [4]. Also, information-theoretical measures reflecting molecular symmetry have been found to correlate with entropy values of a series of compounds [8]. Therefore, it is apparent that looking into some fundamental characteristics of gas phase entropy and identification/use of relevant molecular descriptors may be helpful in developing methods for predicting this thermodynamic property. One of such fundamental aspects is that gas phase entropy may not be a purely additive property as mentioned by Domalski et al. [2] and may depend on different structural characteristics. Statistically significant correlations between information-theoretical molecular descriptors which are computed from the partition of the elements of a molecular graph into disjoint classes by defining an equivalence relation on molecular structural properties seem to indicate that such descriptors have the capability to take care of those aspects of gas phase entropy which make it different from just an additive thermodynamic property. It, therefore, seems reasonable to carry out further studies with those information-theoretical indices that reflect relevant molecular structural characteristics essential for developing statistically significant predictive regression models for predicting gas phase entropy of chemical compounds.

In order to carry out studies in this direction, we have collected from the literature [2] a series of 170 compounds composed of various classes of cyclic and acyclic hydrocarbons along with their experimentally determined gas phase entropy (S°) values. We have then considered five specific information-theoretical molecular descriptors for the present study. These five descriptors translate various structural aspects which are believed to be relevant for explaining gas phase entropy of chemical compounds. For example, the descriptor total information on atomic number (TI^Z) reflecting molecular size has been considered since this descriptor has been found to contribute highly in producing statistically significant results in our previous studies [4] for predicting gas phase thermal entropy of a series of chemical compounds. However, we have also considered the descriptor information on atomic number (I^Z) to investigate its usefulness for the present study.

It also seems reasonable to consider some information-theoretical descriptor that takes care of the topological characteristics of chemical structure effectively. In this regard, an information-theoretical molecular descriptor defined on graph-theoretical distances between pairs of vertices in a molecular graph viz., graph vertex complexity (H^V) [9] has been considered in the present study. The index H^V is obtained as an average of the information content measures (local indices) for individual vertices in a molecular graph [9] and is therefore believed to take care of molecular topological architecture of chemical compounds in greater detail. Furthermore, the other important aspect of considering this index is that since H^V is computed from the information content measures of individual vertices, it may also be regarded as obtained from the contribution of individual groups in a compound. Since group contribution has been considered in various ways [2,6] in developing methods for predicting entropy of chemical compounds, consideration of H^V for the present purpose seems to be meaningful.

In addition to that, it is also quite apparent that information-theoretical indices giving measures for the diverse nature of covalent bonds connecting two atoms in chemical compounds may also be taken into consideration since such indices take care of a different and a very important aspect of chemical structure. It is entirely desirable

that the scope of the predictive power of molecular indices be expanded to cater to more diversified groups of hydrocarbons; therefore, although few indices for the bonds of chemical compounds are available in the literature [8], we have proposed here two new information-theoretical indices on the edges of molecular graph viz., information on bonds from atom type (IB^{AT}) and total information on bonds from atom type (TIB^{AT}) which are based on different types of covalent bonds (single, double, triple) connecting atoms of different chemical nature such as the bonds like C-N, C-O, C = O, C-H, etc. The present investigation suggests that the information-theoretical indices used in this study can produce statistically significant correlation in explaining gas phase entropy of the hydrocarbons under consideration.

Materials and methods

Gas phase entropy (S°) values of a series of 170 hydrocarbons (17 alkanes, 20 alkenes, 14 alkynes, 20 substituted tertiary alkanes, 13 substituted quaternary alkanes, 29 substituted and unsubstituted cycloalkanes, 33 aromatic hydrocarbons as substituted benzenes and 24 aromatic hydrocarbons as substituted naphthalenes) were gathered from the literature [2]. We have also considered five specific information-theoretical indices for explaining gas phase entropy of these compounds viz., information on atomic number (I^Z), total information on atomic number (TI^Z) [4], graph vertex complexity (H^V) [9], and the two newly proposed indices on chemical bonds – information on bonds from atom type (IB^{AT}) and total information on bonds from atom type (TIB^{AT}). The methods of computing the information-theoretical indices, considered here, are based on Shannon's measure of information [7] applied to the elements of suitably weighted connected graphs or, vertex labelled graphs or, simple graph/multigraph models [10] of molecular structure. We, therefore, describe here the methods of computing the indices I^Z and TI^Z [4], H^V [9] and the two newly defined indices IB^{AT} and TIB^{AT} which have contributed in obtaining statistically significant correlations in multiple regression analysis with the gas phase entropy values of the hydrocarbons considered for the present study. The values of these information-theoretical indices have been computed by considering hydrogen-filled molecular graphs of the corresponding compounds using a computer program developed in our laboratory.

Now, since the information-theoretical indices are defined based on Shannon's measure of information [7], we first describe here the method of computing Shannon's measure of information for a system S to be followed by the methods of computing the indices mentioned above.

Let there are n elements in a system S . If these n elements are partitioned into k disjoint classes as $(n_1, n_2, n_3, \dots, n_k)$ on the basis of an equivalence relation applied onto the elements of S , then Shannon's information content (I_S) of S is given by:

$$I_S = - \sum_{i=1}^k p_i \log_2(p_i) \quad (1)$$

In Equation (1), $p_i = (n_i/n)$, $p_i \geq 0$, n_i being the cardinality of i^{th} partitioned class. It may be noted that by considering the logarithm base 2, I_S is expressed in bits. However, we will skip using this unit against the computed information content values for the sake of

simplicity. Now, by substituting p_i in terms of n and n_i in equation – 1, I_S may be computed from:

$$I_S = \sum_{i=1}^k \left(\frac{n_i}{n}\right) \log_2 \left(\frac{n}{n_i}\right) \quad (2)$$

Clearly, Equation (2) can be used more conveniently to compute information content values once the total number of elements and those in partitioned classes are known. Furthermore, a measure of total information content [11] of S , TI_S may be obtained as:

$$TI_S = n \times I_S \quad (3)$$

We now describe below the methods of computing three different types of information-theoretical indices – (A) I^Z and TI^Z ; (B) H^V ; and (C) IB^{AT} and TIB^{AT} (newly defined):

Information (I^Z) and total information (TI^Z) on atomic number (Z)

Let there are N atoms in a chemical compound C and $Z_j, j = 1, 2, \dots, N$ be the atomic numbers of those N atoms of C . Also, let Z be the sum of all the atomic numbers Z_j given by (4):

$$Z(C) = \sum_{j=1}^N Z_j \quad (4)$$

Now, considering Z_j as a partition of $Z(Z_1, Z_2, \dots, Z_N)$, one can have a measure of information on atomic number I^Z for compound C following Equation (2) as:

$$I^Z(C) = \sum_{j=1}^N \left(\frac{Z_j}{Z}\right) \log_2 \left(\frac{Z}{Z_j}\right) \quad (5)$$

Subsequently, one can also have a measure of total information on atomic number (TI^Z) for compound C following Equation (3) as:

$$TI^Z(C) = Z \times I^Z(C) \quad (6)$$

It may be noted that the indices I^Z and TI^Z may also be obtained by considering vertex weighted connected molecular graph of a compound where atomic numbers can be used as the weights assigned to the vertices for the corresponding atoms in a molecule [4].

Computation of I^Z and TI^Z indices

In order to illustrate the computational procedure for these two information indices, we have taken vertex weighted molecular graph, weighted by atomic numbers (Z) of the corresponding atoms of n -butane (Figure 1).

In n -butane, there are four carbon atoms having atomic number 6 for each of them and there are 10 hydrogen atoms having atomic number 1 for each of them. Therefore, the sum of all the atomic numbers of carbon and hydrogen atoms in n -butane is:

$$Z(n\text{-butane}) = (4 \times 6) + (10 \times 1) = 34$$

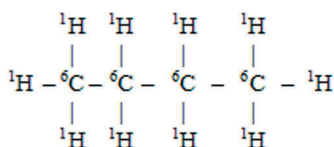


Figure 1. Vertex weighted connected molecular graph of *n*-butane. The numbers given as superscripts before the atomic symbols C and H are the corresponding atomic numbers.

Now, considering the atomic numbers of the atoms as a partition of Z (*n*-butane), one can compute I^Z and TI^Z for *n*-butane using Equations (5) and (6) respectively.

$$I^Z(n\text{-butane}) = \left[4 \times \left(\frac{6}{34} \right) \log_2 \left(\frac{34}{6} \right) \right] + \left[10 \times \left(\frac{1}{34} \right) \log_2 \left(\frac{34}{1} \right) \right] = 3.2628$$

$$TI^Z(n\text{-Butane}) = 34 \times 3.2628 = 110.9352$$

Graph vertex complexity (H^V)

This (H^V) information-theoretical topological index (ITTI) has been developed in connection with studies for discrimination of isomeric structures and H^V has been found to be quite discriminating and is believed to take care of the topological properties of molecular structure in greater details [9]. The index H^V is computed from the graph-theoretical distances of all the vertices in a molecular graph from individual vertices of the graph and information content of individual vertices are computed and summed up to get an average information content measure for the molecular graph. To put it mathematically, let there are N' vertices in a connected molecular graph G_M . Also, let there be $(1, n'_1, n'_2, \dots, n'_{max})$ vertices at topological distances $0, 1, 2, \dots, d_{max}$ from v in G_M where 0 distance stands for the vertex under consideration and d_{max} is the maximum topological distance in G_M from v where n' vertices are situated. Thus, considering N' vertices of G^M be partitioned into disjoint classes on the basis of equivalence of their distances from vertex v with cardinalities $1, n'_1, \dots, n'_{max}$, information content of vertex v , say V_v^c , may be obtained using Equation (2) as:

$$V_v^c = \left[\left(\frac{1}{N'} \right) \log_2 \left(\frac{N'}{1} \right) \right] + \left[\left(\frac{n'_1}{N'} \right) \log_2 \left(\frac{N'}{n'_1} \right) \right] + \dots + \left[\left(\frac{n'_{max}}{N'} \right) \log_2 \left(\frac{N'}{n'_{max}} \right) \right] \quad (7)$$

The measure V^c is known as Vertex Complexity [9]. Now, if $V_{v_1}^c, V_{v_2}^c, \dots, V_{v_{N'}}^c$ be the vertex complexities of N' vertices of G_M , then the measure H^V for G_M , say $H^V(G_M)$, may be obtained using Equation (8):

$$H^V(G_M) = \left(\frac{1}{N'} \right) \sum_{i=1}^{N'} V_{v_i}^c \quad (8)$$

The index H^V is known as Graph Vertex Complexity [9].

Computation of H^V index

To illustrate the computation of H^V index, we have considered the hydrogen-filled connected molecular graph model of n -butane shown in Figure 2. The vertices of this molecular graph have been labelled for the convenience of identifying the vertices in computing their indices.

Now, for vertex 1, there are four vertices at (topological) distance 1, three vertices at distance 2, three vertices at distance 3 and three vertices at distance 4 from vertex 1 besides vertex 1 being at a distance 0 from itself. In this way, the 14 vertices of the molecular graph of n -butane is partitioned into five disjoint classes with cardinalities (1, 4, 3, 3, 3) from the equivalence of their topological distance from vertex 1 in the molecular graph of n -butane. Therefore, vertex complexity of vertex 1 may be obtained using Equation (7) as:

$$V_{v_1}^c = \left[\left(\frac{1}{14} \right) \log_2 \left(\frac{14}{1} \right) \right] + \left[\left(\frac{4}{14} \right) \log_2 \left(\frac{14}{4} \right) \right] + \left\{ 3 \times \left[\left(\frac{3}{14} \right) \log_2 \left(\frac{14}{3} \right) \right] \right\} = 2.2170$$

One can easily see in the molecular graph of n -butane that vertex 4 is in such a position that it is topologically similar with vertex 1. Therefore, V^c value of vertex 4 will be the same as that of vertex 1 i.e., 2.2170.

Thus, following the same procedure of finding vertices at different topological distances from a vertex under consideration, one can compute V^c index values for all the 14 vertices of the molecular graph of n -butane. These V^c values are given below:

$$V_{v_1}^c = V_{v_4}^c = 2.2170;$$

$$V_{v_2}^c = V_{v_3}^c = 1.7885;$$

$$V_{v_5}^c = V_{v_8}^c = V_{v_9}^c = V_{v_{12}}^c = V_{v_{13}}^c = V_{v_{14}}^c = 2.4486;$$

$$V_{v_6}^c = V_{v_7}^c = V_{v_{10}}^c = V_{v_{11}}^c = 2.0201.$$

It can be readily seen that there are other vertices in topologically equivalent positions with each other in the molecular graph of butane and therefore have got the same V^c index values.

Now, considering the V^c index values of 14 vertices of the molecular graph of n -butane, one can compute H^V index value for the molecular graph of n -butane using Equation (8) as:

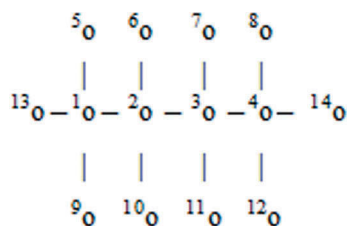


Figure 2. Hydrogen-filled molecular graph of n -butane with the vertices (shown by 'o') labelled by the preceding superscript numbers.

$$H^V(n\text{-butane}) = [(2 \times 2.2170) + (2 \times 1.7885) + (6 \times 2.4486) + (4 \times 2.0201)]/14 \\ = 2.1988.$$

Information (IB^{AT}) and total information (TIB^{AT}) on bonds

Let there are M covalent bonds between pairs of atoms in a compound C_1 where 1 is added for each single bond, 2 for each double bond and 3 for each triple bond. Such a representation of a chemical compound is a connected multigraph [10] which is a more appropriate representation of molecular structures that contain various types of covalent bonds connecting two atoms of different chemical nature (C, N, O, H etc.). Here, we intend to develop an information content measure based on the edges in a connected graph/multigraph where the vertices are labelled by chemical nature of the atoms they are representing. The idea is to get such a measure for various kinds of bond in molecular structures. Therefore, a single bond between two carbon atoms and that between a carbon atom and a nitrogen atom will be considered as different types of bonds which is the case in real situation. Such an information-theoretical measure will, thus, reflect the variety in bonding patterns present in a molecule.

Now, let there are m_i types of single edges, m_j types of double edges and m_k types of triple edges between different pairs of vertices in a molecular graph G_1 where i, j and k can have zero as well as non-zero integer values. Again, let $m_{11}, m_{12}, \dots, m_{1i}$ are ' i ' types of single edges, $m_{21}, m_{22}, \dots, m_{2j}$ are ' j ' types of double edges and $m_{31}, m_{32}, \dots, m_{3k}$ are ' k ' types of triple edges. Therefore, if M is the total number of edges (bonds) in a molecular graph present in the form of single, double and triple edges connecting vertices representing different types of pairs of atoms in G_1 , then M may be given by:

$$M = (m_{11} + m_{12} + \dots + m_{1i}) + (m_{21} + m_{22} + \dots + m_{2j}) \\ + (m_{31} + m_{32} + \dots + m_{3k}) \quad (9)$$

Now, considering M to be partitioned into several types of single, double and triple edges from the equivalence of edge type, one can compute information on bonds from atom type (IB^{AT}) for G_1 using Equation (2) as:

$$IB^{AT}(G_1) = \left\{ \left(\frac{m_{11}}{M} \log_2 \frac{M}{m_{11}} \right) + \left(\frac{m_{12}}{M} \log_2 \frac{M}{m_{12}} \right) + \dots + \left(\frac{m_{1i}}{M} \log_2 \frac{M}{m_{1i}} \right) \right\} \\ + \left\{ \left(\frac{m_{21}}{M} \log_2 \frac{M}{m_{21}} \right) + \left(\frac{m_{22}}{M} \log_2 \frac{M}{m_{22}} \right) + \dots + \left(\frac{m_{2j}}{M} \log_2 \frac{M}{m_{2j}} \right) \right\} \quad (10) \\ + \left\{ \left(\frac{m_{31}}{M} \log_2 \frac{M}{m_{31}} \right) + \left(\frac{m_{32}}{M} \log_2 \frac{M}{m_{32}} \right) + \dots + \left(\frac{m_{3k}}{M} \log_2 \frac{M}{m_{3k}} \right) \right\}$$

Now, since total information content [11] is obtained by multiplying the information content of a system by the number of elements in the system (in our case the total number of edges M in G_1), 'Total Information on Bonds from Atom Type (TIB^{AT})' for G_1 may be obtained using Equation (3) as:

$$TIB^{AT}(G_1) = M \times IB^{AT}(G_1) \quad (11)$$

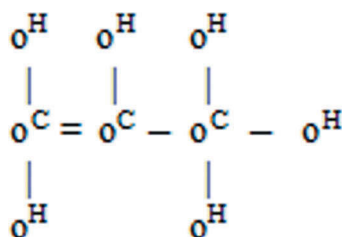


Figure 3. Hydrogen-filled multigraph model of Propylene where atom labels C and H have been assigned to the corresponding vertices indicated by 'o'.

Computation of IB^{AT} and TIB^{AT} indices

To illustrate the derivation and computation of the indices IB^{AT} and TIB^{AT} , we have considered the molecular graph model of propylene shown in Figure 3 where the vertices are labelled with the chemical nature of the corresponding atom (i.e., atom type).

Now, there are nine edges in the molecular multigraph of propylene. Among these edges, there are two edges representing a double bond and seven edges representing seven single bonds in this molecular graph. However, all the single bonds are not connecting the same type of atoms viz., one single bond is connecting two carbon atoms; whereas, there are six single bond which are connecting a carbon atom with a hydrogen atom. These two types of bonds are considered different for computing these two newly defined edge information indices. Therefore, the values of these information-theoretical indices for propylene may be computed as:

$$IE^{AT}(\text{propylene}) = \left(\frac{1}{9} \log_2 \frac{9}{1}\right) + \left(\frac{2}{9} \log_2 \frac{9}{2}\right) + \left(\frac{6}{9} \log_2 \frac{9}{6}\right) = 1.2244$$

$$IE^{AT}(\text{propylene}) = 9 \times 1.2244 = 11.0196$$

Results

In the present study, we have carried out statistical analyses with gas phase entropy (S°) values of the 170 hydrocarbons under consideration and the values of the indices I^Z , TI^Z , IB^{AT} , TIB^{AT} and H^V for these compounds using statistical software Minitab-18 [12]. We have divided 170 compounds into a training set and a test set keeping a reasonable ratio in the number of compounds in each set with 130 compounds in the training set and randomly chosen 40 compounds in the test set for external validation. We have then carried out stepwise regression analyses using the default parameters set in the software [12] which adds or, removes a term (independent variable) based on the parameters chosen and have found from this study with 130 training set compounds that the best correlation ($r^2 = 0.97$) for these hydrocarbons are obtained when TI^Z , TIB^{AT} and H^V indices are considered together. Subsequently, we predicted gas phase entropy values of the test set compounds from the corresponding regression equation obtained using the training set and the correlation between the experimental and predicted gas phase entropy values of the test set compounds has been found to be very high ($r^2 =$

0.975) indicating that the regression model comprising the three indices is capable of predicting gas phase entropy of the hydrocarbon classes under consideration with an acceptable outcome.

However, we have carried out a systematic correlation study with 130 training set compounds for individual indices as well as two or, more indices taken together from the selected five indices. From conceptual point of view, it is important to consider meaningful molecular descriptors as independent variables for developing predictive regression models. Keeping that in mind, we have considered three different types of meaningful information-theoretical indices which can have contributory roles to play in obtaining gas phase entropy values for the hydrocarbons since such indices take into account important/relevant molecular structural properties like molecular size (I^Z , TI^Z), skeletal branching and other topological characteristics of molecular structure along with some sort of group contribution aspect (H^V) and the varieties in the types of covalent bonds that chemical compounds can have between the atoms of different chemical nature (IB^{AT} , TIB^{AT}).

Thus, it seems important to investigate how these indices perform in statistical correlation studies. It is interesting to note that TI^Z , an index reflecting molecular size, alone gives a high correlation ($r^2 = 0.91$) and it is better than what is obtained ($r^2 = 0.77$) from the other index I^Z which too reflects molecular size. This finding seems to support the notion that molecular size is an important factor in obtaining (gas phase) entropy values of chemical compounds. At the same time the index H^V , reflecting skeletal branching and other topological characteristics of molecular structure together with some sort of group contribution aspect, also correlates reasonably well ($r^2 = 0.85$) with the entropy values and gives higher correlations ($r^2 = 0.94$ and $r^2 = 0.89$) when used along with TI^Z and I^Z respectively in multiple regression analyses. It is interesting to note, that the newly defined edge index TIB^{AT} alone can produce a very high correlation ($r^2 = 0.91$), indicating its usefulness in predicting gas phase entropy and it contributes to improved correlation ($r^2 = 0.97$) when used together with the indices TI^Z and H^V as independent variables in multiple regression equation and this betterment in the correlation value seems to be quite significant at this high level of correlation. It is interesting to note that although the other index for bonds (IB^{AT}) does not produce linear correlation with entropy values, it also helps improve the correlation ($r^2 = 0.94$) when used together with TI^Z and H^V indices. It appears, therefore, that the two newly defined information-theoretical indices for chemical bonds have a role to play in predicting gas phase entropy of the hydrocarbons under consideration using multiple regression models. Moreover, the least standard deviation ($s = 22.08$), obtained by considering three indices, TI^Z , H^V and TIB^{AT} together out of that obtained using other combinations of indices further supports the usefulness of these indices in predicting gas phase entropy of the studied hydrocarbons.

We, therefore, report here the results obtained from the statistical analyses carried out with the gas phase entropy values and the values of these three information-theoretical indices for the 130 training set hydrocarbons, considered for the present study, using statistical software Minitab-18 [12]. The observed and predicted gas phase entropy (S°) values, the residual of observed and predicted values and the values of the information indices TI^Z , H^V and TIB^{AT} of the studied hydrocarbon compounds (training set and test set) are given in Table 1. The important statistical

Table 1. Observed (Obs.) and predicted (Pred.) gas phase entropy (S°) values with residuals (Res.) and the values of the indices H^V , Tl^Z and TlB^{AT} for different classes of hydrocarbons.

	Compounds	PubChem CID	H^V	Tl^Z	TlB^{AT}	Gas phase entropy S° (J/mol.K)		
						Obs.	Pred.	Res.
<i>Training set</i>								
1	Propane	6334	1.93	75.68	7.22	269.91	275.3	-5.39
2	Butane	7843	2.2	110.93	10.13	310.12	317.08	-6.96
3	Pentane	8003	2.38	148.93	12.98	348.95	353.58	-4.63
4	Hexane	8058	2.58	189.13	15.8	388.40	393.57	-5.17
5	Heptane	8900	2.73	231.19	18.6	427.90	430.84	-2.94
6	Octane	356	2.88	274.85	21.39	466.73	469.39	-2.66
7	Decane	15,600	3.13	366.22	26.94	544.63	545.5	-0.87
8	Undecane	14,257	3.24	413.66	29.72	583.58	583.64	-0.06
9	Dodecane	8182	3.35	462.12	32.48	622.5	622.63	-0.13
10	Hexadecane	11,006	3.7	664.75	43.54	778.31	777.78	0.53
11	Tridecane	12,388	3.44	511.53	35.25	661.45	660.64	0.81
12	Pentadecane	12,391	3.62	612.9	40.78	739.35	738.69	0.66
13	Octadecane	11,635	3.85	770.54	49.07	856.21	856.73	-0.52
14	1,2-Pentadiene	11,588	2.37	121.87	19.3	333.46	313.3	20.16
15	1,3-Butadiene	7845	2.25	85.17	14.54	278.74	288.44	-9.70
16	1,4-Pentadiene	11,587	2.38	121.87	19.3	333.46	314.15	19.31
17	1-Butene	7844	2.21	97.96	15.02	305.60	293.62	11.98
18	1-Decene	13,381	3.11	350.66	34.77	540.45	508.94	31.51
19	1-Heptene	11,610	2.7	216.64	25.33	423.59	397.42	26.17
20	1-Hexene	11,597	2.56	175.02	22.04	384.64	362.76	21.88
21	1-Nonene	31,285	2.98	304.65	31.67	501.49	471.15	30.34
22	1-Octene	8125	2.86	259.92	28.53	462.54	435.32	27.22
23	1-Pentene	8004	2.37	135.33	18.63	345.81	325.73	20.08
24	Ethylene	6325	1.77	32.98	5.51	219.45	233.52	-14.07
25	Propylene	8252	1.97	63.51	11.02	266.94	258.16	8.78
26	Allene	10,037	2.00	51.58	8.00	243.93	260.18	-16.25
27	2-Methyl propene	8255	1.89	97.96	15.02	293.59	266.59	27.00
28	2-Methyl-1-pentene	12,986	2.36	175.02	22.04	382.17	345.87	36.30
29	2-Methyl-2-butene	10,553	2.19	135.33	18.63	338.57	310.53	28.04
30	1-Butyne	7846	2.14	85.17	15.79	290.83	275.51	15.32
31	1-Decyne	12,997	3.04	335.16	37.07	524.51	484.27	40.24
32	1-Hexadecyne	12,396	3.63	630.98	55.8	758.22	709.95	48.27
33	1-Hexyne	12,732	2.48	161.03	23.51	368.74	340.84	27.90
34	1-Octyne	12,370	2.78	245.08	30.47	446.64	411.38	35.26
35	1-Pentyne	12,309	2.28	121.87	19.79	329.78	304.28	25.50
36	2-Butyne	10,419	2.2	85.17	15.79	283.3	280.58	2.72
37	Acetylene	6326	1.75	22.28	4.85	200.83	225.42	-24.59
38	Propyne	6335	1.91	51.58	11.25	248.11	243.14	4.97
39	3-Methyl-1-butyne	69,019	2.12	121.87	19.79	318.95	290.77	28.18
40	Butadiyne	9997	2.25	60.17	11.02	250.04	279.20	-29.16
41	2,3,4-Trimethyl pentane	11,269	2.33	274.85	21.39	428.07	422.94	5.13
42	2,3-Dimethyl hexane	11,447	2.53	274.85	21.39	443.96	439.83	4.13
43	2,3-Dimethyl pentane	11,260	2.32	231.19	18.6	414.05	396.22	17.83
44	2,4-Dimethyl hexane	11,511	2.53	274.85	21.39	445.64	439.83	5.81
45	2,5-Dimethyl hexane	11,592	2.56	274.85	21.39	439.03	442.37	-3.34
46	2-Methyl butane	6556	2.16	148.93	12.98	343.59	335.00	8.59
47	2-Methyl hexane	11,582	2.56	231.19	18.6	419.99	416.49	3.50
48	2-Methyl nonane	13,379	3.00	366.22	26.94	534.46	534.52	-0.06
49	2-Methyl octane	18,591	2.87	319.91	24.17	495.89	495.54	0.35
50	2-Methyl pentane	7892	2.35	189.13	15.8	380.53	374.15	6.38
51	2-Methyl propane	6360	1.87	110.93	10.13	294.64	289.22	5.42
52	3-Ethyl-2-methyl pentane	11,863	2.32	274.85	21.39	441.12	422.10	19.02
53	3-Ethyl hexane	12,096	2.55	274.85	21.39	458.19	441.52	16.67
54	3-Ethyl pentane	12,048	2.35	231.19	18.6	411.5	398.75	12.75
55	4-Methyl heptane	11,512	2.66	274.85	21.39	453.34	450.81	2.53
56	2,2,3,3-Tetra methyl butane	11,675	2.07	274.85	21.39	389.36	400.99	-11.63
57	2,2,3-Trimethyl pentane	11,255	2.27	274.85	21.39	425.18	417.88	7.30
58	2,2,4-Trimethyl pentane	10,907	2.29	274.85	21.39	423.21	419.56	3.65

(Continued)

Table 1. (Continued).

	Compounds	PubChem CID	H^V	Tf^Z	TfB^{AT}	Gas phase entropy S° (J/mol.K)		
						Obs.	Pred.	Res.
59	2,2-Dimethyl butane	6403	2.06	189.13	15.8	358.23	349.66	8.57
60	2,2-Dimethyl hexane	11,551	2.49	274.85	21.39	431.2	436.45	-5.25
61	2,2-Dimethyl propane	10,041	1.76	148.93	12.98	306.39	301.23	5.16
62	2,3,3-Trimethyl pentane	11,215	2.24	274.85	21.39	431.54	415.34	16.20
63	3,3-Dimethyl hexane	11,233	2.43	274.85	21.39	438.06	431.39	6.67
64	3-Ethyl-3-methyl pentane	14,018	2.26	274.85	21.39	432.96	417.03	15.93
65	3,3-Diethyl pentane	14,020	2.26	319.91	24.17	461.54	444.03	17.51
66	1,1-Dimethyl cyclohexane	11,549	2.25	259.92	22.04	365.01	402.67	-37.66
67	1-Methyl cyclopentene	12,746	2.13	161.03	22.66	326.35	313.75	12.60
68	3-Methyl cyclopentene	14,263	2.12	161.03	22.66	330.54	312.91	17.63
69	4-Methyl cyclopentene	15,658	2.07	161.03	22.66	328.86	308.69	20.17
70	Butyl cyclohexane	15,506	2.81	350.66	27.55	458.48	504.59	-46.11
71	Cyclobutane	9250	1.96	97.96	11.02	265.39	284.13	-18.74
72	Cyclobutene	69,972	1.96	85.17	15.79	263.51	260.31	3.20
73	Cycloheptane	9265	2.26	216.64	19.28	342.33	377.85	-35.52
74	Cyclohexene	8079	2.22	161.03	22.66	310.75	321.35	-10.60
75	Cyclooctane	9266	2.48	259.92	22.04	366.77	422.1	-55.33
76	Cyclopentane	9253	1.95	135.33	13.77	292.88	304.39	-11.51
77	Cyclopentene	8882	1.95	121.87	19.3	289.66	277.83	11.83
78	Cyclopropane	6351	1.63	63.51	8.26	237.44	237.47	-0.03
79	Decyl cyclopentane	137,211	3.44	596.18	41.32	689.9	708.9	-19.00
80	Ethyl cyclohexane	15,504	2.47	259.92	22.04	382.58	421.25	-38.67
81	Heptyl cyclopentane	138,541	3.12	446.04	33.06	573.04	589	-15.96
82	Hexyl cyclopentane	138,257	2.99	397.82	30.3	534.13	548.51	-14.38
83	Methyl cyclohexane	7962	2.29	216.64	19.28	343.34	380.38	-37.04
84	Nonyl cyclopentane	137,755	3.34	545.29	38.57	650.95	668.83	-17.88
85	Octyl cyclopentane	137,210	3.23	495.22	35.81	611.99	628.59	-16.60
86	Pentyl cyclohexane	20,284	2.95	397.82	30.3	497.44	545.13	-47.69
87	Propyl cyclohexane	15,505	2.65	304.65	24.79	419.53	463.28	-43.75
88	Propyl cyclopentane	16,270	2.54	259.92	22.04	417.27	427.16	-9.89
89	1,2,3,4-Tetramethyl benzene	10,263	2.5	304.4	39.92	416.52	406.44	10.08
90	1,2,3,5-Tetramethyl benzene	10,695	2.46	304.4	39.92	422.54	403.06	19.48
91	1,2,3-Trimethyl benzene	10,686	2.4	259.34	36	384.84	374.31	10.53
92	1,2,4,5-Tetramethyl benzene	7269	2.59	304.4	39.92	418.53	414.04	4.49
93	1,2,4-Triethyl benzene	13,415	2.83	398.15	47.37	518.15	485.64	32.51
94	1,2,4-Trimethyl benzene	7247	2.5	259.34	36	395.76	382.76	13.00
95	1,3,5-Triethyl benzene	7602	2.74	398.15	47.37	507.69	478.04	29.65
96	1,3,5-Trimethyl benzene	7947	2.36	259.34	36	385.3	370.93	14.37
97	1,3-Dimethyl benzene	7929	2.38	215.68	31.9	357.69	350.55	7.14
98	1,4-Dimethyl benzene	7809	2.48	215.68	31.9	352.42	359	-6.58
99	1-Methyl-4-ethyl benzene	12,160	2.64	259.34	36	398.9	394.58	4.32
100	Benzene	241	2.24	133.42	22.83	269.2	301.06	-31.86
101	Butyl benzene	7705	2.83	304.4	39.92	439.49	434.31	5.18
102	Dodecyl benzene	31,237	3.66	701.8	67.97	751.57	732.24	19.33
103	Ethyl benzene	7500	2.51	215.68	31.9	360.45	361.53	-1.08
104	Hexaethyl benzene	11,791	2.78	701.8	67.97	697.14	657.92	39.22
105	Hexamethyl benzene	6908	2.56	398.15	47.37	452.37	462.84	-10.47
106	Isopropyl benzene	7406	2.5	259.34	36	388.57	382.76	5.81
107	Nonyl benzene	14,126	3.4	546.3	57.91	634.75	618.47	16.28
108	Octyl benzene	16,607	3.3	496.02	54.46	595.8	580.91	14.89
109	Pentaethyl benzene	11,794	2.79	597.39	61.31	647.89	596.84	51.05
110	Pentyl benzene	10,864	2.96	350.71	43.7	478.94	470.35	8.59
111	Propyl benzene	7668	2.68	259.34	36	400.66	397.96	2.70
112	Undecyl benzene	23,194	3.58	649.24	64.66	712.62	694.19	18.43
113	Ethynyl benzene	10,821	2.45	186.69	36.94	321.67	319.25	2.42
114	1,2-Dimethyl naphthalene	11,317	2.64	350.84	46.97	406.81	433.92	-27.11
115	1,4-Dimethyl naphthalene	11,304	2.61	350.84	46.97	401.08	431.39	-30.31
116	1,5-Dimethyl naphthalene	11,306	2.61	350.84	46.97	401.08	431.39	-30.31
117	1-Butyl naphthalene	15,414	2.96	447.25	55.68	497.18	510.68	-13.5
118	1-Ethyl naphthalene	14,315	2.69	350.84	46.97	418.15	438.15	-20.00

(Continued)

Table 1. (Continued).

	Compounds	PubChem CID	H^V	Tf^Z	TfB^{AT}	Gas phase entropy S° (J/mol.K)		
						Obs.	Pred.	Res.
119	1-Methyl naphthalene	7002	2.56	304.23	42.29	377.44	404.49	-27.05
120	1-Pentyl naphthalene	518,732	3.09	496.88	59.8	536.64	548.32	-11.68
121	2,3-Dimethyl naphthalene	11,386	2.72	350.84	46.97	410.95	440.68	-29.73
122	2,6-Dimethyl naphthalene	11,387	2.78	350.84	46.97	408.69	445.75	-37.06
123	2-Butyl naphthalene	14,339	3.08	447.25	55.68	499.82	520.82	-21.00
124	2-Ethyl-6-methyl naphthalene	10,942,881	2.92	398.54	51.41	455.18	481.8	-26.62
125	2-Ethyl naphthalene	13,652	2.8	350.84	46.97	420.74	447.44	-26.70
126	2-Methyl naphthalene	7055	2.65	304.23	42.29	380.03	412.09	-32.06
127	2-Propyl naphthalene	519,754	2.94	398.54	51.41	460.99	483.49	-22.50
128	Biphenyl	7095	2.80	335.20	44.99	392.67	441.02	-48.35
129	1,7-Dimethyl naphthalene	11,326	2.66	350.84	46.97	409.45	435.61	-26.16
130	2-Ethyl-7-methyl naphthalene	11,095,066	2.88	398.54	51.41	455.18	478.42	-23.24
<i>Test set</i>								
1	Ethane	6324	1.71	44.04	4.14	229.49	241.04	-11.55
2	Nonane	8141	3.01	319.91	24.17	505.68	507.37	-1.69
3	Tetradecane	12,389	3.53	561.81	38.02	700.40	699.34	1.06
4	Heptadecane	12,398	3.77	717.31	46.31	817.26	816.56	0.70
5	1,2-Butadiene	11,535	2.23	85.17	14.54	293.01	286.75	6.26
6	1-Hexadecene	12,395	3.68	647.84	52.78	774.12	736.07	38.05
7	2,3-Pentadiene	136,378	2.44	121.87	19.3	324.68	319.21	5.47
8	2-Methyl-1-butene	11,240	2.18	135.33	18.63	339.53	309.68	29.85
9	1-Heptyne	12,350	2.62	202.2	27.05	407.69	374.42	33.27
10	1-Nonyne	18,937	2.91	289.46	33.81	485.6	447.2	38.40
11	2-Pentyne	12,310	2.39	121.87	19.79	331.79	313.57	18.22
12	2,3-Dimethyl butane	6589	2.16	189.13	15.8	365.77	358.1	7.67
13	2,4-Dimethyl pentane	7907	2.35	231.19	18.6	396.64	398.75	-2.11
14	2-Methyl heptane	11,594	2.71	274.85	21.39	455.26	455.03	0.23
15	3,4-Dimethyl hexane	11,412	2.49	274.85	21.39	448.32	436.45	11.87
16	3-Methyl heptane	11,519	2.69	274.85	21.39	461.58	453.34	8.24
17	2,2,3-Trimethyl butane	10,044	2.1	231.19	18.6	383.6	377.64	5.96
18	2,2-Dimethyl pentane	11,542	2.27	231.19	18.6	392.88	392	0.88
19	3,3-Dimethyl pentane	11,229	2.22	231.19	18.6	399.7	387.77	11.93
20	1,1-Dimethyl cyclopentane	15,421	2.06	216.64	19.28	359.28	360.96	-1.68
21	Butyl cyclopentane	16,269	2.7	304.65	24.79	456.22	467.50	-11.28
22	Cyclohexane	8078	2.23	175.02	16.53	298.24	350.91	-52.67
23	Ethyl cyclopentane	15,431	2.34	216.64	19.28	378.32	384.6	-6.28
24	Methyl cyclopentane	7296	2.1	175.02	16.53	339.91	349.22	-9.31
25	Pentyl cyclopentane	19,540	2.86	350.66	27.55	495.18	508.81	-13.63
26	1,2,3-Triethyl benzene	39,149	2.67	398.15	47.37	507.23	472.13	35.10
27	1,2-Dimethyl benzene	7237	2.37	215.68	31.9	352.75	349.71	3.04
28	1,4-Diethyl benzene	7734	2.81	304.4	39.92	434.01	432.62	1.39
29	Decyl benzene	7716	3.49	597.39	61.31	673.71	655.96	17.75
30	Heptyl benzene	14,115	3.2	446.61	50.95	556.85	544.2	12.65
31	Hexyl benzene	14,109	3.09	398.15	47.37	517.9	507.59	10.31
32	Pentamethyl benzene	12,784	2.54	350.71	43.7	443.88	434.88	9.00
33	Toluene	1140	2.32	173.62	27.55	320.66	325.39	-4.73
34	1,3-Dimethyl naphthalene	11,327	2.6	350.84	46.97	409.45	430.55	-21.10
35	1,6-Dimethyl naphthalene	11,328	2.67	350.84	46.97	409.45	436.46	-27.01
36	1-Propyl naphthalene	33,800	2.82	398.54	51.41	458.36	473.35	-14.99
37	2-Ethyl-3-methyl naphthalene	13,070,203	2.79	398.54	51.41	457.44	470.82	-13.38
38	2-Pentyl naphthalene	523,053	3.19	496.88	59.8	539.28	556.77	-17.49
39	Naphthalene	931	2.54	258.85	37.31	335.64	381.95	-46.31
40	2,7-Dimethyl naphthalene	11,396	2.73	350.84	46.97	408.69	432.81	-24.12

results along with the multiple regression equation obtained from Minitab-18 software [12] are given below:

$$S^{\circ} = 74.4 + 0.78 (TI^Z) + 84.45 (H^V) - 2.91 (TIB^{AT}) \quad (12)$$

$n = 130$, $r^2 = 0.9699$, r^2 (pred.) = 0.9676, $s = 22.08$, F -value = 1351.25

Here, n is the number of data points (compounds) used, r^2 is the coefficient of determination, r^2 (pred.) is obtained by systematically removing one compound from the data set at a time and making the prediction for this compound by estimating the regression model with the remaining compounds; s is the standard deviation, F -value is the F statistic of regression.

In order to explore further the predictive power of the three descriptors for data sets of different sizes, we carried out their correlation with gas phase entropy values of all the 170 compounds considered (training set and test set compounds taken together) and the study has returned an equally high correlation ($r^2 = 0.97$) which was obtained for 130 training set compounds. This indicates that the three indices together are capable of making acceptable gas phase entropy prediction for the hydrocarbon classes under consideration even if the size of the data set/training set is increased/decreased.

Discussion

The statistical significance of this result is clear from the magnitude of correlation and the F -value which is highly significant. Moreover, the extreme closeness in the values of r^2 and r^2 (pred.) indicates that the regression model is not overfitted and has acceptable predictive power for a compound not in the data set. The acceptable prediction of gas phase entropy values of the studied compounds is also apparent from the standard deviation ($s = 22.08$) which may be regarded as a low value for a data set of 130 training set compounds most of which have high gas phase entropy (S°) values and only 7 out of 130 compounds have got large residuals as reported by Minitab-18 statistical software [12] and the predictions for many compounds from different classes of hydrocarbon are well within the standard deviation value (Table 1) particularly those for the alkanes and substituted alkanes. Furthermore, the high correlation ($r^2 = 0.975$) between the experimental and predicted gas phase entropy values for the test set compounds further supports the predictive power of the regression model containing three relevant information-theoretical molecular descriptors. Moreover, the finding that the high correlation ($r^2 = 0.97$) which is obtained even when the size of the data set is changed seems to indicate the usefulness of these three indices in predicting gas phase entropy for the studied compounds.

Also, one of the advantages of the present study over the method of Domalski et al. [2] (from where the data have been taken) is that in the present work we have developed statistically significant regression model using molecular descriptors for the prediction of gas phase entropy and the descriptors can be easily computed for any molecule; whereas the method of Domalski et al. [2] is based on adding several group contribution values to predict gas phase entropy of the hydrocarbons studied although their group additivity approach [2] also gives the predicted values very close to the experimentally determined values for most of the compounds studied. The present approach also has another advantage in that one can work conveniently with large number of compounds at a time while other methods based on fundamental approaches [1,13] are usually done only for

a limited number of compounds having the predicted values close to experimentally determined values. On the other hand, other approaches based on regression analyses [3,6] have been used to carry out studies with various kinds of organic compounds that include both hydrocarbon and non-hydrocarbons compounds using different kinds of descriptors which is beyond the scope of the present study. However, such indices too have produced high correlation in the respective studies [3,6].

It is, therefore, quite apparent from the present study that the three information-theoretical indices, Tf^Z , H^V and TIB^{AT} , translating various relevant structural properties of chemical compounds, taken together (Equation (12)) can predict gas phase entropy of hydrocarbons in a statistically significant manner. While, Tf^Z has produced high correlation with gas phase thermal entropy values in an earlier study [4], H^V and the newly defined information-theoretical indices on edges, particularly the index TIB^{AT} , have also contributed in obtaining higher correlations and may emerge as useful molecular descriptors for explaining gas phase entropy of hydrocarbons. It would also be interesting to see how the index H^V can be related to group contribution so that its usefulness as an information-theoretical molecular descriptor can be explored further in explaining gas phase entropy of hydrocarbons. It also seems to be worth noting that the method of computing information-theoretical indices from the partition of the elements of a system (molecular system in our case) on the basis of an equivalence relation defined on them with respect to some molecular structural properties may be an underlying factor for obtaining significant correlation with gas phase entropy of hydrocarbons which may be perceived more of an additive-constitutive type of thermodynamic property than a purely additive one. Moreover, the mathematical results on information-theoretical measures obtained earlier from our studies [4,5] may further help the analysis of the increase or, decrease in gas phase entropy values with the change in molecular structures which are encoded by the information-theoretical molecular descriptors used in the regression model. Such studies, thus, may help one interpret gas phase entropy of hydrocarbons and perhaps thermodynamic entropy of chemical compounds in general in terms of information-theoretical formalism.

Acknowledgements

We thankfully acknowledge the financial support obtained from the Department of Science and Technology (DST), Government of India, New Delhi, for carrying out the present study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Science and Engineering Research Board [EMR/2015/001977].

ORCID

D. Pal  <http://orcid.org/0000-0002-3591-5978>

References

- [1] D.F. DeTar, *Calculation of entropy and heat capacity of organic compounds in the gas phase. Evaluation of a consistent method without adjustable parameters. Applications to hydrocarbons*, J. Phys. Chem. A 111 (2007), pp. 4464–4477. doi:10.1021/jp066312r.
- [2] E.S. Domalski and E.D. Hearing, *Estimation of the thermodynamic properties of hydrocarbons at 298.15 K*, J. Phys. Chem. Ref. Data 17 (1988), pp. 1637–1678. doi:10.1063/1.555814.
- [3] L. Mu and H. He, *Quantitative structure–property relations (QSPRs) for predicting the standard absolute entropy ($S_{298 K}$) of gaseous organic compounds*, Ind. Eng. Chem. Res. 50 (2011), pp. 8764–8772. doi:10.1021/ie2003335.
- [4] C. Raychaudhury and D. Pal, *Information content of molecular graph and prediction of gas phase thermal entropy of organic compounds*, J. Math. Chem. 51 (2013), pp. 2718–2730. doi:10.1007/s10910-013-0233-9.
- [5] C. Raychaudhury and D. Pal, *Information content measures and prediction of physical entropy of organic compounds*, in *Mathematical Foundations and Applications of Graph Entropy*, M. Dehmer, F. Emmert-Streib, Z. Chen, X. Li, and Y. Shi, eds., Wiley-VCH, Weinheim, 2016, pp. 233–258.
- [6] L. Zhao, P. Li, and S.H. Yalkowsky, *Predicting the entropy of boiling for organic compounds*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 1112–1116. doi:10.1021/ci990054w.
- [7] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, USA, 1949.
- [8] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures*, Research Studies Press, Chichester, UK, 1983.
- [9] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, and S.C. Basak, *Discrimination of isomeric structures using information theoretic topological indices*, J. Comput. Chem. 5 (1984), pp. 581–588. doi:10.1002/jcc.540050612.
- [10] F. Harary, *Graph Theory*, Addison-Wesley, Reading, MA, USA, 1969.
- [11] L. Brillouin, *Science and Information Theory*, Academic Press, New York, 1956.
- [12] Minitab-18.1, *Minitab Statistical Software*, PA, USA, 2017.
- [13] M.K. Sabbe, F. De Vleeschouwer, M.F. Reyniers, M. Waroquier, and G.B. Marin, *First principles based group additive values for the gas phase standard entropy and heat capacity of hydrocarbons and hydrocarbon radicals*, J. Phys. Chem. A 112 (2008), pp. 12235–12251. doi:10.1021/jp807526n.