On Entropy-Based Molecular Descriptors: Statistical Analysis of Real and Synthetic Chemical Structures

Matthias Dehmer,*[,†,‡] Kurt Varmuza,*[,¶] Stephan Borgert,*[,§] and Frank Emmert-Streib*[,||]

Institute for Bioinformatics and Translational Research, UMIT, Eduard Wallnoefer Zentrum 1, A-6060, Hall in Tyrol, Austria, Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Wiedner Hauptstrasse 8−10, A-1040 Vienna, Austria, Laboratory for Chemometrics, Vienna University of Technology, Institute of Chemical Engineering, Getreidemarkt 9/166, A-1060 Vienna, Austria, Darmstadt University of Technology, Department of Computer Science, Hochschulstrasse 10, D-64289 Darmstadt, Germany, and Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, 97 Lisburn Road, Belfast, BT9 7BL, U.K.

Received February 19, 2009

This paper presents an analysis of entropy-based molecular descriptors. Specifically, we use real chemical structures, as well as synthetic isomeric structures, and investigate properties of and among descriptors with respect to the used data set by a statistical analysis. Our numerical results provide evidence that synthetic chemical structures are notably different to real chemical structures and, hence, should not be used to investigate molecular descriptors. Instead, an analysis based on real chemical structures is favorable. Further, we find strong hints that molecular descriptors can be partitioned into distinct classes capturing complementary information.

INTRODUCTION

In mathematical chemistry, investigating topological indices as molecular descriptors that operate on graph-based representations of molecules is a major research field of ongoing interest.[1−6] For example, an important task is to predict biological activities, as well as toxicological or physicochemical properties of chemical compounds from chemical structure data. This area is commonly named quantitative structure−activity/property relationship (QSAR/QSPR) and is highly relevant for instance in drug design.[4,7] In this paper, we are particularly interested in describing the topological complexity of chemical structures[8−10] which is relevant for empirical QSAR/QSPR models. A variety of graph measures have been used so far to characterize molecular complexity that usually takes connectivity and structural features of chemical structures into account.[9−12] Beside various graph complexity measures,[10,11,13,14] information-theoretic concepts for determining the structural information content of a graph have been intensely applied within QSAR/QSPR.[8,10,15,16] As pointed out,[8,17] the starting point of this interdisciplinary research field was initiated[18−22] to investigate biological and chemical systems. Then, Trucco[23] and Rashevsky[22] developed the first method to define a structural information content of a graph. The crucial step was to find distinguishable vertices of a graph and, then, to apply Shannon's entropy[24] for determining the entropy of this graph interpreted as its information content. As an extension, Rashevsky's method[22] was strengthened by Mowshowitz.[25−28] For this, he used algebraic principles (e.g., determining automorphism groups or chromatic decompositions) for inducing vertex partitions and defined the corresponding graph entropy measures. Later, Bonchev et al.[8,15,29] developed a variety of different information indices based on finding a weighted probability scheme. Apart from only characterizing graphs by using entropy measures, entropic measures such as *relative entropy* have been recently applied to determine the local information spread within gene networks.[30] Further entropy measures for networks can be found in.[31]

In this paper, we present a novel family of information indices that uses the construction of a recently proposed entropy measure for graphs,[32] which is a successor of that in ref 33. The definition of this measure is based on using a special information functional that captures structural information of a graph. By applying this concept, we define a family of information indices for determining the structural information content of molecules represented by graphs. So far, this approach has been applied to develop methods to characterize complex networks structurally.[17,32] Moreover we already used a special information functional for inferring so-called information inequalities starting from hierarchical molecular graphs.[34]

One of the main contributions of this paper is to investigate the mentioned family by studying their characteristic value distributions. More precisely, we will consider two types of chemical structures: One set of chemical structures originates from a widely used mass

* To whom correspondence should be addressed. E-mail: matthias.dehmer@umit.at (M.D.); kvarmuza@email.tuwien.ac.at (K.V.); borgert@tk.informatik.tu-darmstadt.de (S.B.); v@bio-complexity (F.E.-S.).
† Institute for Bioinformatics and Translational Research, UMIT.
‡ Institute of Discrete Mathematics and Geometry, Vienna University of Technology.
¶ Laboratory for Chemometrics, Vienna University of Technology.
§ Department of Computer Science, Darmstadt University of Technology.
|| Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast.

10.1021/ci900060x CCC: $40.75 © 2009 American Chemical Society
Published on Web 06/16/2009

spectral database;[35] that means the chemical structures are from existing compounds from which mass spectra have been experimentally measured. Other sets with chemical structures have been created by isomer generation for selected molecular formulas[36] and defined structural restrictions. Then, we will calculate the entropies by varying parameters of the involved information functional and compare the resulting value distributions of the information indices for different sets of chemical structures. Exploring such value distributions is important to obtain an insight into the structural factors which contribute to the complexity of a molecule and, hence, to understand the meaning of our information indices. Moreover, we investigate for every considered graph class the extremal behavior (minimum and maximum entropy) with respect to the chosen family. Finally, we study the relatedness of our entropy-based descriptor to some other selected molecular descriptors.

METHODS

Mathematical Preliminaries. Before starting with our analysis, we first express some mathematical preliminaries.[32,37−39] We start with defining an undirected, finite, and connected graph by $G = (V,E), |V| < \infty, E \subseteq \binom{V}{2}$. G is called connected if for arbitrary vertices v_i and v_j there exists an undirected path from v_i to v_j. Otherwise, we call G unconnected. $\mathscr{G}_{UC}$ denotes the set of finite, undirected, and connected graphs. We call the quantity $\delta(v)$ the degree of a vertex $v \in V$. $\delta(v)$ equals the number of edges $e \in E$, which are incident with v. For measuring distances between vertices in a graph, we denote $d(u,v)$ as the distance between $u \in V$ and $v \in V$ expressed as the minimum length of a path between u,v. $d(u,v)$ is an integer metric. To introduce some further metrical properties of graphs, we call the quantity $\sigma(v) = \max_{u \in V} d(u,v)$ the eccentricity of $v \in V$. $\rho(G) = \max_{v \in V} \sigma(v)$ is called the diameter of G. Finally, the j-sphere of a vertex v_i regarding $G \in \mathscr{G}_{UC}$ is defined as the set,

$$S_j(v_i, G) := \{v \in V \mid d(v_i, v) = j,\ j \geq 1\} \qquad (1)$$

$S_j(v_i,G)$ is the set of vertices whose distances are equal to j starting from $v_i \in V$.

Short Review of Existing Information Indices. In this section, we give a short review of the most known information indices, which have been widely used for quantifying structural information of chemical structures. In QSAR and QSPR, partition-based information measures have been used for characterizing molecular graphs structurally by using a graph invariant X and a certain equivalence criterion α. The main step to construct these measures is as follows: The application of the equivalence criterion produces a partitioning with respect to X into k subsets whose cardinalities are denoted by $|X_i|$. Starting from such a partitioning, the structural information content of a chemical graph G can be defined by[8]

$$I(G, \alpha) = |X| \log(|X|) - \sum_{i=1}^{k} |X_i| \log(|X_i|) \qquad (2)$$

$$\bar{I}(G, \alpha) = -\sum_{i=1}^{k} p_i \log(p_i) = -\sum_{i=1}^{k} \frac{|X_i|}{|X|} \log\left(\frac{|X_i|}{|X|}\right) \qquad (3)$$

Equations 2 and 3 are graph entropy measures for quantifying structural information of G. More precisely, $I(G,\alpha)$ and $\bar{I}(G,\alpha)$ represent the total and the mean information content of G, respectively.[8] As a technical note, we will always take the logarithms to the base 2 because we express the structural information contents in bits. For example, concrete graph invariants were studied by Rashevsky[22] and Trucco[23] and led to well-known specializations of eqs 2 and 3. In fact, Mowshowitz[25−28] was the first who expressed a mathematically rigorous approach for determining the structural information content of a graph by further developing the method of Rashevsky.[22] Further, he explored graph operations like complement, sum, join, etc., and investigated the change of the corresponding information index. This examination was of considerable interest for the information-based modeling of chemical reactions.[8] More classical and advanced information indices which are based on similar construction principles have been investigated by refs 9, 29, and 40−46.

The well-known *magnitude-based* information indices for graphs have been introduced by Bonchev et al.[29] We notice that they can be considered as a generalization of the measures of Rashevsky and Mowshowitz. Concrete magnitude-based information indices were obtained by considering the distribution of distances in a graph, the decomposition of graph distances in the associated distance matrix and degree distributions,[8] etc.

Konstantinova et al. introduced further information indices which are based on graph distances.[47,48] The main difference to the just stated ones is that these measures were developed to determine the information distance of a vertex $v_i \in V$ depending on the remaining vertices of a graph G. This implies that these measures are suitable to measure the information content of a graph element (e.g., a vertex) and not to characterize G globally. Starting from the same principle, Balaban et al.[16] also derived distance-based information measures to determine the information content of chemical structures.

To finalize our short review of information indices to characterize chemical structures, we mention the entropy-based descriptors recently introduced by Gregori-Puigjané et al.[49] These information measures are based on using atom-centered feature pairs, which have been also used as molecular descriptors.[50] Here, the shortest path length between atom-centered features is used to create a feature−pair distribution. From this, a probability distribution and, hence, entropy measures for chemical structures to quantify the variability in a feature-pair distribution are obtained.

Information Indices Based on the Full Topological Neighborhood of All Atoms. In this section, we define a novel family of information indices which are based on the full topological neighborhood of all involved atoms of the molecule. For this, we use a method for deriving graph entropy measures for arbitrary undirected and connected networks that was recently proposed in ref 32. The basic method was used to infer information inequalities for hierarchical graphs.[34] As expressed in ref 32, the main idea for defining novel information measures for graphs is to use

an information functional approach. An information functional represents a positive and monotonous mapping that quantifies structural information of a graph G. A key feature of this approach is that the resulting information indices do not depend on induced vertex partitions because we assign a probability value to every vertex (atom) in a graph. In the following, we use the j-sphere cardinalities of a graph to define a parametric information functional. Hence, we obtain a family of information indices. To start, we briefly repeat some definitions that were presented in ref 32.

Definition 1. Let $G = (V,E) \in \mathscr{G}_{UC}$. The vertex probabilities for each $v_i \in V$ are defined by the quantities

$$p(v_i) := \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \quad (4)$$

where f represents an arbitrary information functional.

Definition 2. Let $G = (V,E) \in \mathscr{G}_{UC}$. Then, the entropy of G is defined by

$$I_f(G) := -\sum_{i=1}^{|V|} p(v_i)\log(p(v_i)) \quad (5)$$

Now, we define a novel information functional by using metrical properties of graphs.

Definition 3. Let $G \in \mathscr{G}_{UC}$ with arbitrary vertex labels. For a vertex $v_i \in V$, the information functional $f_\star^V$ is defined as

$$f_\star^V(v_i) := c_1|S_1(v_i,G)| + c_2|S_2(v_i,G)| + \ldots + c_{\rho(G)}|S_{\rho(G)}(v_i,G)|, \quad c_k > 0, 1 \le k \le \rho(G) \quad (6)$$

We notice that this information functional is similar to the information functional

$$f^V(v_i) := \alpha^{c_1|S_1(v_i,G)|+c_2|S_2(v_i,G)|+\ldots+c_{\rho(G)}|S_{\rho(G)}(v_i,G)|}, \quad c_k > 0, \quad 1 \le k \le \rho(G), \quad \alpha > 0 \quad (7)$$

that has been presented in ref 32. The difference is that $f_\star^V$ does not depend on the free parameter $\alpha > 0$ for evaluating the local information spread in a graph. Also, we already mentioned[32] that the parameters c_k have to be chosen such that they are all not equal, e.g., $c_1 > c_2 > \ldots > c_{\rho(G)}$. Finally, we define the novel family of information measures by using the concept of Shannon's mean information[24] as follows.

Definition 4. Let $G = (V,E) \in \mathscr{G}_{UC}$. Then, we define the information indices

$$I_{f_\star^V}(G) := -\sum_{i=1}^{|V|} p_\star^V(v_i)\log(p_\star^V(v_i)) \quad (8)$$

and

$$I_{f_\star^V}^\lambda(G) := \lambda\left(\log(|V|) + \sum_{i=1}^{|V|} p_\star^V(v_i)\log(p_\star^V(v_i))\right) \quad (9)$$

where $\lambda > 0$ is a scaling constant.

Clearly, $I_{f_\star^V}(G)$ represents the mean structural information content of G by incorporating the probability values $p_\star^V(v_i) := (f_\star^V(v_i))/(\sum_{j=1}^{|V|} f_\star^V(v_j))$. $I_{f_\star^V}^\lambda(G)$ is achieved by subtracting $I_{f_\star^V}(G)$ from the maximum entropy value $\log(|V|)$ and multiplying the result by a scaling constant λ. In this paper, we use the special family of information indices $I_{f_\star^V}^\lambda(G)$ for characterizing molecular structures by calculating their structural information content. To tackle this problem, we consider each atom of a given molecular structure as a subsystem. Our chosen information functional $f_\star^V$ possesses the property that it characterizes the complete neighborhood for each atom by determining the number of atoms in all possible j-spheres around an atom. The already mentioned weighting scheme represented by the vector $(c_1,\ldots, c_{\rho(G)})$ combines the number of atoms in the different spheres resulting in a characteristic property of the atom. Finally, to define an entropy measure for graphs, we use the property of each atom to define normalized values which can be interpreted as vertex probabilities. On the basis of these values, which form a probability distribution, we finally obtain the information indices $I_{f_\star^V}(G)$ and $I_{f_\star^V}^\lambda(G)$. As a remark, we note that in this paper, as in many other contributions dealing with topological indices, we only consider skeletons of the chemical structures, that is., all atoms and all bonds are equal (the graphs represent elements of $\mathscr{G}_{UC}$). An extension of this concept for weighted graphs is in development.

RESULTS AND DISCUSSION

The aim of this section is to investigate the defined family of graph entropy measures numerically. One important problem is to study the distributions of the descriptor values for different graph classes for fixed parameters in the descriptor definition. From this, we will gain novel insights to understand which structural features are crucial for interpreting the resulting (concrete) information indices. Additionally, we also interpret the relatedness to some other molecular descriptors. For finalizing our numerical section, we present a plot that shows the extremal behavior of a concretely chosen information index (see Definitions 5 and 6) by considering certain graph classes.

Software and Data Processing. For generation and handling of chemical structures, the Molfile format (a text format) has been used.[51] For the generation of exhaustive sets of isomeric chemical structures from a given molecular formula and defined structural restrictions, we used software Molgen.[36] Because only skeletons have been considered, only the element carbon has been used in molecular formulas and only single bonds have been allowed. All structures generated under these conditions have different skeletons. The check for isomorphic skeletons, of real chemical structures taken from a spectral database, has been performed by software SubMat.[52,53] Our entropy-based descriptors were calculated from Molfile structures by using a Python program. First, the Python program parses a Molfile and creates an adjacency matrix representing an undirected and connected graph (skeleton of the chemical structure). Second, the program determines the j-spheres for every vertex of such a graph and, hence, the entropy according to eq 9. Furthermore, the software Dragon[40,54] has been used for the calculation of a few molecular descriptors applied to 2265 structures (2-dimensional, H-depleted) selected from the mass spectral database. For statistical evaluation[55] the programming language R[56] has been used.

Concrete Information Indices. We first derive two concrete information indices by using our definition of the family of information measures, see equation 9.

Definition 5. Let $G \in \mathcal{G}_{UC}$ be a graph. We define $f^{V}_{\star,\mathrm{lin}}(v_i)$ if the coefficients $c_1, c_2, ..., c_{\rho(G)}$ satisfy the property

$$c_1 := \rho(G),\ c_2 := \rho(G) - 1, ..., c_{\rho(G)} := 1 \quad (10)$$

Then, $I^{\lambda}_{f_{\star,\mathrm{lin}}{}^{V}}(G)$ represents an information index with coefficients linearly decreasing with increasing topological distance. We use the notation "I lin 01" for this descriptor.

Definition 6. Let G $\in \mathcal{G}_{UC}$ be a graph. We define $f^{V}_{\star,\mathrm{quad}}(v_i)$ if the coefficients $c_1, c_2, ..., c_{\rho(G)}$ satisfy the property

$$c_1 := \rho(G)^2,\ c_2 := (\rho(G) - 1)^2, ...,$$
$$c_{\rho(G)} := (\rho(G) - (\rho(G) - 1))^2 = 1 \quad (11)$$

Then, $I^{\lambda}_{f_{\star,\mathrm{quad}}{}^{V}}(G)$ represents an information index with coefficients quadratically decreasing with increasing topological distance. We use the notation "I quad 01" for this descriptor.

Graph Classes. *Definition 7.* To perform our analysis, we further define the following graph classes:

- MS 2265*:* 2265 selected chemical structures with different skeletons originating from the mass spectral database NIST.[35] The whole database contains approximately 100 000 chemical structures from organic compounds that have been typically analyzed by MS or GC/MS during the last decades. A set of 4000 structures with 4−19 non-hydrogen atoms have been randomly selected from the database; 2265 of them have different skeletons (all atoms and all bonds considered to be equal, all hydrogen atoms removed) and formed the graph class MS 2265. Thus, the number of atoms per structure (vertices per graph) is between 4 and 19 (mean 13.6).
- C10 trees−C14 trees: These five graph classes consist of all alkane isomers with 10−14 carbon atoms corresponding to molecular formulas C_nH_{2n+2} (n = 10−14). The number of structures with 10, 11, 12, 13, and 14 carbon atoms is 75, 159, 355, 802, and 1858, respectively.
- C10 ring 1−C14 ring 1: These five graph classes consist of all hydrocarbon isomers with 10−14 carbon atoms containing one ring (cycle) and only single bonds, corresponding to molecular formulas C_nH_{2n} (n = 10−14). The number of structures with 10, 11, 12, 13, and 14 carbon atoms is 475, 1231, 3232, 8506, and 22 565, respectively.
- C10 ring 2−C14 ring 2: These five graph classes consist of all hydrocarbon isomers with 10−14 carbon atoms containing two rings (cycles) and only single bonds, corresponding to molecular formulas C_nH_{2n-2} (n = 10−14). The number of structures with 10, 11, 12, 13, and 14 carbon atoms is 1792, 5533, 16 977, 51 652, and 156 291, respectively.

We want to emphasize that by using the term "exhaustive", we do not mean all possible graphs up to a certain number of vertices. Instead, it means that only these graphs are included which are conform with the rigid boundaries of the given graph classes, for example, C10 ring1: We only consider these (but all) graphs containing exactly one ring with 10 vertices. Also, this procedure led us to use the term "synthetic" chemical structures for avoiding that the reader could think the whole space (exhaustive) of all possible graphs up to a certain number of vertices is meant.

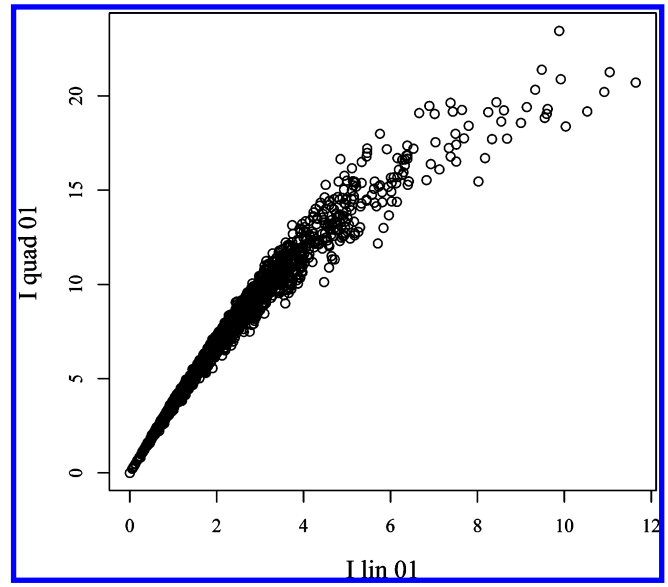

Figure 1. Comparison of the indices "I lin 01" and "I quad 01" for the structure set MS 2265.

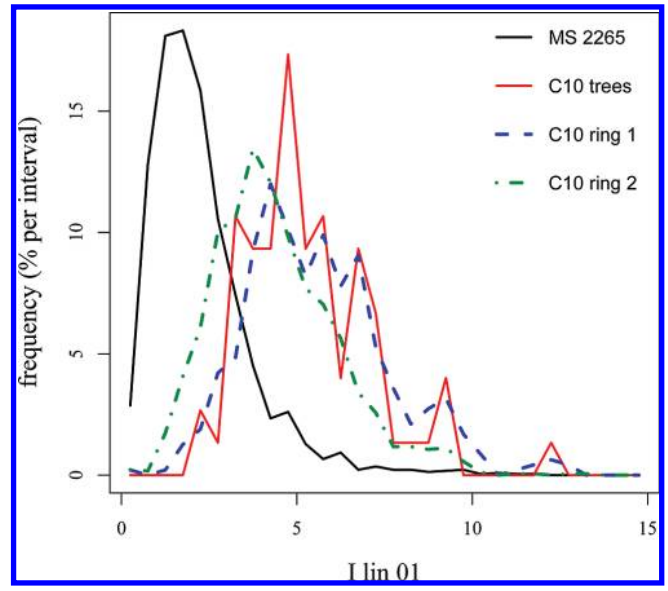

Figure 2. Distributions for structure set MS 2265 and generated isomers with 10 vertices.

Numerical Results. Figure 1 compares the values for index "I lin 01" (with coefficients linearly decreasing with increasing topological distance) and for index "I quad 01" (with coefficients quadratically decreasing with increasing topological distance) for the structure set MS 2265. The indices show a high correlation which is almost linear up to values of about 5 for "I lin 01". For the investigated isomeric structures the relationship is very similar (not shown). For this reason, in our further considerations, we only use the entropy-based descriptor "I lin 01".

In the following, we discuss the characteristic value distributions calculated for the index "I lin 01" and for the just defined graph classes. In the Figures 2−5, the x-axis represents the range of the index, and the y-axis represents the frequency of graph structures (in percent) per interval of width 0.5.

Note that the used index characterizes the diversity of the vertices in terms of their neighborhood. The higher the value

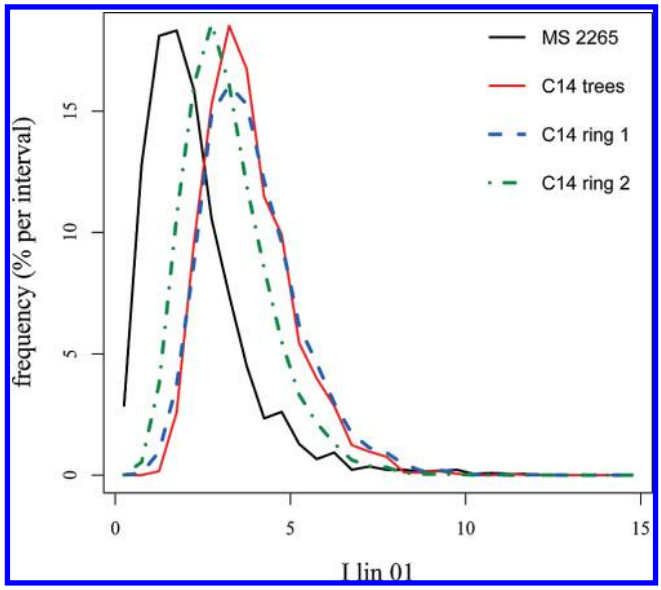

Figure 3. Distributions for structure set MS 2265 and generated isomers with 14 vertices.

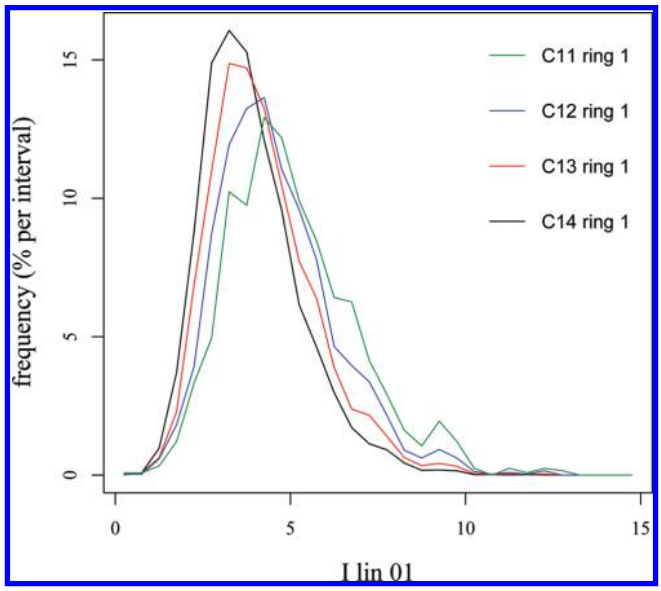

Figure 4. Distributions for generated isomers containing one ring.

of the index, the more topologically different vertices are in the graph - and the lower is the inner symmetry of the corresponding chemical structure. The distributions in Figure 2 and Figure 3 show that the synthetic (and exhaustively generated skeletons have in general more diverse vertices (higher values of "I lin 01") than the skeletons taken from the spectroscopic database (MS 2265). This difference is larger for isomeric skeletons with 10 vertices than for isomeric skeletons with 14 vertices. Reasons for the different distributions may be 2-fold: (1) Isomer generation does not consider any stability of the created structures, and of course not all created isomers may be stable chemical compounds. (2) The skeletons from chemical structures present in the spectroscopic database have in general a higher inner symmetry (more topologically equivalent or similar atoms) than the exhaustively created isomers.

From the distributions in Figures 4 and 5 the dependence of the used index "I lin 01" on the number of vertices can be seen. For structures with one ring and for tree structures

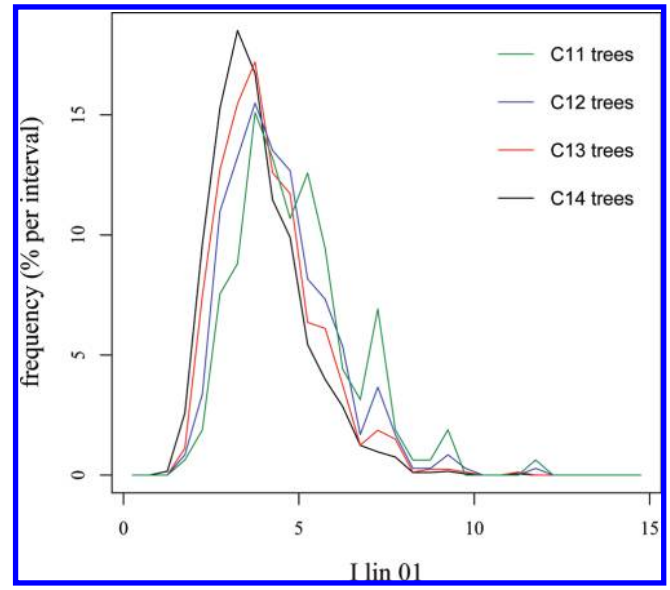

Figure 5. Distributions for generated isomers representing trees.

the distribution is shifted to lower values with increasing number of vertices. The same trend is with structures containing two rings (not shown). However, the dependence of the index value on the number of vertices is small, and we conclude that the index is prominently influenced by the inner symmetry but not by the size of a structure.

The interpretation from Figures 2 and 3 is that real chemical structures are a nontrivial mixture of different classes of synthetic chemical structures and none of these *pure* classes seems to correspond well to MS 2265. The conclusion from this is that it does not seem to make sense to study descriptors by using synthetic graph structures because they have characteristics which are different to real chemical structures. This seems illogical at first because, in principle, each real chemical network structure can be found in one class of the generated data. However, the crucial point is practically we did not consider an exhaustive set consisting of all possible graph skeletons. Instead, subsets have been used, for example, tree or ring structures with a certain number of vertices. The reason for using just subsets is for combinatorial reasons. The number[37] of all possible structures consisting of up to 19 vertices (the maximal number of vertices found in the MS 2265 data set) gives $2^{\binom{19}{2}}$. That means an exhaustive set containing all different graphs up to this number of vertices is practically intractable. For this reason we use for the following investigations the MS 2265 data set because synthetic data, characterizing a subset of the complete set of different graphs, lack important characteristics of real chemical structures.

Before comparing "I lin 01" with other known molecular descriptors, we examine its degeneracy. This relates to investigate the discriminating power of a descriptor, that is, its ability to distinguish nonisomorphic graphs.[47] To tackle this problem, we evaluate the sensitivity measure[47]

$$S(I) = \frac{|N| - |N_i|}{|N|} \tag{12}$$

of a topological index I with respect to a set of graphs denoted by N. Generally, $|N|$ stands for the cardinality of N and $|N_i|$ stands for the number of graphs $G_j \in N$, which can

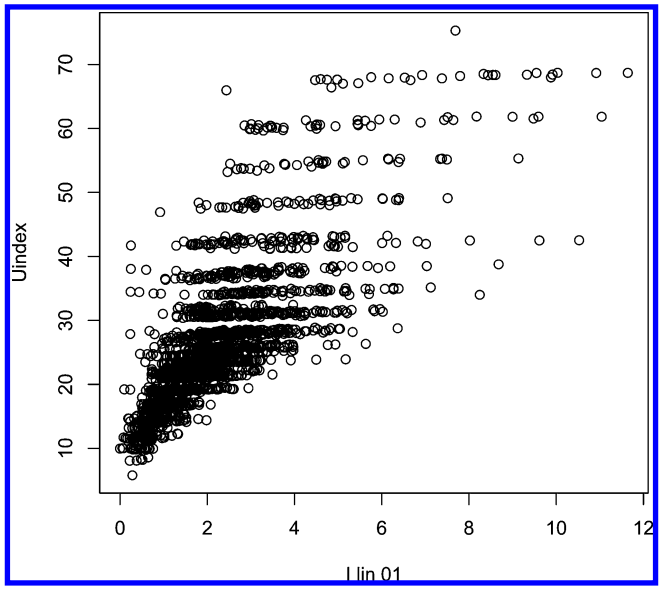

Figure 6. Balaban's information index "*U* index" versus our "I lin 01" for the structure set MS 2265.

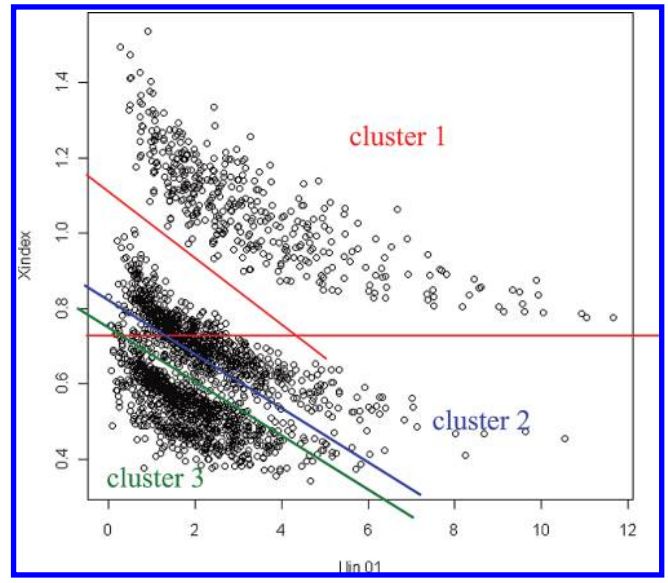

Figure 7. Balaban's information index "*X* index" versus our "I lin 01" for the structure set MS 2265.

not be distinguished (by using *I*), respectively. By definition, it holds $S(I) = 1$ iff it does not exist any pair of nonisomorphic graphs $G \in N$ whose graphs possess the same value of *I*. Now, starting from $I =$ "I lin 01" and $N =$ MS 2265, we calculate S (I lin 01)$= (2265 - 4)/(2265) = 0.998233$. [The comparison of the numerical values of "I lin 01" was performed by considering 6 decimal places.] That means there are only 2 pairs of graphs (4 graphs) whose values of "I lin 01" are pairwise equal. Further, we found for the Wiener index W, $S(W) =(2265 - 2113)/(2265) = 0.067108$. This result is not surprising because it is known that W is highly degenerated.[47] However, note that "I lin 01" is zero for all vertex transitive graphs (simple ring structures, prisman, tetrahedron, cube). The sensitivity measure to detect degeneracy depends of the structure set under consideration, and it is important to use real chemical structures. Figure 11 shows the two pairs of graphs explicitly. As a final note, the high value for sensitivity, even for skeletons, makes the descriptor "I lin 01" a potential candidate for a new identification code for chemical structures that could be used for fast structure presearches in databases.

From the many molecular descriptors described in ref 40, five have been selected for a comparison with our new index "I lin 01". The descriptors have been calculated by the software Dragon[54] for the structure set MS 2265. Figure 6 − Figure 10 show the relationships between "I lin 01" and the five selected descriptors. Figures 6 and 7 contain the information indices "*U* index" and "*X* index" suggested by Balaban[16,40] (notation as used in software Dragon), which are related but differently defined as our entropy-based descriptor. The "*U* index" is defined by[16]

$$U(G) = \frac{|E|}{\mu + 1} \sum_{(v_i,v_j)\in E} [u(v_i)u(v_j)]^{-\frac{1}{2}} \qquad (13)$$

where

$$u(v_i) = -\sum_{j=1}^{\sigma(v_i)} \frac{jg_j}{d(v_i)} \log\left(\frac{j}{d(v_i)}\right) \qquad (14)$$

and

$$d(v_i) = \sum_{j=1}^{|V|} d(v_i, v_j) \qquad (15)$$

The "*X* index" is very similarly defined (the arguments $u(v_i)$,$u(v_j)$ are slightly differently defined). The main difference between the indices of Balaban and our "I lin 01" is that the resulting information indices ("*U* index" and "*X* index") are based on local vertex entropies. These entropies can be expressed by eq 14. In contrast, our graph entropy measure is a direct result from applying Shannons's entropy formula considering the full topological neighborhood of all atoms. Interestingly, the local vertex entropies in Balaban's indices, as well as our information functional $f^{V}_{\star}$, have been defined by using the *j*-sphere cardinalities. This could be one reason why "*U* index" has almost the same value for sets of structures with different values for "I lin 01". Further, we observe that "*U* index" is positively correlated with "I lin 01", "*X* index" is negatively correlated, both correlations are very weak.

In the scatter plot "*X* index" versus "I lin 01" (see Figure 7), three clusters appear, two of them are close together. Concretely, we found that cluster 1 does not contain ring structures. Further, within cluster 2 and cluster 3 the structures are more branched than in cluster 1. We want to emphasize that these clusters are the result of exploring the correlation between the two mentioned information indices. This result also shows that "*X* index" and "I lin 01" capture structural information of the considered graphs differently.

The mean Wiener index "WA" (a topological descriptor)[40] is highly correlated with "I lin 01" as shown in Figure 8. This can be interpreted in a way that "WA" and "I lin 01" capture structural information very similarly. We note that both descriptors are generally based on distances in a graph but "I lin 01" is an information index and "WA" is not based on using Shannon entropy measures. In contrast, the molecular walk count index "MWV09"[40] (number of walks of length 9) is highly uncorrelated with "I lin 01", see Figure 9. To finalize this discussion, Figure 10 shows the compari-

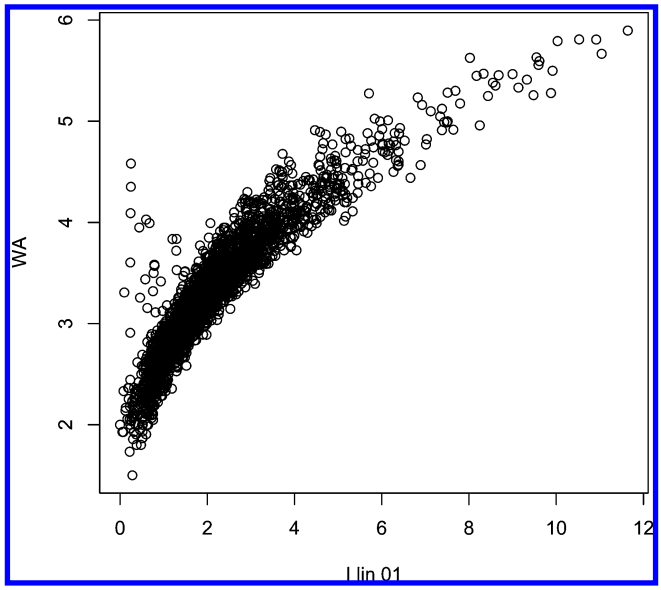

Figure 8. Mean Wiener index "WA" versus our "I lin 01" for the structure set MS 2265.

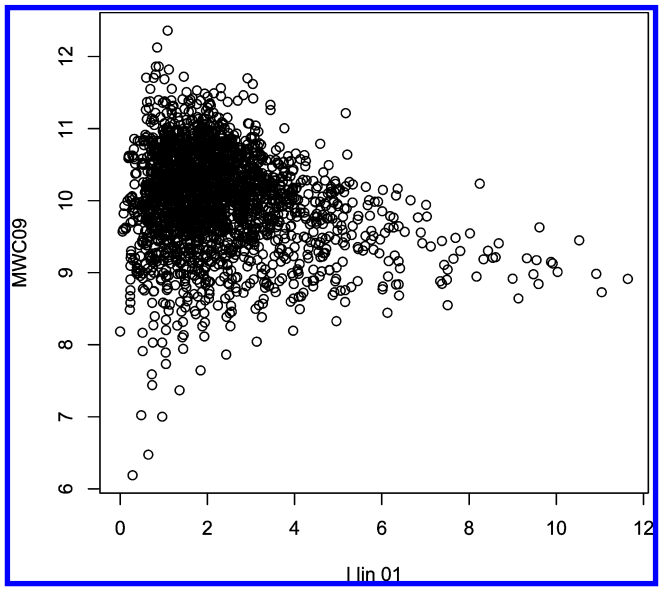

Figure 9. Index "MWV09" versus our "I lin 01" for the structure set MS 2265.

son of the Randić connectivity index ("X1" in Dragon) and "I lin 01". Starting from a graph for $G = (V,E)$, the Randić connectivity index "$X1$" is defined by[4]

$$R(G) = \sum_{(v_i,v_j)\in E} [\delta(v_i)\delta(v_j)]^{-\frac{1}{2}} \qquad (16)$$

Interestingly, we found that graphs with a large value of "$X1$" and a small value of "I lin 01" are rather symmetric structures with multiple rings. The observation that the values of "I lin 01" are very small for these structures corresponds the fact that this information measure ("I lin 01") vanishes for highly symmetric structures, for example, cyclic graphs, k-regular graphs etc. We already described the kind of structural complexity which can be captured by "I lin 01" when interpreting the characteristic value distributions. Once again, this clearly shows that kind of structural complexity that can be captured by "$X1$" and "I lin 01" is different. Similarly as in Figure 9, "$X1$" is highly uncorrelated with "I

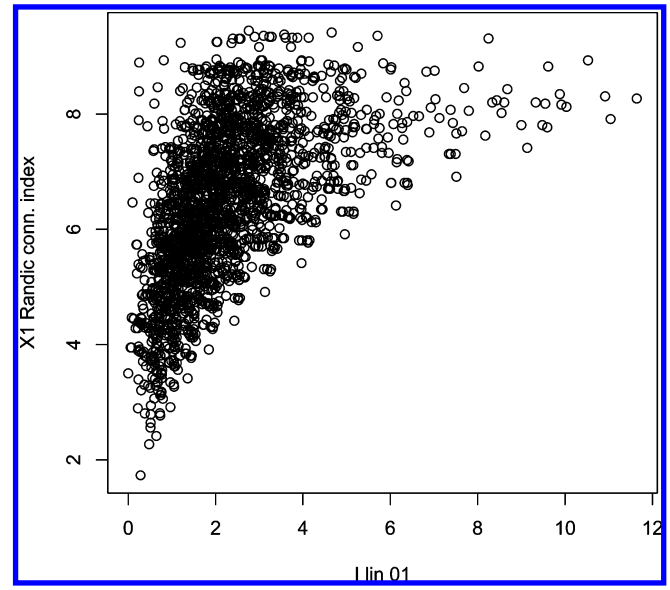

Figure 10. Randić connectivity index and "I lin 01" for the structure set MS 2265.

lin 01", see Figure 10. It is well-known that maximal values of $R(G)$ (i.e., "$X1$") appear for linear chains. Actually, the points on a smooth curved line at the lower border of the scatterplot are from linear chains with 5 vertices (lower left side corner) up to 17 vertices (upper right-hand side corner). For the structures in the graph class MS 2265 the linear chains define an upper limit for "I lin 01".

To interpret the extremal entropy values of the information index "I lin 01", we interpret Equation 9 and search for minimum and maximum values using our graph classes (see Definition 7). The results are summarized in Figure 12. Before expressing the concrete results, we state some general mathematical properties of Equation 9. If all vertices are topologically equivalent (in a vertex-transitive graph), they have the same vertex probabilities $p^V_\star(v_i)$; thus the sum term in eq 9 becomes to $-\log(|V|)$; $|V|$ is the number of vertices in the graph. Consequently, the index value is zero for all vertex-transitive graphs, the smallest possible value. Examples are simple ring structures with any number of atoms. However, also molecular structures like a tetrahedron, a cube, prisman or similar, or C_{60} fullerene and similar have topologically identical atoms and a descriptor value equal to zero. Finally, we find that the value of the information index increases with increasing "neighborhood diversity".

We now start the concrete interpretation of Figure 12 by considering isomeric structures with 12 carbon atoms. For connected trees, a linear chain maximizes the molecular descriptor "I lin 01". In contrast, an extensively branched tree led to the minimal entropy. We notice that this minimal entropy is greater than zero. As discussed above, if we consider structures containing one ring only, the fully symmetric graph represented by a simple ring structure leads to an entropy equal to zero.

The structure with one ring which maximizes "I lin 01" is a chain-like graph but also contains one ring. Clearly, this corresponds to the fact that the atoms (vertices) are much more diverse in terms of their neighborhood relations induced by the j-spheres. The just explained situation holds similarly for structures which contain two rings. We see that the minimal entropy of such a graph is greater than zero because

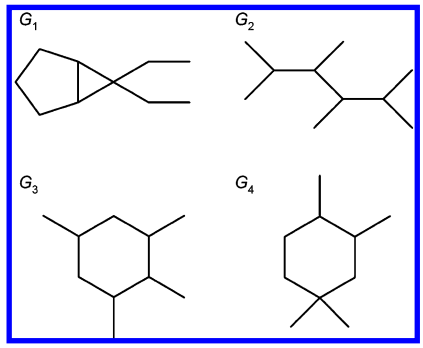

Figure 11. The chemical structures G_1, G_2 and G_3, G_4 contained in MS 2265 can not be distinguished by "I lin 01".

	minimum *I lin 01*	maximum *I lin 01*
trees	1.8413	11.6165
one ring	0 (12-ring)	12.4312
two rings	0.1678 (11-ring and 3-ring)	12.7809
spectral database	0 (7-ring)	11.6431

Figure 12. Extremal entropy values of "I lin 01" regarding the defined graph classes. For trees, one ring graphs and two ring graphs, 12 vertices have been considered.

the structure is not fully symmetric. Maximum entropy holds again for a chain-like graph containing two rings. Regarding the spectral database MS 2265, we obtain the result that a fully cyclic structure with seven vertices leads to an entropy that equals zero. This result is of course included into the above outlined observation that the entropy for fully cyclic structures always vanishes. The chemical structure of MS 2265 that maximizes the entropy is very similar to a chain-graph but contains an inner vertex with degree equal to three.

SUMMARY AND CONCLUSION

In this paper, we studied entropy-based molecular descriptors. In particular, we investigated a family of graph entropy measures by using real chemical and synthetic chemical structures (isomers). Moreover, we compared our entropy-based molecular descriptor with other known molecular descriptors. From our large scale numerical analysis, we obtained the following results:

- The synthetic data corresponding to subclasses (trees, one ring, two rings) of isomeric structures do not capture all structural information contained in real chemical structures. This could be revealed by comparing the distributions of descriptor values in dependence on the data set (Figures 2 and 3). This result suggests not to use synthetic data for the analysis of molecular descriptors because, as discussed in the result section, one would need a mixture of different classes (whose compositions is unknown currently) to obtain characteristics that correspond to real chemical structures. The reason therefore was that the synthetic generated structures are not enumerative due to the combinatoric explosion of possibilities and, hence, highly incomplete. A subset of randomly generated isomeric structures (not restricted to certain subclasses) would increase the structural diversity. But on the other hand, such a set would also include thermodynamically unstable molecules.
- We used a random sample of chemical structures present in a spectroscopic library for practical tests of our descriptors; structure libraries with drugs is alternative.
- With respect to MS 2265, it turned out that the degeneracy of "I lin 01" is low. We found that only 4 graphs could not distinguished. As expected, the degeneracy of the Wiener index for MS 2265 is very high.
- Entropy-based molecular descriptors seem to be classifiable according to their behavior with respect to the analyzed data set. For example, the mean Wiener index (WA) and "I lin 01" are strongly positive correlated for the MS 2265 data set. That means they are in the same class because both descriptors provide very similar results. In contrast, "MWV09" and "I lin 01", "*X*1" and "I lin 01" are highly uncorrelated. Hence, they reveal complementary structural information from the underlying network. From this result an intriguing question to study would be how many different descriptors exist among all descriptors developed so far. Because this would not only be interesting from a theoretical point of view but also allow to obtain a lower bound on the *dimension* of the chemical structures corresponding to the number of distinct classes found. Further studies are necessary to shed light on this exiting point.
- The introduced entropy-based descriptor can be considered as a generalization of some classical ones.[40] Classical graph entropy measures are often based on the problem of inducing vertex partitions of a graph in question (e.g., Trucco, Rashevsky, Mowshowitz). For example, such classical graph entropy measures were obtained by determining automorphism groups of graphs (vertex orbits) and chromatic decompositions. It is well-known that the time complexity of the underlying algorithms is very high (often NP-complete). Dehmer[17] proved that the computation of the graph entropy measure by using our information functional is polynomial. For large networks, none of the just mentioned classical information indices (based on algebraic principles) has polynomial time complexity. Further, our presented information measure possesses two important properties: (i) it is not based on inducing vertex partitions (because a probability value is assigned to every vertex in a graph) and (ii) we use parametric information functionals. From this, it follows that the corresponding graph entropy measures generalize the most known information indices used in mathematical chemistry.[8] The fact that our graph entropy measures are parametric also leads to an interesting connection to machine learning problems (by using appropriate data sets, the parameters could be learned). To examine this connection and to compare our entropy-based descriptors with other existing topological descriptors mathematically will be a part of our future research. This thread of thoughts has been already contemplated.[17]

ACKNOWLEDGMENT

We thank Alexandru T. Balaban, Subhash Basak, Danail Bonchev, Armin Graber, Abbe Mowshowitz, Roberto Todes-

chini, and the anonymous referees for fruitful discussions and valuable comments. Also, we are grateful to A. Kerber and R. Laue for providing the software Molgen, S. E. Stein for the NIST mass spectral database, and H. Sesibrany for development of the software SubMat. Parts of this work have been supported by the Center for Mathematics (CMUC), University of Coimbra, Portugal. Moreover, this work was supported by the COMET Center ONCOTYROL and funded by the Federal Ministry for Transport Innovation and Technology (BMVIT) and the Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ), the Tiroler Zukunftsstiftung (TZS), and the State of Styria represented by the Styrian Business Promotion Agency (SFG) [and supported by the University for Health Sciences, Medical Informatics and Technology and BIOCRATES Life Sciences AG].

REFERENCES AND NOTES

(1) Basak, S. C.; Magnuson, V. R. Molecular Topology and Narcosis. *Arzeim. Forsch./Drug Design* **1983**, *33*, 501–503.
(2) Basak, S. C.; Gute, B. D.; Balaban, A. T. Interrelationship of Major Topological Indices Evidenced by Clustering. *Croat. Chem. Acta* **2004**, *77*, 331–344.
(3) Bonchev, D.; Rouvray, D. H. *Chemical Graph Theory. Introduction and Fundamentals*; Abacus Press: New York, 1991.
(4) Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers: Amsterdam, The Netherlands, 1999.
(5) Diudea, M. V.; Gutman, I.; Jäntschi, L. *Molecular Topology*; Nova Publishing: New York, 2001.
(6) Trinajstíc, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.
(7) Diudea, M. V. *QSPR/QSAR Studies by Molecular Descriptors*; Nova Publishing: Huntington, NY, 2001.
(8) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K., 1983.
(9) Bonchev, D. Overall Connectivities and Topological Complexities: A New Powerful Tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci* **2000**, *40*, 934–941.
(10) Bonchev, D. *Complexity in Chemistry. Introduction and Fundamentals*; Taylor and Francis: Boca Raton, FL, 2003.
(11) Minoli, D. Combinatorial graph complexity. *Atti. Accad. Naz. Lincei, VIII. Ser., Rend., Cl. Sci. Fis. Mat. Nat.* **1975**, *59*, 651–661.
(12) Randić, M.; Plavšić, D. P. On the Concept of Molecular Complexity. *Croat. Chem. Acta* **2002**, *75*, 107–116.
(13) Bertz, S. H.; Zamfirescu, C. M. New Complexity Indices based on Edge Covers. *MATCH* **2000**, *42*, 39–70.
(14) Bonchev, D.; Rouvray, D. H. *Complexity in Chemistry, Biology, and Ecology*; Mathematical and Computational Chemistry Springer: New York, 2005.
(15) Bonchev, D. Information Indices for Atoms and Molecules. *MATCH* **1979**, *7*, 65–113.
(16) Balaban, A. T.; Balaban, T. S. New Vertex Invariants and Topological Indices of Chemical Graphs Based on Information on Distances. *J. Math. Chem.* **1991**, *8*, 383–397.
(17) Dehmer, M. Information-theoretic Concepts for the Analysis of Complex Networks. *Appl. Artif. Intell.* **2008**, *22*, 684–706.
(18) Morowitz, H. Some order-disorder considerations in living systems. *Bull. Math. Biophys.* **1953**, *17*, 81–86.
(19) Quastler, H. Information Theory in Biology. *Bull. Math. Biol.* **1954**, 183–185.
(20) Dancoff, S. M.; Quastler, H. Information Content and Error Rate of Living Things. In *Essays on the Use of Information Theory in Biology*; Quastler, H., Ed.; University of Illinois Press: Urbana, IL, 1953; pp 262−273.
(21) Linshitz, H. The Information Content of a Battery Cell. In *Essays on the Use of Information Theory in Biology*; Quastler, H., Ed.; University of Illinois Press: Urbana, IL, 1953; pp 251−262.
(22) Rashevsky, N. Life, Information Theory, and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229–235.
(23) Trucco, E. A note on the information content of graphs. *Bull. Math. Biol.* **1956**, *18*, 129–135.
(24) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1997.
(25) Mowshowitz, A. Entropy and the complexity of the graphs I: An index of the relative complexity of a graph. *Bull. Math. Biophys.* **1968**, *30*, 175–204.
(26) Mowshowitz, A. Entropy and the complexity of graphs II: The information content of digraphs and infinite graphs. *Bull. Math. Biophys.* **1968**, *30*, 225–240.
(27) Mowshowitz, A. Entropy and the complexity of graphs III: Graphs with prescribed information content. *Bull. Math. Biophys.* **1968**, *30*, 387–414.
(28) Mowshowitz, A. Entropy and the complexity of graphs IV: Entropy measures and graphical structure. *Bull. Math. Biophys.* **1968**, *30*, 533–546.
(29) Bonchev, D.; Trinajstić, N. Information theory, distance matrix and molecular branching. *J. Chem. Phys.* **1977**, *67*, 4517–4533.
(30) Emmert-Streib, F.; Dehmer, M. Fault tolerance of information processing in gene networks. *Phys. A* **2009**, *338*, 541–548.
(31) Solé, R. V.; Valverde, S. Information Theory of Complex Networks: On Evolution and Architectural Constraints. *Lect. Notes Phys.* **2004**, *650*, 189–207.
(32) Dehmer, M.; Emmert-Streib, F. Structural Information Content of Networks: Graph Entropy based on Local Vertex Functionals. *Comput. Biol. Chem.* **2008**, *32*, 131–138.
(33) Emmert-Streib, F.; Dehmer, M. Information Theoretic Measures of UHG Graphs with Low Computational Complexity. *Appl. Math. Comput.* **2007**, *190*, 1783–1794.
(34) Dehmer, M.; Borgert, S.; Emmert-Streib, F. Entropy Bounds for Molecular Hierarchical Networks. *PLoS ONE* **2008**, *3*, e3079.
(35) *NIST, Mass Spectral Database 98*; National Institute of Standards and Technology: Gaithersburg, MD, 1998; www.nist.gov/srd/nist1a.htm.
(36) *Molgen Isomer Generator Software*; Institute of Mathematics II, University of Bayreuth: Bayreuth, Germany, 2000; www.molgen.de.
(37) Harary, F. *Graph Theory*; Addison Wesley Publishing Company: Reading, MA, 1969.
(38) Halin, R. *Graphentheorie*; Akademie Verlag: Berlin, Germany, 1989.
(39) Skorobogatov, V. A.; Dobrynin, A. A. Metrical Analysis of Graphs. *MATCH* **1988**, *23*, 105–155.
(40) Todeschini, R.; Consonni, V.; Mannhold, R. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2002.
(41) Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3241–3243.
(42) Bonchev, D. The Overall Topological Complexity Indices. In *Advances in Computational Methods in Science and Engineering*, Simos, T., Maroulis, G., Eds.; VSP Publications: Zeist, The Netherlands, 2005; Vol. 4B, pp 1554−1557.
(43) Bonchev, D. Kolmogorov's information, Shannon's entropy, and topological complexity of molecules. *Bulg. Chem. Commun.* **1995**, *28*, 567–582.
(44) Bonchev, D. The Overall Wiener Index - A New Tool for Characterization of Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 582–592.
(45) Bonchev, D.; Trinajstić, N. Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR QSAR Environ. Res.* **2001**, *12*, 213–236.
(46) Hendrickson, J. B.; Huang, P.; Toczko, A. G. Molecular complexity: A simplified formula adapted to individual atoms. *J. Chem. Inf. Comput. Sci.* **1987**, *2*, 63–67.
(47) Konstantinova, E. V.; Skorobogatov, V. A.; Vidyuk, M. V. Applications of Information Theory in Chemical Graph Theory. *Indian J. Chem.* **2002**, *42*, 1227–1240.
(48) Konstantinova, E. V.; Paleev, A. A. Sensitivity of topological indices of polycyclic graphs. *Vychisl. Sistemy* **1990**, *136*, 38−48, In Russian.
(49) Gregori-Puigjané, E; Mestres, J. Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.
(50) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(51) Gasteiger, J.; Engel, T. *Chemoinformatics—A Textbook*; Wiley VCH: Weinheim, Germany, 2003.
(52) Scsibrany, H.; Varmuza, K. *Software SubMat*; Vienna University of Technology, Institute of Chemical Engineering, Laboratory for Chemometrics: Vienna, Austria, 2004; www.lcm.tuwien.ac.at.
(53) Varmuza, K.; Scsibrany, H. Substructure isomorphism matrix. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 308–313.
(54) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *Dragon, Software for Calculation of Molecular Descriptors*; Talete srl: Milano, Italy, 2004; www.talete.mi.it.
(55) Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; Francis & Taylor, CRC Press: Boca Raton, FL, 2009.
(56) *R: A language and Environment for Statistical Computing*; R Development Core Team, Foundation for Statistical Computing: Vienna, Austria, 2008; www.r-project.org.

CI900060X