MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **79** (2018) 659-683

ISSN 0340 - 6253

On Distance–Based Topological Indices Used in Architectural Research[☆]

Sanja Stevanović[*], Dragan Stevanović

Mathematical Institute, Serbian Academy of Sciences and Arts, Knez Mihajlova 36, 11000 Belgrade, Serbia
sanja_stevanovic@yahoo.com, dragan_stevanovic@mi.sanu.ac.rs

(Received November 19, 2017)

Abstract

Distance-based topological indices have been used in studies of molecular graphs ever since Harry Wiener introduced his now famous index back in the 1940s, with tens of such indices studied actively in current mathematical chemistry literature. Interestingly, two further distance-based invariants, the difference factor and the intelligibility, have been used in parallel in studies of graphs associated to building and urban plans since the 1970s as well. These invariants are defined in terms of integration values that represent normalized values of the sums of distances from a given vertex to all other vertices in a graph. The difference factor is defined as an entropic measure that quantifies the diversity of the sequence of integration values, while the intelligibility is defined as the Pearson correlation coefficient between sequences of vertex degrees and integration values thus quantifying the extent to which integration values, for which one has to know the structure of the whole graph, can be predicted from vertex degrees, for which one has to know only how many neighbors a vertex has. We perform here a number of computational studies of the difference factor and the intelligibility that reveal to what extent these invariants can be used as topological indices in mathematical chemistry as well.

1 Introduction

Topological index is an invariant of a molecular graph, whose aim is to provide easily calculable way of approximating a physicochemical property that is correlated with the structure of the molecular graph. The oldest topological index is the Wiener index,

[☆]This work was supported by the research project ON174033 of the Ministry of Education and Science of the Republic of Serbia.

[*]Corresponding author

the sum of distances between all pairs of vertices in a graph, that had been shown by Harry Wiener to correlate well with various properties of alkanes in a series of papers in 1947 and 1948 [1–4]. The Wiener index attracted attention of mathematicians in the late 1970s when it was introduced in graph theory under the names distance of a graph and transmission of a graph [5–7], and had been further studied also in the form of average distance of graphs [8]. Nowadays it is widely used in quantitative structure-activity relationship studies, and many of its properties are surveyed in [9–11].

Other early and well studied topological indices include the Hosoya Z index [12], the first and the second Zagreb indices [13], the Randić index [14], the molecular connectivity indices [15–17], the Balaban J index [18] and the information content indices [19]. Many further topological indices have emerged in the meantime, so that, for example, Dragon [20], a QSAR software with the most extensive list of implemented molecular descriptors, nowadays contains routines to calculate more than a thousand topological indices.

However, these new topological indices are not automatically deemed useful. Several researchers had discussed requirements that a satisfactory topological index should satisfy. Randić [21] listed desirable attributes for topological indices, shown in Table 1, that were proposed in analogy to the Read's list [22] of requirements for codings of molecules.

A topological index should:
• have structural interpretation
• have good correlation with at least one property
• preferably discriminate among isomers
• be locally defined
• be generalizable to "higher" analogues
• be preferably independent from other descriptors
• be simple
• not be based on properties
• not be trivially related to other descriptors
• be possible to construct efficiently
• be based on familiar structural concepts
• show a correct size dependence
• change gradually with gradual change in structures

Table 1. Randić's list of desirable attributes for topological indices [21].

In addition to efficient computation, Balaban [23] further stated that a good topological index should have low, or no, degeneracy and should show good correlation with the degree of molecular branching. Degeneracy of a topological index in a set of graphs is

measured in terms of the number of values that appear as values of the topological index for at least two nonisomorphic graphs in that set. Correlation with molecular branching, on the other hand, means that the value of the topological index should increase with molecular branching, so that among trees with given number of vertices, for example, the extremal values of the topological index should be attained at a path and a star, respectively. Note, however, that some of these requirements are stated rather intuitively. For example, Furtula, Gutman and Dehmer [24] have only recently offered a formal definition for the Randić's requirement for "a gradual change of a topological index with a gradual change in structure" in terms of the so-called structure sensitivity and abruptness.

On the other hand, as basic mathematical structures used to visually describe relations within a set of objects, graphs had found numerous applications in many different sciences. These applications started to appear abundantly after the publication of Harary's influential textbook on graph theory [25] in 1969. Early uses of graphs in architecture and urbanism appear, for example, in works by Alexander [26, 27], March and Steadman [28–30], Krüger [31] and Hillier and Hanson [32]. In later publications, Hillier and Hanson and their coworkers expanded their use of graphs in architecture and urbanism to a whole new field called the *space syntax* (see [33] for its overview). In space syntax graphs are primarily used to study people movement tendencies within a building or a settlement, so that most studies are based on distances among vertices in such graphs. Researchers are usually interested to compare spaces within a building between each other, in which case they employ local distance invariants, such as the sum of distances from a vertex to all other vertices in a graph or a variation of it known as the *integration value*. However, when it is needed to compare different buildings or settlements, researchers rely on two global distance invariants called the *difference factor* and the *intelligibility*, both of which were introduced in 1987. The difference factor [34] is an entropy-like measure that quantifies the relation between the minimum, average and maximum integration values, while the intelligibility [35] represents the coefficient of correlation between vertex degrees and their integration values. Hillier et al. [35] introduced intelligibility to denote the strength of predicting integration values, for which the structure of the whole graph structure must be known, solely from vertex degrees, for which only the number of immediate neighbors has to be known (which is easily observable by a person standing in a given building or a settlement). Hence if a graph has high intelligibility, then "the whole can be read from

the parts", in the words of Hillier et al. [35].

The purpose of this paper is to test to what extent difference factor and intelligibility, distance-based invariants that are used in space syntax research for the last 30 years, but do not appear to have been mentioned in mathematical chemistry literature, satisfy desired requirements for topological indices. In the next section we rephrase definitions of these invariants in standard graph theoretical terminology. In subsequent sections we perform computational studies that test adequacy of using the difference factor, the entropy of integration values and the intelligibility as topological indices: in Section 3 we test their usability in predicting physicochemical properties of octanes and correlations to other distance-based indices studied in the literature, then in Section 4 we study the level of degeneracy of these indices, while in Section 5 we report on their extremal graphs for various types of graphs. Conclusions are given in Section 6.

2 Difference factor and intelligibility

There are a couple of ways in space syntax to associate a graph to a given building or a settlement. Each of them first represents a spatial layout through a pattern of basic spatial elements and then represents configurational relations between spatial elements as a graph [36]. In case of justified plan graphs the layout may be partitioned, depending on various definitions, either into convex spaces or into functional units. Each of these elements is then represented as a vertex in a graph, with two vertices being adjacent if there is a direct passage from one to the other element. An example of a justified plan graph for one of the Frank Lloyd Wright's great residential designs, the Arthur Heurtley house, is shown in Fig. 1.b). Axial maps, on the other hand, are obtained from the set of the fewest and longest straight lines required to cover every convex space in a plan, while ensuring that all non-trivial circulation loops in the plan are made. The process of selecting such set of sight-lines is explained in detail in [37], while the result of such process for the same Frank Lloyd Wright design is shown in Fig. 1.c). The actual axial map is obtained by assigning a vertex to each axial line, with two vertices being adjacent if the corresponding axial lines intersect each other (a rather dense axial map for the same design is shown in Fig. 1.d). Further details on other approaches for creating graphs in space syntax may be found, for example, in [38] and the references cited therein.

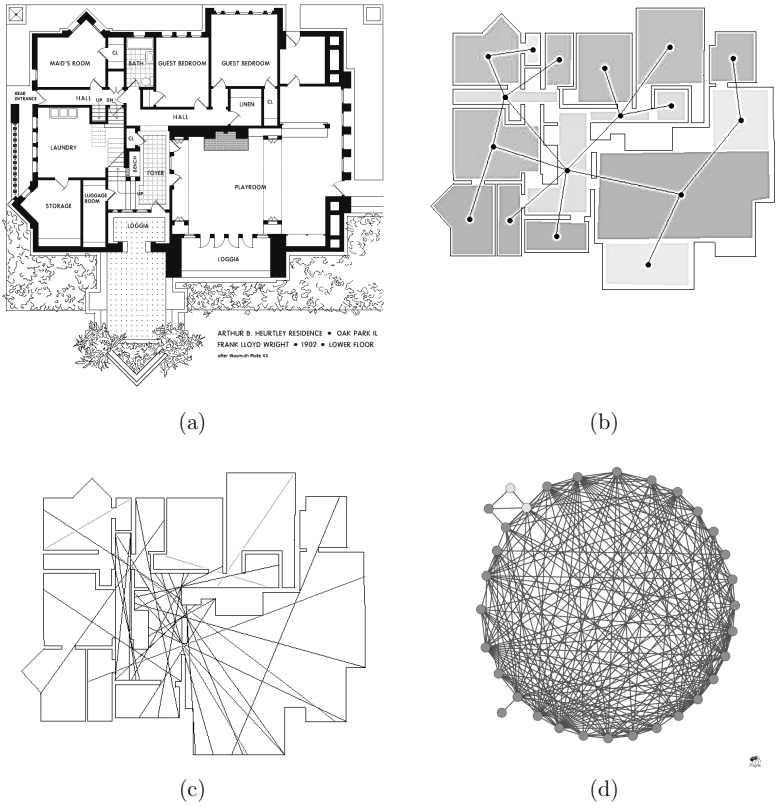

Figure 1. Justified plan graph and axial map associated to a building floor plan: (a) Lower floor plan of the Arthur Heurtley residence, designed by Frank Lloyd Wright in 1902 [39]; (b) Justified plan graph corresponding to the main functional units; (c) The set of axial lines; (d) Axial graph corresponding to the reduced set of axial lines.

We will now move on to rephrase difference factor and intelligibility from their original space syntax definitions to more familiar graph theoretical terms. Let $G = (V, E)$ be a simple graph with $n = |V|$ vertices and $m = |E|$ edges. For a vertex $u \in V$ let d_u denote its degree, i.e., the number of vertices in V that are joined by an edge to u. For two vertices $u, v \in V$ let $d(u, v)$ denote the distance between them, defined as the number of edges in a shortest path between u and v. Note that $d(u, v) = 0$ if $u = v$. For a given vertex $u \in V$, its total depth TD_u, usually called transmission in graph theoretical

literature, is defined as

$$TD_u = \sum_{v \in V \setminus \{u\}} d(u,v).$$

Note that from here for the Wiener index $W = \sum_{\{u,v\} \in \binom{V}{2}} d(u,v)$ we also have

$$W = \frac{1}{2} \sum_{u \in V} TD_u. \tag{1}$$

The mean depth MD_u is the average distance between u and other vertices in V,

$$MD_u = \frac{1}{n-1} \sum_{v \in V \setminus \{u\}} d(u,v) = \frac{TD_u}{n-1}.$$

Integration value i_u is introduced as a way of normalizing MD_u value to the range $[0,1]$:

$$\begin{aligned} i_u = \frac{2[MD_u - 1]}{n-2} &= \frac{2}{n-2} \left(-1 + \frac{1}{n-1} \sum_{v \in V \setminus \{u\}} d(u,v) \right) \\ &= \frac{2}{(n-1)(n-2)} \sum_{v \in V \setminus \{u\}} [d(u,v) - 1] \\ &= \frac{1}{\binom{n-1}{2}} \sum_{v \in V \setminus \{u\}} [d(u,v) - 1] \qquad (2) \\ &= \frac{TD_u - n + 1}{\binom{n-1}{2}}. \qquad (3) \end{aligned}$$

Since $d(u,v) \geq 1$ for each $v \neq u$, i_u is necessarily nonnegative from (2). On the other hand, if the maximum distance from u to all other vertices of V is D, then the sum $\sum_{v \in V \setminus \{u\}} [d(u,v) - 1]$ necessarily contains summands $0, 1, \ldots, D-1$, while its remaining summands may be bounded from above by $D, D+1, \ldots, n-2$. Hence

$$\sum_{v \in V \setminus \{u\}} [d(u,v) - 1] \leq 0 + 1 + \cdots + (n-2) = \binom{n-1}{2},$$

so that $i_u \leq 1$ from (2).

Next, let a, b and c denote, respectively, the minimum, the average and the maximum integration value of vertices in V. Further, let $t = a + b + c$ and $a' = \frac{a}{t}$, $b' = \frac{b}{t}$ and $c' = \frac{c}{t}$. The difference factor H is then defined as

$$H = -a' \ln a' - b' \ln b' - c' \ln c'.$$

Note that, in case one or two of a', b' and c' are zero, we can define that $0 \ln 0 = 0$ since $\lim_{x \to 0+} x \ln x = 0$, so that H is defined in these cases as well. On the other hand, if

$a = b = c = 0$ then $t = 0$ which leaves a', b', c', and consequently H, undefined. However, this happens only if G is a complete graph for which each vertex has integration value of 0. Hence we will exclude complete graphs when considering difference factor in the sequel.

As stated in [34], the difference factor is introduced "as an adaptation of the Shannon entropy [40] to quantify the degree of difference between integration values of any three spaces," although we can see from the above formula that these three spaces are not really arbitrary. The Shannon entropy is one of the most widely used indices of diversity, which for a vector $x = (x_1, \ldots, x_n)$, $x_i \geq 0$, $\sum_i x_i = 1$, is defined as $H_x = -\sum_i x \ln x$. However, Hillier, Hanson and Graham [34] apply the Shannon entropy to three particular values only, which they use to describe the spread of the set of integration values of all vertices. They do not offer explanations why the entropy is not calculated for all integration values at once. As a matter of fact, the entropy of all integration values is conceptually simpler invariant than the difference factor, which requires calculating the maximum, minimum and average of integration values first, and then taking the entropy of these three values. Hence in the sequel we will take into account the entropy of all integration values as a candidate for a topological index as well: with $L = \sum_{u \in V} i_u$ this will be

$$H^S = -\sum_{u \in V} \frac{i_u}{L} \ln \frac{i_u}{L}. \tag{4}$$

Similarly to H, H^S is defined for all connected graphs except for a complete graph.

The intelligibility is defined as the Pearson product-moment correlation coefficient between the sequences of vertex degrees and their integration values. If $\bar{d} = \frac{2m}{n}$ denotes the average vertex degree and $\frac{\sum_{u \in V} i_u}{n} = \frac{L}{n}$ denotes the average integration value, then the intelligibility is defined as

$$I = \frac{\sum_{u \in V} \left(d_u - \bar{d}\right) \left(i_u - \frac{L}{n}\right)}{\sqrt{\sum_{u \in V} \left(d_u - \bar{d}\right)^2} \sqrt{\sum_{u \in V} \left(i_u - \frac{L}{n}\right)^2}}. \tag{5}$$

Note that the Pearson correlation coefficient is not defined if one of the sequences is constant, which happens here if a graph is either regular, with all vertices having the same degree, or transmission-regular, with all vertices having the same transmission. Degree regularity is usually not a problem with justified plan graphs, as there is usually only one entrance to a building which corresponds to an exterior carrier vertex of degree one. Further, since the correlation coefficient is preserved by multiplying each element

of one sequence by a constant or by adding a constant to each element of one sequence, we see from (3) that intelligibility also represents correlation between vertex degrees and transmissions in case when graphs have constant number of vertices.

3 Correlations with physicochemical properties and other distance-based indices

In this section we first correlate the difference factor, the integration entropy and the intelligibility with physicochemical properties of octanes. Values for many such properties are provided in the benchmark dataset suggested by the International Academy of Mathematical Chemistry [41] and the website of the Korean databank of thermophysical properties [42]. Selected physicochemical properties of octanes and the values of their difference factors, integration entropies and intelligibilities are shown in Table 2.

Octane	CT	AP	MR	CP	LD	RI	H	H^S	I
n-octane	296.2	343.75	39.1922	2490	0.006154239	1.39505	1.041989747	2.027291846	-0.881917104
2-methyl-heptane	288	347.15	39.2316	2500	0.006145485	1.39257	1.032469299	2.027844836	-0.757670319
3-methyl-heptane	292	345.35	39.1001	2550	0.006180502	1.3961	1.022202174	2.020678795	-0.825028647
4-methyl-heptane	290	344.75	39.1174	2540	0.006171748	1.39553	1.010745401	2.016362398	-0.837241221
3-ethyl-hexane	292	341.85	38.9441	2610	0.006285553	1.39919	0.993820761	2.011409725	-0.944911183
2,2-dimethyl-hexane	279	351.15	39.2525	2530	0.006084205	1.39104	1.006956973	2.019011752	-0.708901337
2,3-dimethyl-hexane	293	343.75	38.9808	2630	0.006233028	1.3988	0.991970655	2.015377063	-0.85377141
2,4-dimethyl-hexane	282	346.55	39.13	2560	0.006127976	1.39291	1.026545306	2.021365884	-0.859215972
2,5-dimethyl-hexane	279	351.15	39.2596	2490	0.006066697	1.39004	1.053959917	2.029660354	-0.802180629
3,3-dimethyl-hexane	290.84	345.15	39.0087	2650	0.006215519	1.39782	0.968965165	2.006490345	-0.792405816
3,4-dimethyl-hexane	298	341.15	38.8453	2690	0.006294307	1.4018	1.011404265	2.01153134	-0.919866211
2-methyl-3-ethyl-pentane	295	340.35	38.8362	2700	0.006294307	1.40167	0.991872378	2.009296471	-0.961810764
3-methyl-3-ethyl-pentane	305	339.05	38.7171	2810	0.006364341	1.40549	0.962267837	1.999332839	-0.894427191
2,2,3-trimethyl-pentane	294	343.95	38.9249	2730	0.006268045	1.40066	0.988545691	2.008836763	-0.891498707
2,2,4-trimethyl-pentane	271.15	352.65	39.2617	2570	0.006057942	1.38898	1.03132043	2.021062584	-0.844926112
2,3,3-trimethyl-pentane	303	340.15	38.7617	2820	0.006355587	1.40522	0.960683693	2.00461833	-0.918397948
2,3,4-trimethyl-pentane	295	341.45	38.8681	2730	0.006294307	1.40198	1.01330833	2.015583185	-0.962395683
2,2,3,3-tetramethylbutane	270.8			2870			1.012330839	2.006981214	-1

Table 2. Selected physicochemical properties of octanes and their difference factor, integration entropy and intelligibility values. Legend: CT=critical temperature, AP=aniline point, MR=molar refraction, CP=critical pressure, LD=liquid density, RI=refractive index, H=difference factor, H^S=integration entropy, I=intelligibility. Data sources: IAMC [41] for CT and MR and KDB [42] for AP, CP, LD and RI.

Best single correlations are provided by the intelligibility, which yields $R^2 = 0.765$ when correlated with either critical temperature (with positive R) or aniline point (with negative R). However, better models are provided by multiple linear regression when the invariant pairs (H, I) and (H^S, I) are used to model octane properties. In particular, from data shown in Table 2 we get the following models for molar refraction (MR), critical pressure (CP), liquid density (LD) and refractive index (RI), for which R^2 values range

from 0.852 to 0.903:

$$\begin{aligned} MR &\approx 10.97888597 + 14.37814477H^S + 1.08395378I \qquad (R^2 = 0.903), \\ CP &\approx 21742.44664 - 9715.880529H^S - 544.661396I \qquad (R^2 = 0.853), \\ LD &\approx 0.007749596 - 0.002166637H - 0.000745039I \qquad (R^2 = 0.865), \\ RI &\approx 1.480372777 - 0.115162862H - 0.038128809I \qquad (R^2 = 0.852). \end{aligned}$$

Although these simple models are not always better than previously published models for MR, CP, LD and RI (see, for example, [43–52]), their high R^2 values nevertheless suggest that the intelligibility, combined with either the difference factor or the integration entropy, may be used in prediction of physicochemical properties.

Next, we check to what extent the difference factor, the integration entropy and the intelligibility are correlated to some of the existing distance-based topological indices. The following distance-based indices are considered for these correlations:

- the Wiener index $W = \frac{1}{2}\sum_{u\in V} TD_u$;
- the hyper-Wiener index [53, 54] $WW = \frac{1}{2}\sum_{u\in V}\sum_{v\in V\setminus\{u\}} d(u,v)[d(u,v)+1]$;
- the Harary index [55, 56] $Har = \frac{1}{2}\sum_{u\in V}\sum_{v\in V\setminus\{u\}} \frac{1}{d(u,v)}$;
- the Balaban index [18] $J = \frac{m}{m-n+2}\sum_{uv\in E} \frac{1}{\sqrt{TD_u TD_v}}$;
- the degree-distance [57] $DD = \frac{1}{2}\sum_{u\in V}\sum_{v\in V}(d_u + d_v)d(u,v)$;
- the Gutman index [58] $ZZ = \frac{1}{2}\sum_{u\in V}\sum_{v\in V} d_u d_v d(u,v)$;
- the Szeged index [59] $Sz = \sum_{uv\in E} n_u(uv)n_v(uv)$, where $n_u(uv)$ denotes the number of vertices in V that are closer to u than to v (and vice versa for $n_v(uv)$);
- the second geometric-arithmetic index [60] $GA_2 = \sum_{uv\in E} \frac{2\sqrt{n_u(uv)n_v(uv)}}{n_u(uv)+n_v(uv)}$;
- the edge Szeged index [61] $Sz_e = \sum_{uv\in E} m_u(uv)m_v(uv)$, where $m_u(uv)$ denotes the number of edges that are closer to u than to v (and vice versa for $m_v(uv)$). The distance between a vertex u and an edge $e = st$ is defined as $d(u,e) = \min\{d(u,s), d(u,t)\}$;
- the PI index [62] $PI = \sum_{uv\in E}(m_u(uv) + m_v(uv))$;

- the third geometric-arithmetic index [63] $GA_3 = \sum_{uv \in E} \frac{2\sqrt{m_u(uv) m_v(uv)}}{m_u(uv) + m_v(uv)}$;
- the spectral radius of distance matrix $\mathcal{D}$, whose (u, v)-entry represents the distance $d(u, v)$;
- the Kirchhoff index [64] $Kf = \frac{1}{2} \sum_{u \in V} \sum_{v \in V \setminus \{u\}} \Omega_{uv}$, where Ω is the resistance distance matrix whose entries are defined as $\Omega_{uv} = (\Gamma^{-1})_{uu} + (\Gamma^{-1})_{vv} - 2(\Gamma^{-1})_{uv}$, where $\Gamma = L + \frac{J_n}{n}$ for the Laplacian matrix L and the all-one matrix J_n.

Correlations of these indices have been calculated in four sets of graphs (recall that the attribute *chemical* denotes a graph in which all vertex degrees are at most four):

- T_{15}, trees on 15 vertices (7,741 trees),
- CT_{16}, chemical trees on 16 vertices (10,359 trees),
- G_8^-, connected graphs on 8 vertices (11,100 graphs), and
- CG_9^-, connected chemical graphs on 9 vertices (12,189 graphs).

In order for intelligibility to be defined, we have excluded 17 degree- and transmission-regular graphs from G_8^- and 18 degree- and transmission-regular graphs from CG_9^-. Correlations have been tested with Sage [65] and package MathChem [66], to which we added the code for calculating H, H^S and I.

Among trees and chemical trees results of these correlation tests show that the difference factor, the integration entropy and the intelligibility are not highly correlated to any of the above distance-based topological indices (with $R^2 < 0.58$ for each pair). The difference factor and the integration entropy are mutually correlated with $R^2 \approx 0.8$ in both sets of trees, while they are both uncorrelated to the intelligibility with $R^2 < 0.106$. On the other hand, the PI index, with its constant value for trees (and bipartite graphs in general), is not correlated to any other index, while the remaining distance-based indices are all mutually highly correlated with $R^2 > 0.855$ for each such pair.

Among connected graphs and chemical graphs correlation tests again show that the difference factor, the integration entropy and the intelligibility are not highly correlated to any of the above distance-based topological indices (with $R^2 < 0.50$ for each pair). The difference factor and the integration entropy are mutually correlated with $R^2 = 0.678$ for connected graphs on 8 vertices and $R^2 = 0.795$ for connected chemical graphs on 9

vertices, while they are both uncorrelated to the intelligibility with $R^2 < 0.051$. Nature of correlations among other distance-based indices in these sets of graphs changes considerably from the case of trees and is shown graphically in Fig. 2. We can see from this figure that, in addition to H, H^S and I, also the Balaban index, the Szeged index (and the degree-distance in the case of chemical graphs) are not highly correlated to any of the remaining indices, which all belong to the same connected component in these networks. Moreover, these connected components are edge-partitioned into cliques in almost the same way in both networks: the correlation network for connected graphs consists of cliques $\{W, Kf, WW, Har, Sp.rad.D\}$, $\{Sze, ZZ, PI, GA_2, GA_3, DD\}$ and $\{GA_2, GA_3, PI, Har, W\}$, while the correlation network for chemical graphs consists of $\{W, Kf, WW, Har, Sp.rad.D\}$, $\{Sze, ZZ, PI, GA_2, GA_3\}$ and $\{GA_2, GA_3, PI, Har, Kf\}$.

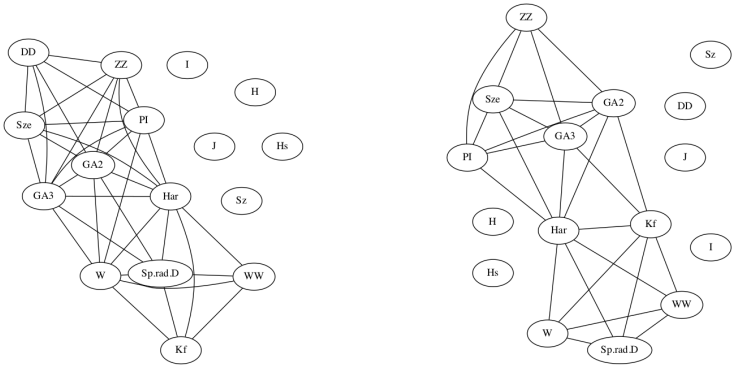

Figure 2. Correlation networks for the sets of connected graphs on 8 vertices (left) and connected chemical graphs on 9 vertices (right). Two topological indices are joined by an edge if $R^2 \geq 0.8$ for the corresponding correlation.

Altogether we can see from these correlations that the difference factor, the integration entropy and the intelligibility represent substantially different invariants from the main distance-based indices studied in the mathematical chemistry literature.

4 Degeneracy

Here we deal with degeneracy of the difference factor, the integration entropy and the intelligibility. Recall that a topological index is called degenerate if it possesses the same value for more than one graph. While one would ideally want to have a topological index

without degeneracy, so that a molecular graph could be uniquely described simply by the numerical value of the topological index, such an index, or even a set of indices whose values would jointly uniquely determine a graph, has not yet been found. Actually, most topological indices have quite a high degeneracy, as has been exemplified recently by Dehmer, Grabner and Furtula [67]. For evaluating degeneracy of an index TI we use the measure $S_{TI} = \frac{|\mathcal{G}_{\text{unique}}|}{|\mathcal{G}|}$, called the sensitivity by Konstantinova [68], where $|\mathcal{G}|$ denotes the number of graphs in the considered set, while $|\mathcal{G}_{\text{unique}}|$ denotes the number of graphs that have unique value of the topological index TI in that set. Values of sensitivity for H, H^S and I in the sets G_8, CG_9, T_{15} and CT_{16} are given in Table 3.

	G_8		CG_9		T_{15}		CT_{16}	
index	$\lvert\mathcal{G}_{\text{unique}}\rvert$	S	$\lvert\mathcal{G}_{\text{unique}}\rvert$	S	$\lvert\mathcal{G}_{\text{unique}}\rvert$	S	$\lvert\mathcal{G}_{\text{unique}}\rvert$	S
H	109	0.00981	161	0.01319	2120	0.27387	1681	0.16227
H^S	1710	0.15383	3621	0.29663	7477	0.96590	9494	0.91650
I	565	0.05090	3059	0.25096	7394	0.95517	8757	0.84535

Table 3. Sensitivity of H, H^S and I in the sets of connected graphs on 8 vertices (G_8), connected chemical graphs on 9 vertices (CG_9), trees on 15 vertices (T_{15}) and chemical trees on 16 vertices (CT_{16}). The complete graph K_8 has been excluded from G_8 in order for H and H^S to be defined. Similarly, I has been computed for the subsets G_8^- and CG_9^- which do not contain degree-regular and transmission-regular graphs.

It can be seen from this table that among these three indices the integration entropy H^S has the lowest degeneracy, followed by the intelligibility I, with the difference factor H having the highest degeneracy. When compared to Tables 2-4 in [67] it can be seen that degeneracies of H^S and I are still much higher than those of Laplacian energy [69] and Laplacian Estrada index [70] for all considered graph types, but that they are on par with most of the other topological indices. Further, degeneracy of each of H^S, I and H is lowest among trees with a slight increase for chemical trees, while it becomes much higher for connected chemical graphs and highest for general connected graphs. This could be understood intuitively by the fact that each pair of vertices is connected by a unique shortest path in a tree and usually via more shortest paths in graphs with many cycles (and edges). Consequently, sets of distances from any given vertex to all other vertices (sums of whose lengths make up the TD values) become subsets of smaller and smaller ranges as new edges are being added to the graph, which in principle makes it easier for

a given graph to find another graph with an appropriate distribution of TD values that yields the same value of H, H^S or I.

This is further exemplified by characteristic value distributions of H, H^S and I in these sets, which are shown in diagrams in Figs. 3–5. In these diagrams values of an index in a given set are depicted on the x-axis, while the number of graphs with that value of the index is depicted on the y-axis, with exclusion of a single point.

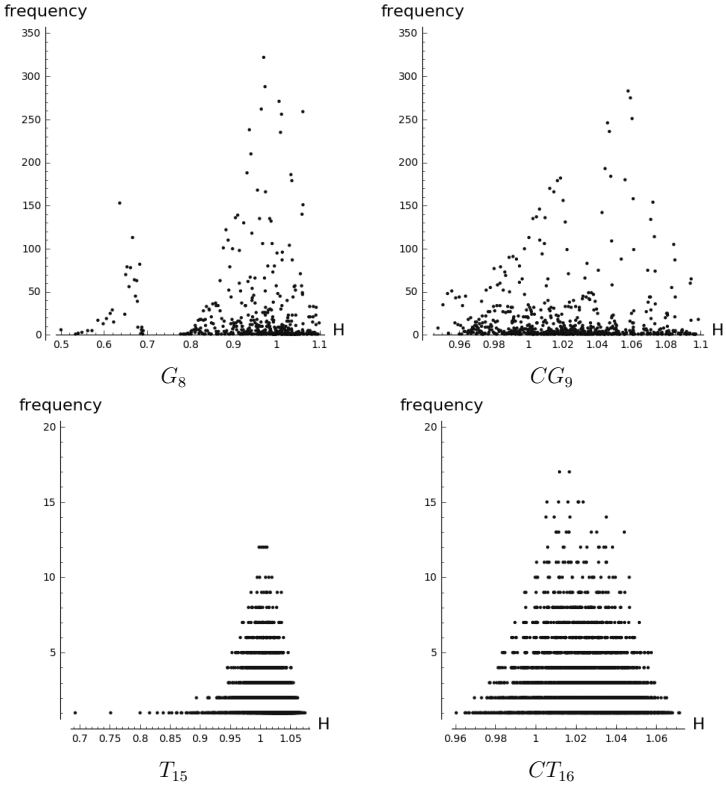

Figure 3. Characteristic value distributions of the difference factor H in the sets of connected graphs on 8 vertices (G_8), connected chemical graphs on 9 vertices (CG_9), trees on 15 vertices (T_{15}) and chemical trees on 16 vertices (CT_{16}).

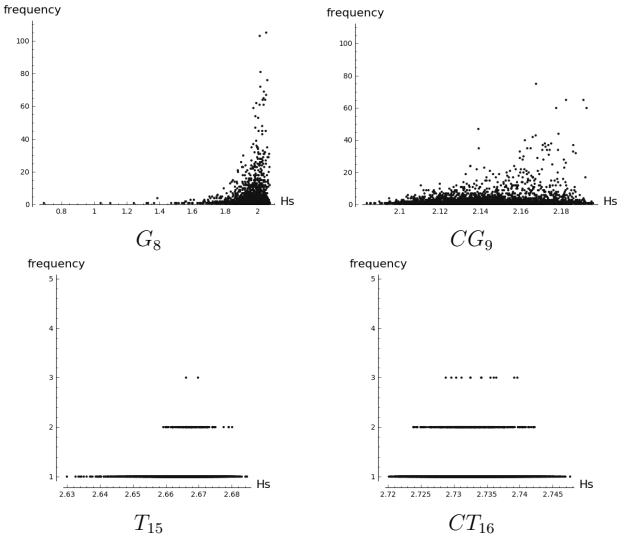

Figure 4. Characteristic value distributions of the integration entropy H^S in the sets of connected graphs on 8 vertices (G_8), connected chemical graphs on 9 vertices (CG_9), trees on 15 vertices (T_{15}) and chemical trees on 16 vertices (CT_{16}).

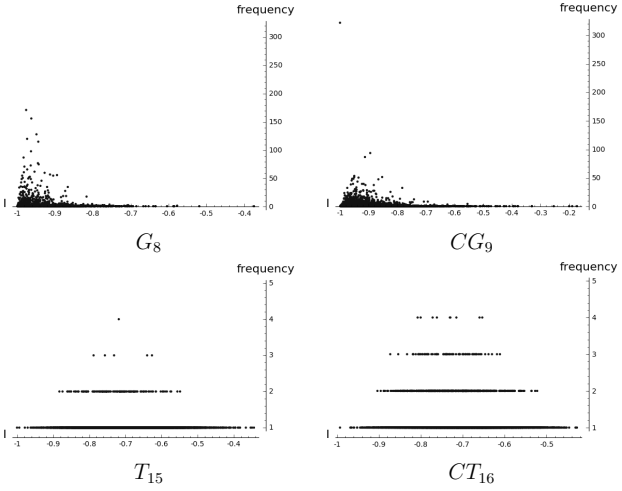

Figure 5. Characteristic value distributions of the intelligibility I in the sets of connected graphs on 8 vertices (G_8^-), connected chemical graphs on 9 vertices (CG_9^-), trees on 15 vertices (T_{15}) and chemical trees on 16 vertices (CT_{16}).

We can observe from diagrams in Figs. 3 and 4 that the restriction of the maximum degree to four in chemical graphs and trees is propagated further to also restrict the range of values that H and H^S have in these sets. These diagrams also more clearly demonstrate that highest frequencies are always encountered in G_8, followed by CG_9, and then followed by a sharp drop in frequencies among trees, emphasizing that majority of trees are determined by their values of H^S or I.

index	G_8	CG_9	T_{15}	CT_{16}
H	322	283	12	17
H^S	110	75	3	3
I	4259	322	4	4

Table 4. Maximum frequencies of values of H, H^S and I in the sets of connected graphs on 8 vertices (G_8), connected chemical graphs on 9 vertices (CG_9), trees on 15 vertices (T_{15}) and chemical trees on 16 vertices (CT_{16}).

The actual values of maximum frequencies are given in Table 4. The most interesting observation from this table is that the intelligibility has enormously high maximum frequency of 4259 in G_8, which means that 69.7% of all graphs in G_8 have the same value of I. This highly frequent value is -1: it is the minimum possible value for I and also the single point that could not fit in the corresponding diagram in Fig. 5 without significantly rescaling its y-axis. The reason for such a high frequency of this extremal value is explained in the following section.

5 Extremal graphs

Here we discuss observations about extremal graphs for the difference factor, the integration entropy and the intelligibility in considered sets of connected graphs, chemical connected graphs, trees and chemical trees. We start the section by explaining high frequency of the minimum value of -1 for the intelligibility I among connected graphs. After an initial review of such graphs in G_8 it became clear that they usually contain a vertex of very high degree, often of the maximum possible value of $n-1$. Certainly, the existence of a high degree vertex means that the diameter of such graph is rather small. Indeed, when we classified the 4,259 graphs with intelligibility -1 in G_8 according to their diameter, it turned out that among them there are:

- 3,594 graphs with diameter two,
- 98 graphs with diameter three,
- 6 graphs with diameter four, and
- 1 graph with diameter five.

Examples of such graphs are shown in Fig. 6.

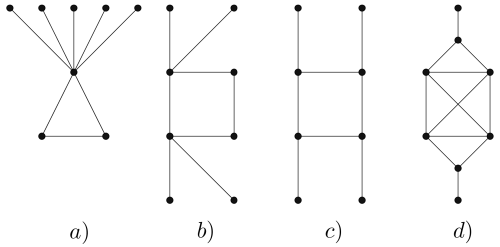

Figure 6. Examples of graphs with intelligibility -1 in G_8 with: a) diameter two, b) diameter three, c) diameter four and d) diameter five.

As a matter of fact, it is easy to see that every graph with diameter two has intelligibility -1. Namely, in a graph with diameter two, for each vertex u there are d_u neighbors at distance one and $n-1-d_u$ vertices at distance two from u. Consequently, $TD_u = d_u \cdot 1 + (n-1-d_u) \cdot 2 = 2(n-1) - d_u$. Hence $d_u + TD_u$ is equal to a constant $2(n-1)$ for each vertex u, which implies perfect linear correlation between degrees and transmissions, i.e., the intelligibility of -1.

On the other hand, in order to maintain a linear relation between degrees and transmissions graphs of larger diameters have to have higher and higher degrees of symmetry as their diameter grows. Consequently, the number of graphs with intelligibility -1 should drop very quickly with the increase of the diameter, as is evident from the numbers of such graphs in G_8. Nevertheless, we need to also take into account here the existing results on the number of graphs with given diameter, which claim that almost all graphs have diameter two [71] and that for each fixed $k \geq 2$ the limit of the ratio of the number of n-vertex graphs with diameter k and the number of n-vertex graphs with diameter $k+1$ is equal to ∞ when n tends to ∞ [72]. Hence, intelligibility will inevitably have high

degeneracy among connected graphs, regardless of the number of vertices, predominantly because of the high frequency of the minimum value -1.

Concerning maximum intelligibility, note first that it is intuitively expected that intelligibility is always negative, since a vertex with larger degree is expected to have smaller transmission and vice versa. This is confirmed by the maximum values of intelligibility in the sets G_8^-, CG_9^-, T_{15} and CT_{16}, which range between -0.4277 and -0.1723. While the stars, as graphs with diameter two, always appear as graphs with the minimum intelligibility, the paths do not even come close to the maximum intelligibility: intelligibility of paths on 8 to 16 vertices ranges from -0.8819 to -0.7001 only. Instead, graphs with the maximum intelligibility among connected graphs and connected chemical graphs appear to consist of two cliques connected by a path, while graphs with the maximum intelligibility among trees and chemical trees appear as a concatenation of a star and a path. For illustration, the two graphs with the maximum value of intelligibility from each of the considered graph sets are shown in Fig. 7.

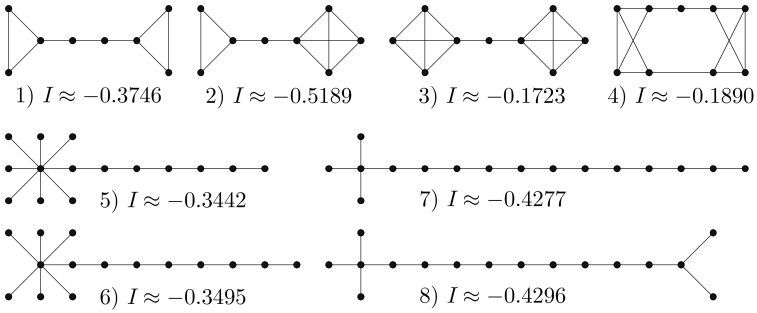

Figure 7. Graphs with the maximum values of intelligibility in the sets G_8^- (1-2), CG_9^- (3-4), T_{15} (5-6) and CT_{16} (7-8).

Since both H and H^S represent entropies of sequences (a three-element sequence for H and an n-element sequence for H^S), their maximum values should be achieved when all the sequence elements are equal. In cases of connected graphs and connected chemical graphs, this happens when the graphs are transmission-regular. In such case the value of H becomes $\ln 3$, while the value of H^S becomes $\ln n$. There are quite a few transmission-regular graphs. They mostly arise from vertex-transitive graphs, which for any two vertices u and v have an automorphism that maps u to v, which implies

that any two locally defined invariants (including degrees and transmissions) are equal for both u and v. Perhaps a simplest example of a vertex-transitive graph is the cycle C_n. However, there are also examples of chemical graphs that are transmission-regular, but not degree-regular (and hence not vertex-transitive), one of which is shown in Fig. 8.

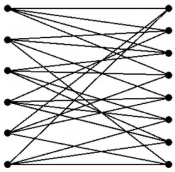

Figure 8. An example of a chemical graph that is transmission-regular, but not degree-regular [73].

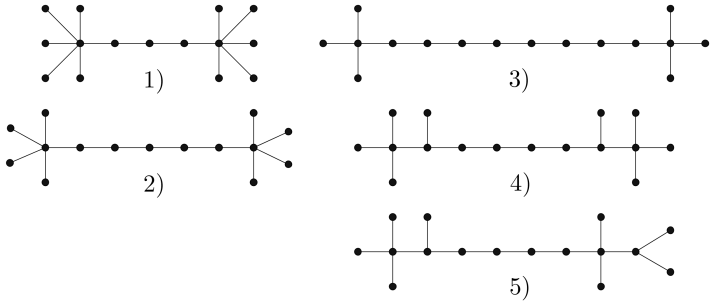

Figure 9. Trees with the two maximum values of H and H^S in the sets T_{15} and CT_{16}: 1) Tree with the maximum value of H and the second maximum value of H^S in T_{15}; 2) Tree with the second maximum value of H and the maximum value of H^S in T_{15}; 3) Chemical tree with the maximum value of H in CT_{16}; 4) Chemical tree with the second maximum value of H and the maximum value of H^S in CT_{16}; 5) Chemical tree with the second maximum value of H^S in CT_{16}.

On the other hand, transmission-regularity cannot be achieved among trees, as leaves always have larger transmission than internal vertices. Trees with the maximum values of H and H^S in T_{15} consist of two stars joined by a path, while such trees in CT_{16} appear to have the same aim, but are restricted in it by their maximum degree. Moreover, some of the extremal trees are shared between H and H^S as can be seen from Fig. 9.

Similarities between entropic definitions of H and H^S are not transferred to graphs with minimum values of these indices. Among connected graphs in G_8, the minimum

value of H is achieved for a complete graph K_8 from which edges of a star have been deleted, while the minimum value of H^S is achieved only for K_8 from which a single edge is deleted. Among chemical graphs in CG_9, the minimum value of H is achieved by eight graphs of diameter four that consist of a seven-vertex subgraph to which a path of length two is appended, while the minimum value of H^S is achieved by an entirely different, single graph. Even among trees in T_{15}, the extremal trees are quite different: the minimum value of H is achieved by a star, while the minimum value of H^S is achieved by a broom—a star and a path in which a pair of leaves have been identified. Moreover, extremal trees in CT_{16} look rather unorthodox, with the one for H apparently trying to emulate a broom under the restriction on the maximum degree at most four, while the one for H^S offers no particular hint for its structure. A selection of these extremal graphs is shown in Fig. 10.

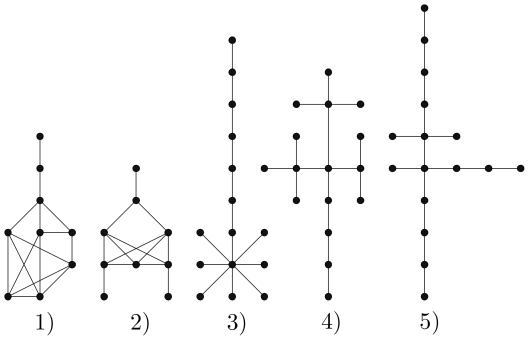

Figure 10. Graphs with minimum values of H or H^S in considered graph sets: 1) Graph with the minimum value of H in CG_9; 2) Graph with the minimum value of H^S in CG_9; 3) Tree with the minimum value of H^S in T_{15}; 4) Tree with the minimum value of H in CT_{16}; 5) Tree with the minimum value of H^S in CT_{16}.

6 Conclusions

The difference factor and the intelligibility have been used in urban and architectural research for more than 30 years. We have performed here a number of computational studies to find out to what extent these invariants, together with the integration entropy, can also serve as topological indices in mathematical chemistry. We have seen that these invariants are correlated to some physicochemical properties of octanes, although they

have to be used in pairs, or in conjuction with other indices, in order to yield predictions of higher precision. The difference factor and the integration entropy employ entropic measures to quantify the diversity of integration values of vertices, while the intelligibility employs the Pearson product-moment correlation coefficient to measure the extent to which integration values may be predicted from vertex degrees. As such, it comes as no surprise that these invariants are uncorrelated to usual distance-based indices studied in mathematical chemistry.

Degeneracy of all three invariants is rather high among general connected graphs. However, degeneracy of intelligibility and integration entropy decreases significantly for chemical graphs and becomes acceptably low for trees. It can thus be expected that degeneracy of intelligibility and integration entropy will remain on acceptably low levels in cases when the number of edges is similar to the number of vertices, i.e., when graphs have low cyclomatic number.

Computational studies of extremal graphs reveal that, from one side, these invariants have trivial extremal graphs among connected graphs: the difference factor and the integration entropy achieve their maximum values for transmission-regular graphs, while the intelligibility achieves the minimum value of -1 for each diameter two graph, which make up majority of connected graphs. For the opposite extrema it is possible to conjecture the structure of extremal graphs in a number of cases: the maximum intelligibility is achieved either by brooms or cliques joined by a path, while the minimum difference factor and integration entropy are achieved, depending on the graph type, by a star, a broom or a complete graph from which edges of a star have been deleted. However, some rather unorthodox graphs can appear as extremal graphs as well, especially among chemical trees.

To conclude, based on the results of computational studies performed here it is our opinion that the difference factor and the intelligibility, together with the integration entropy, satisfy a number of desirable requirements for topological indices and that they may offer to researchers in mathematical chemistry interesting new perspectives and avenues for further research.

Note: Sage code used for calculations in this article, together with an updated copy of MathChem and necessary graph sets, may be obtained from the corresponding author on request.

References

[1] H. Wiener, Structural determination of paraffin boiling points, *J. Am. Chem. Soc.* **69** (1947) 17–20.

[2] H. Wiener, Correlation of heat of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons, *J. Am. Chem. Soc.* **69** (1947) 2636–2638.

[3] H. Wiener, Relationship of physical properties of isomeric alkanes to molecular structure surface tension, specific dispersion and critical solution temperature in aniline, *J. Phys. Colloid Chem.* **52** (1948) 1082–1089.

[4] H. Wiener, Vapour pressure–temperature relations among the branched paraffin hydrocarbons, *J. Phys. Chem.* **52** (1948) 425–430.

[5] R. C. Entringer, D. E. Jackson, D. A. Snyder, Distance in graphs, *Czech. Math. J.* **26** (1976) 283–296.

[6] J. Plesník, On the sum of all distances in a graph or digraph, *J. Graph Theory* **8** (1984) 1–21.

[7] L. Šoltés, Transmission in graphs: A bound and vertex removing, *Math. Slovaca* **41** (1991) 11–16.

[8] F. Chung, The average distance and the independence number, *J. Graph Theory* **12** (1988) 229–235.

[9] M. V. Diudea, I. Gutman, Wiener–type topological indices, *Croat. Chem. Acta* **71** (1998) 21–51.

[10] A. A. Dobrynin, R. Entringer, I. Gutman, Wiener index of trees: theory and applications, *Acta Appl. Math.* **66** (2001) 211–249.

[11] A. A. Dobrynin, I. Gutman, S. Klavžar, P. Žigert, Wiener index of hexagonal systems, *Acta Appl. Math.* **72** (2002) 247–294.

[12] H. Hosoya, Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons, *B. Chem. Soc. Jpn* **44** (1971) 2332–2339.

[13] I. Gutman, N. Trinajstić, Graph theory and molecular orbitals. Total π-electron energy of alternant hydrocarbons, *Chem. Phys. Lett.* **17** (1972) 535–538.

[14] M. Randić, On characterization of molecular branching, *J. Am. Chem. Soc.* **97** (1975) 6609–6615.

[15] L. B. Kier, L. H. Hall, W. J. Murray, M. Randić, Molecular connectivity I: Relationship to local anesthesia, *J. Pharm. Sci.* **64** (1975) 1971–1974.

[16] L. B. Kier, W. J. Murray, M. Randić, L. H. Hall, Molecular connectivity V: Connectivity series concept applied to density, *J. Pharm. Sci.* **65** (1976) 1226–1230.

[17] L. B. Kier, L. H. Hall, Molecular connectivity VII: Specific treatment of heteroatoms, *J. Pharm. Sci.* **65** (1976) 1806–1809.

[18] A. T. Balaban, Highly discriminating distance–based topological index, *Chem. Phys. Lett.* **89** (1982) 399–404.

[19] S. C. Basak, Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach, *Med. Sci. Res.* **15** (1987) 605–609.

[20] Kode Chemoinformatics, Dragon 7, Retrieved from `https://chm.kode-solutions.net/products_dragon.php`

[21] M. Randić, Generalized molecular descriptors, *J. Math. Chem.* **7** (1991) 155–168.

[22] R. C. Read, A new system for the designation of chemical compounds. 1. Theoretical preliminaries and the coding of acyclic compounds, *J. Chem. Inf. Comput. Sci.* **23** (1983) 135–149.

[23] A. T. Balaban, Topological indices and their uses: A new approach for the coding of alkanes, *J. Mol. Struc. (Theochem)* **165** (1988) 243–253.

[24] B. Furtula, I. Gutman, M. Dehmer, On structure–sensitivity of degree–based topological indices, *Appl. Math. Comput.* **219** (2013) 8973–8978.

[25] F. Harary, *Graph Theory*, Addison-Wesley, Reading, 1969.

[26] C. Alexander, A city is not a tree, Part 1, *Archit. Forum* **122** (1965), No. 1, 58–62.

[27] C. Alexander, A city is not a tree, Part 2, *Archit. Forum* **122** (1965), No. 2, 58–62.

[28] L. March, P. Steadman, *The Geometry of Environment*, Royal Inst. British Architects, London, 1971.

[29] P. J. Steadman, Graph–theoretic representation of architectural arrangement, *Archit. Res. Teaching* **2** (1973) 161–172.

[30] L. March, *The Architecture of Form*, Cambridge Univ. Press, Cambridge, 1976.

[31] M. J. T. Krüger, *An approach to built form connectivity at an urban scale*, Ph.D. thesis, Univ. Cambridge, Cambridge, 1977.

[32] B. Hillier, J. Hanson, *Social Logic of Space*, Cambridge Univ. Press, Cambridge, 1983.

[33] S. Bafna, Space syntax: a brief introduction to its logic and analytical techniques, *Environ. Behav.* **35** (2003) 17–29.

[34] B. Hillier, J. Hanson, H. Graham, Ideas are in things: an application of the space syntax method to discovering house genotypes, *Environ. Plann. B* **14** (1987) 363–385.

[35] B. Hillier, R. Burdett, J. Peponis, A. Penn, Creating life: or, does architecture determine anything? *Arch. Behav.* **3** (1987) 233–250.

[36] B. Hillier, K. Tzortzi, Space syntax: the language of museum space, in: S. Macdonald (Ed.), *A Companion to Museum Studies*, Blackwell, London, 2006, pp. 282–301.

[37] M. J. Ostwald, M. Dawes, Axial line analysis revisited: reconsidering its value for architecture, *Int. J. Constr. Environ.* **1** (2011) 219–242.

[38] M. J. Ostwald, The mathematics of spatial configuration: revisiting, revising and critiquing justified plan graph theory, *Nexus Netw. J.* **13** (2011) 445–470.

[39] Wikimedia Commons, File: Wright-Heurtley House Lower Floor.png, `https://commons.wikimedia.org/wiki/File:Wright-Heurtley_House_Lower_Floor.png`, accessed November 2017.

[40] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* **27** (1948) 379–423.

[41] Milano Chemometrics & QSAR Research Group, Datasets for molecular descriptor comparisons, `http://www.moleculardescriptors.eu/dataset/dataset.htm`, accessed September 2017.

[42] Chemical Engineering and Materials Research Center, KDB, `https://www.cheric.org/research/kdb/`, accessed September 2017.

[43] J. Bicerano, *Prediction of Polymer Properties*, Marcel Dekker, New York, 2002.

[44] E. J. Kupchik, Structure–molar refraction relationships of alkylsilanes using molecular connectivity, *Mol. Inform.* **4** (1985) 123–127.

[45] F. Gharagheizi, P. Ilani–Kashkouli, A. Kamari, A. H. Mohammadi, D. Ramjugernath, Group contribution model for the prediction of refractive indices of organic compounds, *J. Chem. Eng. Data* **59** (2014) 1930–1943.

[46] M. A. F. Afzal, C. Cheng, J. Hachmann, Combining first–principles and data modeling for the accurate prediction of the refractive index of organic polymers, *ChemRxiv* 5446564, `https://doi.org/10.26434/chemrxiv.5446564.v1`, accessed October 2017.

[47] K. G. Joback, R. C. Reid, Estimation of pure–component properties from group–contributions, *Chem. Eng. Commun.* **57** (1987) 233–243.

[48] A. K. A. A. Razzak, E. T. Hashim, H. N. Al-Shurchi, New statistical concepts for prediction viscosity, critical temperature, and critical pressure of homologous hydrocarbon series (C_nH_{2n+2}), *Petrol. Sci. Technol.* **24** (2006) 1431–1435.

[49] H. G. Rackett, Equation of state for saturated liquids. *J. Chem. Eng. Data* **15** (1970) 514–517.

[50] T. Yamada, R. D. Gunn, Saturated liquid molar volumes: The Racket equation. *J. Chem. Eng. Data* **18** (1973) 234–236.

[51] R. W. Hankinson, G.H. Thomson, A new correlation for saturated densities of liquids and their mixtures, *AIChE J.* **25** (1979) 653–663.

[52] G. H. Thomson, K. R. Brobst, R. W. Hankinson, An improved correlation for densities of compressed liquids and liquid mixtures, *AIChE J.* **28** (1982) 671–676.

[53] M. Randić, Novel molecular descriptor for structure-property studies, *Chem. Phys. Lett.* **211** (1993) 478–483.

[54] D. J. Klein, I. Lukovits, I. Gutman, On the definition of the hyper-Wiener index for cycle-containing structures, *J. Chem. Inf. Comput. Sci.* **35** (1995) 50–52.

[55] O. Ivanciuc, T. S. Balaban, A. T. Balaban, Reciprocal distance matrix, related local vertex invariants and topological indices, *J. Math. Chem.* **12** (1993) 309–318.

[56] D. Plavšić, S. Nikolić, N. Trinajstić, Z. Mihalić, On the Harary index for the characterization of chemical graphs, *J. Math. Chem.* **12** (1993) 235–250.

[57] A. A. Dobrynin, A. A. Kochetova, Degree distance of a graph: a degree analogue of the Wiener index, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1082–1086.

[58] I. Gutman, Selected properties of the Schultz molecular topological index, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1087–1089.

[59] I. Gutman, A formula for the Wiener number of trees and its extension to graphs containing cycles, *Graph Theory Notes New York* **27** (1994) 9–15.

[60] G. Fath–Tabar, B. Furtula, I. Gutman, A new geometric–arithmetic index, *J. Math. Chem.* **47** (2010) 477–486.

[61] I. Gutman, A. R. Ashrafi, The edge version of the Szeged index, *Croat. Chem. Acta* **81** (2008) 263–266.

[62] P. V. Khadikar, On a novel structure descriptor PI, *Nat. Acad. Sci. Lett.* **23** (2000) 113–118.

[63] B. Zhou, I. Gutman, B. Furtula, Z. Du, On two types of geometric–arithmetic index, *Chem. Phys. Lett.* **482** (2009) 153–155.

[64] D. J. Klein, M. Randić, Resistance distance, *J. Math. Chem.* **12** (1993) 81–95.

[65] W. Stein, Sage: Open Source Mathematics Software, `http://www.sagemath.org/`, accessed Oct 2017.

[66] A. Vasilyev, D. Stevanović, MathChem: A python package for calculating topological indices, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 657–680.

[67] M. Dehmer, M. Grabner, B. Furtula, Structural discrimination of networks by using distance, degree and eigenvalue-based measures, *PLoS ONE* **7** (2012) e38564.

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

[69] I. Gutman, B. Zhou, Laplacian energy of a graph, *Lin. Algebra Appl.* **414** (2006) 29–37.

[70] G. H. Fath–Tabar, A. R. Ashrafi, I. Gutman, Note on Estrada and L-Estrada indices of graphs, *Bull. Cl. Sci. Math. Nat. Sci. Math.* **34** (2009) 1–16.

[71] B. Bollobas, *Graph Theory. An Introductory Course*, Springer, New York, 1979.

[72] I. Tomescu, An asymptotic formula for the number of graphs having small diameter, *Discr. Math.* **156** (1996) 219–228.

[73] M. Aouchiche, P. Hansen, On a conjecture about the Szeged index, *Eur. J. Comb.* **31** (2010) 1662–1666.