MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **67** (2012) 147-172

ISSN 0340 - 6253

Uniquely Discriminating Molecular Structures Using Novel Eigenvalue–Based Descriptors

Matthias Dehmer[a,*], Lavanya Sivakumar[a], Kurt Varmuza[b]

[a] *Institute for Bioinformatics and Translational Research, UMIT,*
Eduard Wallnoefer Zentrum 1, A-6060, Hall in Tyrol, Austria

[b] *Laboratory for Chemometrics, Institute of Chemical Engineering,*
Vienna University of Technology, Getreidemarkt 9/166,
A-1060 Vienna, Austria

(Received March 7, 2011)

Abstract

In this article, we explore novel spectra-based descriptors to discriminate molecular graphs. As known, classical structure descriptors based on the eigenvalues of the underlying adjacency matrix are often insufficient since there exist a large number of isospectral graphs. Briefly recall that the spectrum is the set of eigenvalues of the characteristic polynomial. To tackle the problem, we propose five families of novel descriptors based on the eigenvalues of certain molecular matrices representing chemical structures. Note that in this paper, we only consider the underlying skeleton of a molecular graph. Because it is crucial to study the discrimination power (often called degeneracy) by not merely using synthetic (isomeric) structures, we apply the novel measures to both real and synthetic molecular graphs. Also, we use ten different types of molecular matrices to calculate the novel descriptors and determine correlations between them. It turns out that the novel descriptors possess high discrimination power when being applied to appropriate molecular matrices. Evidently, the study also reveals that special kinds of matrices capture structural information of the molecular graphs more meaningfully than others, particularly the adjacency matrix which turned out often to be insufficient to develop molecular descriptors.

1 Introduction

Structural descriptors for investigating molecular structures have been extensively explored in mathematical, structural and computational chemistry, and related disciplines; see [3–5, 7, 55]. These measures have been often used to characterize molecular networks

[*] Corresponding Author. E-mail: matthias.dehmer@umit.at

since they quantify structural features thereof. In fact, numerous descriptors have been contributed but no such graph invariant can characterize the underlying graph topology completely. Thus, it is still an outstanding problem to investigate structural graph measures having high discrimination power when being applied to various graph classes.

In terms of structural chemistry and drug design, molecular descriptors have also been extensively employed, e.g., for analyzing and determining physico-chemical properties of the underlying compounds. Indeed, QSAR and QSPR are well-known application areas where molecular structure descriptors have been proven useful; see [15]. In particular, the prediction of biological and pharmacological properties using descriptors has been a problem of great interest and importance [16].

If one takes a closer look at the developed descriptors, one can realize that over the years, various mathematical methods from different fields such as combinatorics, statistics and information-theory have been employed to tackle the challenging problem of characterizing the complexity of molecules [6, 11, 52]. Interestingly, the analysis of biological networks has triggered the insight that statistical techniques should be employed because the underlying networks are often affected by measurement errors [20]. In particular, entropy-based graph measures [6, 11] turned out to be meaningful when characterizing graphs structurally. A major reason is that measures based on Shannon's entropy can be understood as a cumulation of other (local) quantities representing probabilities that capture structural information of a graph. Other classical descriptors like the Wiener or Randić index [39, 56] are simply based on deriving structural features (e.g., distances or degrees) to obtain a single numerical value characterizing the complexity of the molecular graph.

In this paper, we focus on developing and investigating spectra-based descriptors using molecular matrices [32]. The main contribution besides examining the novel descriptors is to demonstrate their ability and usefulness to tackle problems in structural chemistry. Among the existing spectra-based measures are indices that only take a single eigenvalue, e.g., the maximal eigenvalue or $p < n$ ($|V| := n$) eigenvalues into account. Hence, we put the emphasis on such measures taking the complete spectrum of different molecular matrices into account. Those measures have not been well explored yet. In any way, note that only measures that capture significant structural information have the potential to be applied. Finally, the degeneracy problem is also a crucial issue in structural chemistry that deserves special attention.

As known, eigenvalue-based measures relying on the adjacency matrix have been extensively explored [31, 40], but as mentioned, they often show lack because of their less discrimination power. Also, note that graph polynomials have been used to characterize chemical structures with the aid of coefficients or the zeros of the polynomials [31, 40]. In order to sketch the most known approaches in this area in brief, we note that many descriptors are generally defined by using either the elements or the characteristic polynomial of a molecular matrix [16, 31, 40]. Early work to study the eigenvalues of graphs dates back to Lovász et al. [37] and Cvetkovic [10]. Lovász et al. suggested that the leading eigenvalue of the adjacency matrix can be used to detect molecular branching of hydrocarbons. Leading eigenvalues of other molecular matrices such as the distance matrix have also been studied [43]. Also, Gutman [24, 25] studied the sum of the absolute values of eigenvalues of the adjacency matrix of a graph (called graph energy) and related it to the total π-electron energy involved in the formation of hydrocarbon molecules. For studying the DNA structure of different species and folding of proteins, Randić et al. [45] proposed descriptors representing sums of positive eigenvalues and also based on the multiplicity of the zero eigenvalue of molecular matrices. Apart from these measures, there are also various measures based on the statistical quantities, such as mean, absolute deviation, variance of the eigenvalues inferred from molecular matrices [9, 29, 30, 53]. Estrada [22, 23] dealt with studying exponential functions when analyzing the degree of folding of proteins to accommodate both positive and negative eigenvalues of an adjacency matrix. Recently, Consonni and Todeschini [9] proposed two classes of eigenvalue based descriptors namely the absolute deviation and mean absolute deviation of eigenvalues derived from matrices of weighted molecular graphs. But as already mentioned, most of the known eigenvalue-based descriptors are not discriminative enough to characterize graphs meaningfully. Also, they do not capture structural information meaningfully and, hence, they can not be applied to tackle important problems in quantitative graph theory [21, 38] such as graph classification and related problems.

The paper is organized as follows: In Section 2, we introduce some graph-theoretical terminology. Section 3 sketches the most important contribution towards the uniqueness of descriptors. In Section 4, we propose novel descriptors based on the eigenvalues of molecular matrices. The molecular matrices to be used in this paper are described in

Section 5. Section 6 outlines the databases we are going to use to perform our study and provides numerical results when determining the uniqueness of the descriptors as well as correlations between them. The paper ends with a summary and conclusion in Section 7.

2 Graph-theoretical Terminology

We start this section by defining some terminology for chemical graphs and molecular matrices to be used in subsequent sections [11,12,47,49].

Let $G = (V, E)$ be a finite connected graph without loops and multiple edges. Here V is the set of elements called *vertices* and E is the set of unordered pairs of distinct elements of V, called *edges* [28]. G is a *molecular graph* representing the chemical structures. Again, we only consider the skeleton of the underlying graph, i.e., we do not take heteroatoms and bond types into account [26,54]. In view of the above definitions, we interchangeably use the terms structure and graph.

Let $deg(v)$ denote the number of edges incident with v, called the *degree of a vertex.* Let $d(u, v)$ denote the length of the shortest path connecting the vertices u and v, called as *distance between two vertices u and v*, and let $\rho(G) = \max\{d(u, v) : u, v \in V\}$ denote the *diameter* of G. Let $S_r(u; G)$ denote the *r-sphere* of a vertex u defined as $S_r(u; G) = \{x \in V : d(u, x) = r\}$.

3 Uniqueness of Descriptors

The uniqueness (or degree of degeneracy) is an important property of a descriptor and expresses its discrimination power [8, 34, 36]. In general, a descriptor is *degenerated* if there are at least two non-isomorphic graphs possessing the same value. However, it is well-known that every index is degenerated to some extent. Hence to quantify the degree of degeneracy of a given index, a non-information-theoretic measure [34, 35] and an information-theoretic measure [51] have been developed. To generate the numerical results, we use the sensitivity index $S(I)$ of a descriptor I due to Konstantinova [34]:

$$S(I) = \frac{|\mathcal{G}| - |\mathcal{G}_i|}{|\mathcal{G}|}, \tag{1}$$

where $|\mathcal{G}|$ denotes the cardinality of a given set of graphs $\mathcal{G}$ and $|\mathcal{G}_i|$ is the number of graphs $\mathcal{G}_i \in \mathcal{G}$ that cannot be distinguished by the descriptor I. It is evident that $S(I)$ relies on the given collection of graph structures. By definition $S(I) = 1$ implies that there does not exist any pair of non-isomorphic graphs possessing the same value of I. Hence when $S(I)$ is close to one, then I is said to be highly discriminative. Seminal work to evaluate the discrimination power of topological indices has been done by Bonchev et al. [8] when analyzing information-theoretic measures on alkane structures. This study has been then further extended by Raychaudhri et al. [46] to graphs containing one ring. Later Konstantinova et al. [34–36] studied the sensitivity of various measures using polycyclic structures such as hexagonal and square lattices that represent the class of cata-condensed benzenoid hydrocarbons. More recent results have been achieved by Diudea et al. [17] when examining a novel super index based on Shell-matrices and polynomials. By using real and synthetic molecular graphs, it turned out that this new descriptor can distinguish all graphs uniquely [17].

4 Molecular Descriptors

In this section, we define novel molecular descriptors derived by using the spectra of molecular matrices. Let $G = (V, E)$ be a molecular graph on n vertices. Let M be a molecular matrix defined on G and let $s > 0$. Let $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be the non-zero eigenvalues of M. We define

$$H_{M,s}(G) = -\sum_{i=1}^{k} \frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_j|^{\frac{1}{s}}} \log_2 \left(\frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_j|^{\frac{1}{s}}} \right), \tag{2}$$

$$S_{M,s}(G) = |\lambda_1|^{\frac{1}{s}} + |\lambda_2|^{\frac{1}{s}} + \cdots + |\lambda_k|^{\frac{1}{s}}, \tag{3}$$

$$IS_{M,s}(G) = \frac{1}{|\lambda_1|^{\frac{1}{s}} + |\lambda_2|^{\frac{1}{s}} + \cdots + |\lambda_k|^{\frac{1}{s}}}, \tag{4}$$

$$P_{M,s}(G) = |\lambda_1|^{\frac{1}{s}} \cdot |\lambda_2|^{\frac{1}{s}} \cdots |\lambda_k|^{\frac{1}{s}}, \tag{5}$$

$$IP_{M,s}(G) = \frac{1}{|\lambda_1|^{\frac{1}{s}} \cdot |\lambda_2|^{\frac{1}{s}} \cdots |\lambda_k|^{\frac{1}{s}}}. \tag{6}$$

Note that $H_{M,s}(G)$ defined by Equation (2) is based on Shannon's entropy [48]. Here, $H_{M,s}(G)$ represents the *mean information content* of G with respect to the distribution of the eigenvalues of the molecular matrix M. A comprehensive and almost up to date survey on information-theoretic indices for graphs has been recently published by Dehmer et al. [11].

The just defined families generalize some measures known in the literature. When $M = A(G)$ and $s = 1$, $S_{M,s}(G)$ is the well-known *graph energy* proposed by Gutman

[24,25]. Also, $S_{M,s}(G)$ (see Equation (3)) is a special case of the measure

$$SpSum^k(M,w) := \sum_{i=1}^{n} |\lambda_i|^k, \tag{7}$$

proposed in [9]. When $s = 2$, $S_{M,s}(G)$ represents a measure defined by Dehmer et al. [14] for analyzing the zeros of a special matrix (defined in the next section as $IM_1(G)$ in Equation (18)). In this paper, we refer to Equation (2) as an entropy-(based) measure and Equations (3) to (6) as algebraic measures.

5 Molecular Matrices

Numerous graph-theoretical matrices have been defined by using the structural properties of a molecular graph [32]. In this article, we mainly consider ten types of matrices defined by using the adjacency between vertices, the degree of a vertex and/or the distance between two vertices of a graph. Reasons why we have used these particular matrices are the reasonable computational complexity to derive the matrices from given molecular graphs and the fact that their underlying structural features have been extensively investigated and are well understood (interpretability). Now let $V = \{1, 2, \ldots, n\}$ be the vertex set of the graph G. Each of the matrices defined below is real and a square $n \times n$ matrix:

1. $A(G)$ is the adjacency matrix of G.

2. The Laplacian matrix, $L(G)$, is a symmetric matrix which is based on the degrees and adjacency relations [28]. For $1 \leq i, j \leq n$, the $(i,j)^{th}$ entry of $L(G)$ is given by,

$$[L]_{ij} := \begin{cases} -1, & \text{if vertices } i \text{ and } j \text{ are adjacent,} \\ deg(i), & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

3. The distance matrix, $D(G)$, is a symmetric matrix defined by using the distance between the vertices [28]:

$$[D]_{ij} := \begin{cases} d(i,j), & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

4. The distance path matrix, $DP(G)$, is a symmetric matrix defined by using the elements of the distance matrix [18,19]:

$$[DP]_{ij} := \binom{[D]_{ij} + 1}{2}. \tag{10}$$

Each element of this matrix counts all internal paths included in the shortest paths between the pair of vertices.

5. The augmented vertex degree matrix, $AV(G)$, is a non-symmetric matrix defined by using the degree and distance between the vertices [41, 42]:

$$[AV]_{ij} := deg(j)/2^{[D]_{ij}}. \tag{11}$$

6. The extended adjacency matrix, $EA(G)$, is a symmetric matrix defined by using the degree of the vertices [57]:

$$[EA]_{ij} := \begin{cases} \frac{1}{2}\left(\frac{deg(i)}{deg(j)} + \frac{deg(j)}{deg(i)}\right), & \text{if vertices } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

7. The vertex connectivity matrix, $VC(G)$, is a symmetric matrix defined by using the degrees of the vertices [44]:

$$[VC]_{ij} := \begin{cases} \frac{1}{\sqrt{deg(i)deg(j)}}, & \text{if vertices } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

8. The random walk Markov matrix, $MM(G)$, is a non-symmetric matrix defined by using the vertex degrees [33]. Firstly, it is assumed that every neighboring vertex can be reached from a given vertex u with the same probability, that is, the probability of going from u to one of its neighbors is $1/deg(u)$. The resulting Markov matrix is defined as follows:

$$[MM]_{ij} := \begin{cases} \frac{1}{deg(j)}, & \text{if vertices } i \text{ and } j \text{ are adjacent,} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The walks of the generated distribution are called *simple random walks*. Note that random walks can also be obtained by calculating powers of $MM(G)$. That is the entry at $[MM^\lambda]_{ij}$ represent the probability for a λ-step simple random walk starting at vertex j and ending at vertex i.

To describe two more graph-theoretical matrices, we need further definitions. Let $G = (V, E)$ be a graph with n vertices. We start by defining a probability distribution $P_G(V)$ on the vertices of a given graph using an arbitrary information functional that captures the structural information of a graph, see [11, 12]. Then, the quantities [12]

$$p_f(i) = \frac{f(i)}{\sum_{j=1}^{n} f(j)}, \tag{15}$$

form a probability distribution over the set of vertices $V = \{1, 2, \dots n\}$. Again $f : V \to R^+$ is an arbitrary information functional that maps a set of vertices to non-negative real numbers.

Now we can define the so-called *weighted structure function matrix* denoted by $IM(G)$:

$$[IM]_{ij} := 1 - |p_f(i) - p_f(j)|\beta(d(i,j)). \tag{16}$$

This matrix is symmetric and is based on the inferred probability distribution $P_G(V)$. β is a weighting function. Evidently, concrete information functionals and parameters lead to special matrices. By using the information functional

$$f(i) = \sum_{j=1}^{\rho(G)} c_j |S_j(i;G)|, \tag{17}$$

we yield the following matrices:

9. Weighted structure function matrix, $IM_1(G)$ [12, 14]:

$$[IM_1]_{ij} := 1 - \frac{|p_f(i) - p_f(j)|}{2^{d(i,j)}}, \tag{18}$$

   where $f(i) = \sum_{j=1}^{\rho(G)} (\rho(G) + 1 - j)|S_j(i;G)|$ and $\beta(d(i,j)) = \frac{1}{2^{d(i,j)}}$.

10. Weighted structure function matrix, $IM_2(G)$ [12, 14]:

$$[IM_2]_{ij} := 1 - \frac{|p_f(i) - p_f(j)|}{2^{d(i,j)}}, \tag{19}$$

   where $f(i) = \sum_{j=1}^{\rho(G)} (\rho(G)e^{-j+1})|S_j(i;G)|$ and $\beta(d(i,j)) = \frac{1}{2^{d(i,j)}}$.

More technical details and other information functionals can be found in [12].

6 Analysis and Numerical Results

In this section we exhibit the behavior of the newly defined descriptors (from Section 4) by applying them to various databases containing a large number of molecular structures. Concretely, we investigate what kind of structural information the measures do detect by determining correlations and cumulative distributions.

From the newly defined descriptors in Section 4, we arbitrarily choose eight descriptors by varying the value of s, namely $H_{M,1}(G)$, $H_{M,2}(G)$, $S_{M,2}(G)$, $S_{M,3}(G)$, $IS_{M,2}(G)$,

$P_{M,1}(G)$, $P_{M,2}(G)$ and $IP_{M,2}(G)$. Here M denotes the matrix that is being considered. For example, $H_{D,2}(G)$ denotes the entropy measure at $s = 2$ and the eigenvalues are derived using the distance matrix $D(G)$. Apart from these eight measures, we also consider two more statistical measures from the literature [14,37], namely the leading eigenvalue (also called the maximum) denoted by $TI^1_M(G)$ and the variance of the eigenvalues denoted by $TI^2_M(G)$.

6.1 Databases

We have considered both real and synthetic chemical structures. Note that all databases only contain the skeletons of the underlying chemical structures (all bond and atom types are equal). Chemical structures with isomorphic structures are represented only once. The synthetic graphs have been generated by using the software Molgen [1].

MS2265 This database has been extracted from the commercially available mass spectral database NIST [13,50]. It contains 2265 non-isomorphic chemical structures. Also $4 \leq |V| \leq 19$; $2 \leq \rho(G) \leq 15$ holds for all $G \in$ MS2265.

AG3981 The database has been generated from a freely available database called Ames Genetoxicity [13,27]. It contains 3981 non-isomorphic chemical structures. Also $2 \leq |V| \leq 109$ and $2 \leq \rho(G) \leq 47$ holds $\forall G \in$ AG3981.

APL91075 The database has been generated from a freely available database called ASINEX Platinum Collection [2,13] which contains in-house designed and synthesized collection of drug-like compounds. Also $6 \leq |V| \leq 60$ and $3 \leq \rho(G) \leq 36$ holds $\forall G \in$ APL91075.

C12 Trees This is a synthetic graph class consisting of 355 (exhaustive) alkane isomers with 12 carbon atoms.

C12 Ring1 This synthetic graph class consists of 3232 (exhaustive) hydrocarbon isomers with 12 carbon atoms such that each molecular structure contains one ring.

C12 Ring2 This synthetic graph class consists of 16977 (exhaustive) hydrocarbon isomers with 12 carbon atoms such that each molecular structure contains two rings.

C13 Ring2 This synthetic graph class consists of 51652 hydrocarbon isomers with 13 carbon atoms such that each molecular structure contains two rings.

C14 Trees This synthetic graph class consists of 1858 alkane isomers with 14 carbon atoms.

C14 Ring1 This synthetic graph class consists of 22565 hydrocarbon isomers with 14 carbon atoms such that each molecular structure contains one ring.

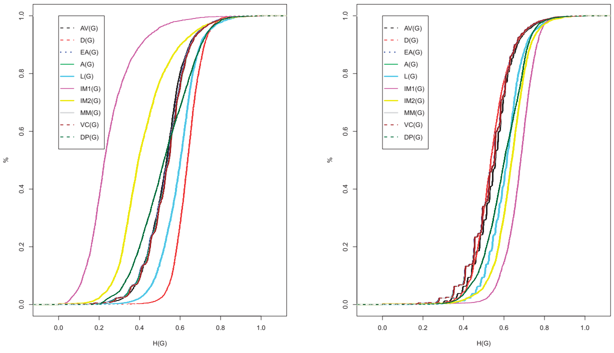
C15 Trees This is also a synthetic graph class consisting of 4347 alkane isomers with 15 carbon atoms.

6.2 Entropy Based Measures

6.2.1 Cumulative Distributions

In the following, we interpret the obtained results by using cumulative distributions of the measures. Also, we determine correlations between different measures. To interpret the cumulative distributions, consider Figure 1. They show the cumulative distributions of entropy measures $H_{M,1}(G)$ and $H_{M,2}(G)$ for AG3981. Here the x-axis represents the normalized entropy values and the y-axis represents the percentage rate of chemical structures having a normalized value less or equal TI. The measures were normalized according to the following scheme:

$$\overline{TI} = \frac{TI - \min(TI)}{\max(TI) - \min(TI)}. \tag{20}$$



(a) $H_{M,1}(G)$ (b) $H_{M,2}(G)$

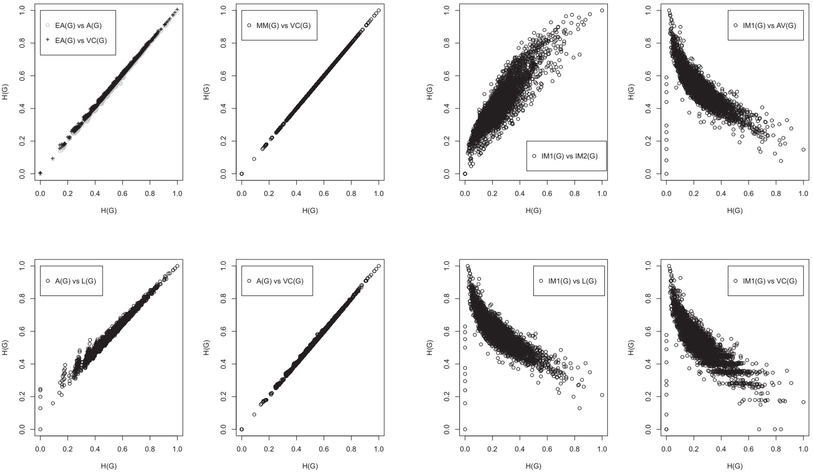
Figure 1: Cumulative distributions of entropy based measures for AG3981.

Indices		ρ when		Indices		ρ when	
		$s=1$	$s=2$			$s=1$	$s=2$
$H_{A,s}(G)$	$H_{AV,s}(G)$	0.987	0.9792	$H_{IM_1,s}(G)$	$H_{A,s}(G)$	-0.8405	0.3807
$H_{A,s}(G)$	$H_{EA,s}(G)$	0.9987	0.9999	$H_{IM_1,s}(G)$	$H_{AV,s}(G)$	-0.8442	0.358
$H_{A,s}(G)$	$H_{L,s}(G)$	0.9898	0.9806	$H_{IM_1,s}(G)$	$H_{EA,s}(G)$	-0.8317	0.3847
$H_{A,s}(G)$	$H_{VC,s}(G)$	0.9993	0.9999	$H_{IM_1,s}(G)$	$H_{L,s}(G)$	-0.8375	0.3751
$H_{A,s}(G)$	$H_{MM,s}(G)$	0.9993	0.9999	$H_{IM_1,s}(G)$	$H_{VC,s}(G)$	-0.8387	0.3831
$H_{AV,s}(G)$	$H_{EA,s}(G)$	0.9829	0.9723	$H_{IM_1,s}(G)$	$H_{D,s}(G)$	-0.3489	0.3435
$H_{AV,s}(G)$	$H_{L,s}(G)$	0.9955	0.9989	$H_{IM_1,s}(G)$	$H_{DP,s}(G)$	0.0903	0.5587
$H_{AV,s}(G)$	$H_{VC,s}(G)$	0.9855	0.979	$H_{IM_1,s}(G)$	$H_{IM_2,s}(G)$	0.8846	0.7159
$H_{AV,s}(G)$	$H_{MM,s}(G)$	0.9855	0.979	$H_{DP,s}(G)$	$H_{D,s}(G)$	0.4489	0.64241
$H_{EA,s}(G)$	$H_{L,s}(G)$	0.9864	0.9796	$H_{L,s}(G)$	$H_{VC,s}(G)$	0.9884	0.9804
$H_{EA,s}(G)$	$H_{VC,s}(G)$	0.9992	0.999	$H_{L,s}(G)$	$H_{MM,s}(G)$	0.9884	0.9804
$H_{EA,s}(G)$	$H_{MM,s}(G)$	0.9992	0.999	$H_{VC,s}(G)$	$H_{MM,s}(G)$	1	1

Table 1: Correlation coefficient ρ if $s=1$ and $s=2$ for $H_{M,s}(G)$ by using AG3981.

In Figure 1(a), we observe that the eigenvalues of the matrices $A(G)$ and $EA(G)$ possess almost identical distributions. In addition, the eigenvalues of $VC(G)$ and $MM(G)$ also have identical distribution and are almost identical with $A(G)$ and $EA(G)$. The eigenvalues of $AV(G)$ and $L(G)$ also possess a similar distribution to the aforementioned matrices. But this does not mean that these molecular matrices are per se useless. Indeed, the resulting eigenvalues can be used to define measures that capture structural information meaningfully, e.g., see [14].

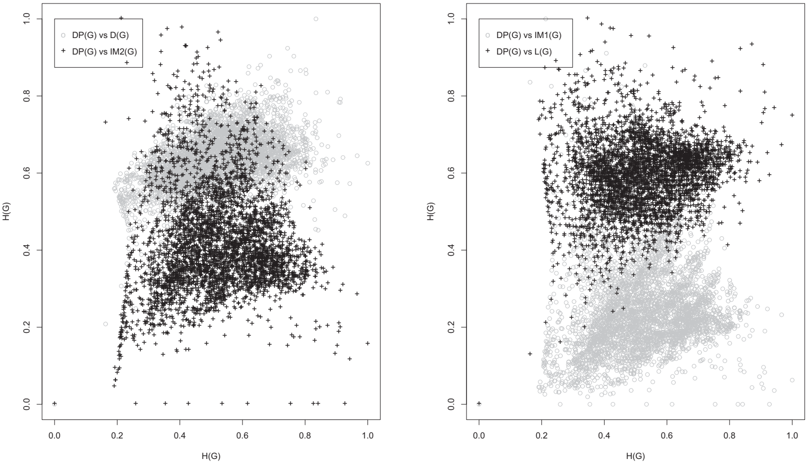
We also present the results when determining the correlations of the entropy measures $H_{M,s}$ between various matrices in Table 1. The resulting scatter plots are shown in Figure 2. In support of the above mentioned observation, we see again that the correlation between the measures $H_{M,s}(G)$ for every pair of the matrices, $A(G)$, $L(G)$, $AV(G)$, $EA(G)$, $VC(G)$ and $MM(G)$, is greater than 0.986. The correlation for descriptors using the matrices $VC(G)$ and $MM(G)$ equals one. Further, the values of the entropy measure based on the eigenvalues of $DP(G)$ are uniformly distributed where the values range between 0.2 to 0.8. We also observe that the correlation between $DP(G)$ and any matrix is around zero while the maximum correlation is around 0.44 with $D(G)$. That means, the corresponding measures capture structural information of the molecular graphs differently. The scatter plot of the entropy measure by using $DP(G)$ vs. other molecular matrices is shown in Figure 3. It is worth mentioning that the mean of these distributions range between 0.524 and 0.536. The distribution of the eigenvalues of the matrices $D(G)$ and $L(G)$ show that $80\% - 90\%$ of the underlying chemical structures possess large entropy values having means of 0.63 and 0.59, respectively. The eigenvalue distribution of $D(G)$



(a)

(b)

Figure 2: Scatter plot of $H_{M,1}(G)$ for different combinations of M.



(a)

(b)

Figure 3: Scatter plot of $H_{M,1}(G)$ for different combinations of M.

has a correlation coefficient ranging between 0.66 and 0.69 with the matrices described above. In contrast, $IM_1(G)$ and $IM_2(G)$ possess small entropy values for nearly 90% of the molecular graphs having means of 0.25 and 0.42, respectively. From Table 1, we also observe that the entropy measure $H_{M,1}$ shows negative correlation between $IM_1(G)$ and all the other measures except $IM_2(G)$.

Clearly, Figure 1(b) shows that $H_{M,2}(G)$ has a different impact on $D(G)$, $IM_1(G)$ and $IM_2(G)$ as compared to $H_{M,1}(G)$. This fact can also be underpinned by the correlation analysis, wherein the correlation between $IM(G)$(both $IM_1(G)$ and $IM_2(G)$) and other matrices is positive having correlation coefficients between 0.3 and 0.76. Finally, we gain that all the matrices used for calculating $H_{M,2}(G)$ have a quite similar pattern of distribution. Note that we obtain similar results by using MS2265 and APL91075 (not shown here).

6.2.2 Discrimination Power

	MS2265	AG3981	APL91075		MS2265	AG3981	APL91075
$H_{A,1}$	0.925828	0.962823	0.899292	$H_{A,2}$	0.925386	0.955288	0.8818666
$H_{D,1}$	0.993819	0.989952	0.809409	$H_{D,2}$	0.999117	0.997488	0.89714
$H_{L,1}$	0.985431	0.988696	0.886753	$H_{L,2}$	0.984106	0.985933	0.783003
$H_{IM_1,1}$	0.947461	0.944235	0.092704	$H_{IM_1,2}$	0.992936	0.989199	0.832698
$H_{IM_2,1}$	0.980574	0.966591	0.188614	$H_{IM_2,2}$	0.998234	0.993971	0.823431
$H_{AV,1}$	0.996468	0.991962	0.892133	$H_{AV,2}$	0.979691	0.9578	0.624167
$H_{EA,1}$	0.997351	0.992213	0.914093	$H_{EA,2}$	0.992936	0.990204	0.890684
$H_{MM,1}$	0.993819	0.990204	0.914455	$H_{MM,2}$	0.988521	0.987189	0.890716
$H_{VC,1}$	0.993819	0.990204	0.914455	$H_{VC,2}$	0.988521	0.987189	0.890716
$H_{DP,1}$	0.996468	0.997488	0.871776	$H_{DP,2}$	0.997351	0.996483	0.89401

Table 2: Value of sensitivity index for the descriptor $H_{M,s}(G)$ when $s = 1$ and $s = 2$, with various matrices for real chemical structures.

As a next step of our analysis, we study the discrimination power of the measures using our databases. At the outset, we analyze Table 2 showing the calculated sensitivity indices $(S(I))$ concerning the entropy measures $H_{M,1}(G)$ and $H_{M,2}(G)$ computed for MS2265, AG3981 and APL91075. In general, all measures using the defined matrices show a high sensitivity value for AG3981 and MS2265 and, hence, can discriminate more than 92% of the graphs uniquely. In particular, the measures defined by the matrices $DP(G)$ and $AV(G)$ (that are based on distances) can discriminate more than 99% of the graphs uniquely. The measures using the adjacency matrix and the weighted structure function matrices possess the lowest discrimination power. Also, the entropy measure $H_{M,s}(G)$

does possess a high discrimination power if $s = 1$ compared to the case $s = 2$, except for the distance and the weighted structure function matrices. In contrast, the matrices $D(G)$, $IM_1(G)$ and $IM_2(G)$ do have a better discrimination power if $s = 2$ (this also supports the observation from Figure 1).

Next we observe that the measures (based on the molecular matrices) possess relatively less discrimination power for APL91075. The measure $H_{M,1}(G)$ has the highest sensitivity value of approximately 0.915 when using the matrices based on degrees (such as $EA(G)$, $VC(G)$, $MM(G)$). The matrices defined using distances follow with around $0.80 - 0.88$ while the weighted structure function matrices show quite low sensitivity values (0.09 and 0.18). But when applying $H_{M,2}(G)$, the values of $S(I)$ range from $62\% - 89\%$ for all the matrices. However, it is worth mentioning that the weighted structure function matrices show a tremendous improvement of the discrimination power (83% as compared to ca. 18% for $H_{M,1}(G)$).

Next we note that the discrimination power of the measure based on the matrix $VC(G)$ is as good as when using $MM(G)$, for all the graph classes (as observed earlier by performing the correlation analysis; $\rho = 1$). However, there is a major difference between the matrices: $VC(G)$ is symmetric and, hence, has only real eigenvalues. $MM(G)$ is non-symmetric and its spectrum contains complex-valued eigenvalues too. Interestingly, when the absolute values of the eigenvalues are taken into consideration, the spectra of both matrices become identical and, hence, the entropy values are identical. This fact can also be observed from the Figure 1 since the cumulative entropy distributions for $MM(G)$ and $VC(G)$ are almost identical. In contrast to the above situation, even though the matrices $EA(G)$ and $A(G)$ show identical distribution of measure values, the matrix $EA(G)$, has higher discrimination power than $A(G)$, $VC(G)$ or $MM(G)$.

We yield another interesting finding if $s = 1$ for MS2265, see Table 2. By calculating the matrices $D(G)$ and $VC(G)$, we find equal sensitivity values (0.994), but the structures that are not distinguishable by them are completely different. This shows that the underlying measures captures structural information of the graphs significantly different.

In Tables 3 and 4, we evaluate the discrimination power of the entropy measures $H_{M,1}(G)$ and $H_{M,2}(G)$ for synthetic structures (see Section 6.1). For $H_{M,1}(G)$, the matrices when being ranked in terms of their resulting sensitivity values (based on the underlying measure) reveal that $DP(G) \geq EA(G) \geq D(G) \geq AV(G)$ for all synthetic (isomer)

	C12Ring1	C12Ring2	C13Ring2	C14Ring1	C12Trees	C14Trees	C15Trees
$H_{A,1}$	0.756807	0.774283	0.676643	0.601773	0.743662	0.710441	0.625489
$H_{D,1}$	0.987005	0.938093	0.826415	0.771638	1	0.992465	0.982977
$H_{L,1}$	0.808478	0.713789	0.548962	0.583204	0.983099	0.958019	0.943179
$H_{IM_1,1}$	0.933478	0.696236	0.339561	0.519211	0.983099	0.95479	0.904992
$H_{IM_2,1}$	0.95823	0.80986	0.510861	0.626679	0.994366	0.982239	0.933747
$H_{AV,1}$	0.98453	0.918949	0.768005	0.741724	0.994366	0.989236	0.980676
$H_{EA,1}$	0.990718	0.954232	0.870848	0.782717	1	0.989236	0.990798
$H_{MM,1}$	0.909963	0.917241	0.824634	0.728917	0.943662	0.951561	0.950311
$H_{VC,1}$	0.909963	0.917241	0.824634	0.728917	0.943662	0.951561	0.950311
$H_{DP,1}$	0.996287	0.977028	0.930342	0.79743	1	0.996771	0.992179

Table 3: Value of the sensitivity index for $H_{M,1}(G)$ by using various matrices and synthetic structures.

graph classes. Then, other matrices follow while the adjacency matrix $A(G)$ shows the lowest performance in most of the cases. By using $H_{M,2}(G)$, the ordering of the matrices changes to $DP(G) \geq IM_1(G) \geq IM_2(G) \geq D(G)$ followed by other matrices. The matrices $AV(G)$, $L(G)$ and $A(G)$ lead to low discrimination power among others. Even for isomers, we observe (as before) that the entropy measures have higher discrimination power if $s = 1$ than $s = 2$ for all matrices except the distance matrix and the weighted structure function matrices. In contrast, using the distance matrix $D(G)$, weighted structure function matrices $IM_1(G)$ and $IM_2(G)$ lead to a better discrimination power if $s = 2$. In particular, $IM_2(G)$ has higher discrimination power than $IM_1(G)$, in all the cases. A reason for this is the presence of exponential terms in the definition of the matrix $IM_2(G)$. Also, using $VC(G)$ and $MM(G)$ lead to similar results when applying the measures to synthetic chemical structures.

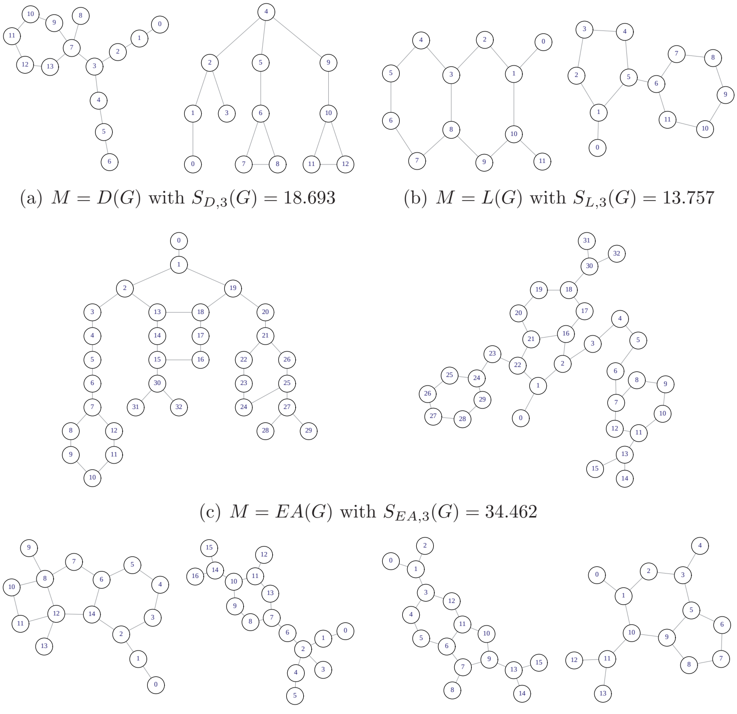
	C12Ring1	C12Ring2	C13Ring2	C14Ring1	C12Trees	C14Trees	C15Trees
$H_{A,2}$	0.741027	0.699535	0.516398	0.529803	0.743662	0.707750	0.609616
$H_{D,2}$	0.989480	0.94469	0.836676	0.756348	0.994366	0.989236	0.972625
$H_{L,2}$	0.744121	0.573187	0.281228	0.372568	0.926761	0.76211	0.541753
$H_{IM_1,2}$	0.993193	0.958827	0.871176	0.780545	1	0.992465	0.990798
$H_{IM_2,2}$	0.993812	0.963362	0.881166	0.782185	1	0.996771	0.991718
$H_{AV,2}$	0.865408	0.607646	0.278073	0.370796	0.963380	0.836921	0.690821
$H_{EA,2}$	0.97401	0.899158	0.736177	0.713982	1	0.981701	0.968484
$H_{MM,2}$	0.889542	0.853095	0.65709	0.648837	0.938028	0.946717	0.924776
$H_{VC,2}$	0.889542	0.853095	0.65709	0.648837	0.938028	0.946717	0.924776
$H_{DP,2}$	0.996287	0.965424	0.904166	0.78799	1	0.995695	0.990338

Table 4: Value of $S(I)$ for $H_{M,2}(G)$ using various matrices and synthetic structures.

From the sensitivity values shown in Tables 3 and 4, it is important to emphasize that among all synthetic graph classes, the alkane isomers (C12Trees, C14Trees and C15Trees) could be discriminated significantly as compared to the other hydrocarbon isomer classes.

6.3 Algebraic and Statistical Measures

Next, we analyze the algebraic measures $S_{M,2}(G)$ and $S_{M,3}(G)$ (see Equation (3)) . The results are shown by Tables 5, 6 and 7. Here, we only consider the symmetric matrices. From Tables 5 and 2, we observe that $S_{M,s}(G)$ has a better discrimination power than $H_{M,s}(G)$. Also, $S_{M,3}(G)$ discriminates better than $S_{M,2}(G)$. Again, the distance and degree- based matrices $DP(G)$, $D(G)$ and $EA(G)$ capture structural information more significantly and possess a higher discriminating power. By using AG3981, $S_{M,3}(G)$ has the same sensitivity values (0.9995) when applied to the matrices $D(G)$, $L(G)$, $IM_1(G)$, $IM_2(G)$ and $EA(G)$. But interestingly, the pairs of graphs that can not be discriminated by these matrices are notably different, see Figure 4.



(a) $M = D(G)$ with $S_{D,3}(G) = 18.693$

(b) $M = L(G)$ with $S_{L,3}(G) = 13.757$

(c) $M = EA(G)$ with $S_{EA,3}(G) = 34.462$

(d) $M = IM_1(G)$ with $S_{IM_1,3}(G) = 4.75$

(e) $M = IM_2(G)$ with $S_{IM_2,3}(G) = 5.313$

Figure 4: Pairs of graphs having same values using $S_{M,3}(G)$.

	MS2265	AG3981	APL91075		MS2265	AG3981	APL91075
$S_{A,2}$	0.928918	0.965084	0.94431	$S_{A,3}$	0.977925	0.990455	0.948251
$S_{D,2}$	1	1	0.980006	$S_{D,3}$	1	0.999498	0.996904
$S_{L,2}$	0.99117	0.996986	0.922701	$S_{L,3}$	0.998234	0.999498	0.877068
$S_{IM_1,2}$	0.997351	0.997991	0.961713	$S_{IM_1,3}$	0.999117	0.999498	0.964524
$S_{IM_2,2}$	0.999117	0.999498	0.966676	$S_{IM_2,3}$	0.996468	0.999498	0.97098
$S_{EA,2}$	1	1	0.996014	$S_{EA,3}$	1	0.999498	0.996114
$S_{VC,2}$	0.992936	0.994223	0.993741	$S_{VC,3}$	0.993819	0.994725	0.994049
$S_{DP,2}$	1	1	0.999231	$S_{DP,3}$	1	1	0.99765

Table 5: Value of $S(I)$ for $S_{M,s}(G)$ if $s = 2$ and $s = 3$ by using various matrices and real chemical structures.

	C12Ring1	C12Ring2	C13Ring2	C14Ring1	C12Trees	C14Trees	C15Trees
$S_{A,2}$	0.762686	0.820758	0.793638	0.644538	0.749296	0.717976	0.635611
$S_{D,2}$	0.999381	0.993698	0.984144	0.81755	1	1	0.9977
$S_{L,2}$	0.820235	0.789068	0.727213	0.655965	0.983099	0.974704	0.973315
$S_{IM_1,2}$	0.989171	0.949108	0.859328	0.782008	1	0.997847	0.991258
$S_{IM_2,2}$	0.993193	0.962066	0.895048	0.784445	1	0.998924	0.995399
$S_{EA,2}$	0.997525	0.989987	0.975683	0.816132	1	1	0.99954
$S_{VC,2}$	0.914913	0.954114	0.936169	0.774075	0.943662	0.960172	0.966644
$S_{DP,2}$	0.998762	0.996937	0.99245	0.820474	1	1	1

Table 6: Value of $S(I)$ for $S_{M,2}(G)$ using various matrices and synthetic molecular structures.

In Tables 8 and 9, we present the sensitivity values for $TI^1_M(G)$, $TI^2_M(G)$, $P_{M,1}(G)$, $P_{M,2}(G)$ and $IP_{M,2}(G)$. This table shows only the best sensitivity values in conjunction with the underlying molecular matrix, i.e., only those values which represent a competitively higher discriminating power among all values (and their corresponding matrices) when applying the underlying measures to our databases. Similarly, we present in Table 10 the sensitivity values for $IS_{M,2}(G)$ using selected symmetric matrices in the order of their discrimination power. From Table 8, we observe that $IP_{M,2}(G)$ is highly sensitive for the matrices $IM_1(G)$, $IM_2(G)$ and $L(G)$, and can discriminate almost every graph

	C12Ring1	C12Ring2	C13Ring2	C14Ring1	C12Trees	C14Trees	C15Trees
$S_{A,3}$	0.940594	0.924368	0.915918	0.768314	0.949296	0.944564	0.938348
$S_{D,3}$	0.998762	0.990104	0.976651	0.813915	1	0.998924	0.994019
$S_{L,3}$	0.943379	0.909525	0.814509	0.733658	0.988732	0.986007	0.968715
$S_{IM_1,3}$	0.998762	0.973847	0.921842	0.799778	1	0.998924	0.994479
$S_{IM_2,3}$	0.998144	0.978324	0.936518	0.806116	1	0.995694	0.995859
$S_{EA,3}$	0.998762	0.988455	0.977407	0.816131	1	1	0.996779
$S_{VC,3}$	0.913676	0.951051	0.929935	0.771460	0.943662	0.961249	0.966874
$S_{DP,3}$	0.999381	0.994817	0.982634	0.815644	1	0.996771	0.99908

Table 7: Value of $S(I)$ for $S_{M,3}(G)$ using various matrices and synthetic molecular structures.

	$IP_{IM_1,2}$	$IP_{IM_2,2}$	$IP_{L,2}$	$P_{DP,1}$	$P_{EA,1}$	$P_{DP,2}$	$P_{EA,2}$
MS2265	1	1	1	1	0.9766	0.999117	0.902429
AG3981	1	1	1	1	0.990706	0.999497	0.954032
APL91075	1	0.999495	0.989844	1	0.994697	1	0.939753
C12Ring1	1	1	1	1	0.950186	0.999381	0.769183
C12Ring2	1	1	1	0.999647	0.858043	0.995641	0.587265
C13Ring2	1	1	1	0.999884	0.855824	0.995663	0.560733
C14Ring1	0.822734	0.822734	0.822734	0.822734	0.76499	0.821183	0.559982
C12Trees	1	1	1	1	0.994366	1	0.95493
C14Trees	1	1	1	1	0.997847	0.998924	0.911195
C15Trees	1	1	1	1	0.994479	0.99954	0.89188

Table 8: Best values of $S(I)$ for algebraic descriptors $IP_{M,s}(G)$ and $P_{M,s}(G)$ using various matrices.

	TI^1_{DP}	TI^1_{D}	TI^1_{EA}	$TI^1_{IM_2}$	$TI^2_{IM_2}$	$TI^2_{IM_1}$	TI^2_{DP}
MS2265	1	1	0.978367	0.997351	0.998234	0.993819	0.873289
AG3981	1	1	0.937704	0.994725	0.997991	0.991962	0.942477
APL91075	0.99989	0.955334	0.606961	0.149108	0.278035	0.661685	0.979171
C12Ring1	1	0.995359	0.996287	0.939975	0.968441	0.916151	0.174505
C12Ring2	0.999647	0.996938	0.973553	0.716734	0.841256	0.658126	0.030394
C13Ring2	0.999342	0.994386	0.951173	0.366007	0.580307	0.309068	0.023136
C14Ring1	0.822424	0.820519	0.797252	0.551607	0.666652	0.466342	0.120629
C12Trees	1	1	0.994366	0.994366	0.994366	0.977465	0.890141
C14Trees	1	1	0.982239	0.961787	0.975242	0.927341	0.784715
C15Trees	1	0.99954	0.959282	0.921785	0.958822	0.848861	0.617207

Table 9: Best values of $S(I)$ for algebraic descriptors $TI^1_M(G)$ and $TI^2_M(G)$ using various matrices.

	$IS_{IM_2,2}$	$IS_{IM_1,2}$	$IS_{EA,2}$	$IS_{D,2}$	$IS_{DP,2}$	$IS_{L_1,2}$
MS2265	0.973068	0.969978	0.947461	0.927152	0.907726	0.89404
AG3981	0.961316	0.965838	0.919116	0.892238	0.868626	0.81889
APL91075	0.216887	0.243184	0.081032	0.053132	0.03843	0.061104
C12Ring1	0.856745	0.843441	0.789913	0.580446	0.612933	0.143874
C12Ring2	0.462862	0.408847	0.33298	0.1395432	0.126465	0.015904
C13Ring2	0.126926	0.110489	0.072408	0.025827	0.023639	0.003311
C14Ring1	0.294882	0.282296	0.167029	0.062752	0.064259	0.007401
C12Trees	0.960563	0.994366	0.960563	0.898592	0.938029	0.771831
C14Trees	0.914424	0.911195	0.850915	0.58127	0.699139	0.143703
C15Trees	0.776628	0.781459	0.651944	0.267541	0.413159	0.037037

Table 10: Best values of $S(I)$ for the algebraic descriptor $IS_{M,s}(G)$ using various matrices.

in all the databases except C14Ring1. However in terms of C14Ring1, only $IP_{IM_1,2}(G)$, $IP_{IM_2,2}(G)$ and $IP_{L,2}(G)$ have high discrimination power. We also discover that there exists a set of 4000 structures (out of 22565 structures) that can not be discriminated by these measures. Interestingly, we contemplate that this set of 4000 non-isomorphic structures appear in the set of non-distinguishable structures for all combinations of the measures and matrices.

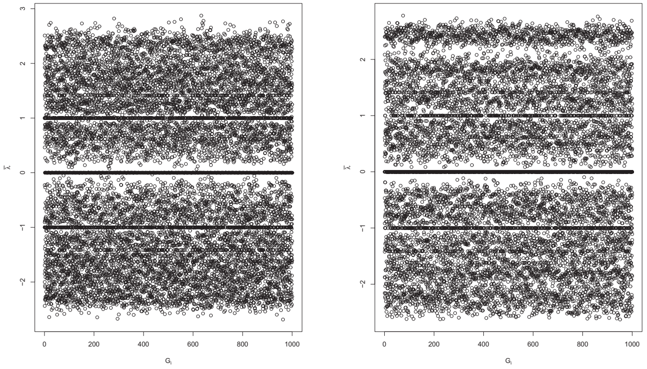
The measure $P_{M,s}(G)$ is sensitive by only applying it to matrices $DP(G)$ and $EA(G)$; see Table 8. When using other matrices, it shows very little discrimination power, less than 1%. This can be understood by the fact that the values of the eigenvalues of the underlying matrices are very small and that the value of the measure itself is close to zero. In general, $P_{M,1}(G)$ shows higher sensitivity when compared with $P_{M,2}(G)$.

From Table 10, we observe that $IS_{M,2}(G)$ shows an average discrimination power when compared with all other measures. In particular, this measure has low discrimination power when applied to synthetic structures.

$TI_M^1(G)$ shows a better result by using the distance-based matrices $DP(G)$ and $D(G)$, followed by the weighted structure function matrices $IM(G)$ and $EA(G)$; see Table 9. Further, this measure has a very little discrimination power by using $VC(G)$. For all the other matrices, it has a better discrimination power. This holds due to the fact that the leading eigenvalue for the matrix $VC(G)$ is always equal to one for every graph and, hence, the measure possesses the lowest discrimination power when compared to all other matrices.

$TI_M^2(G)$ is highly sensitive for $IM_2(G)$ followed by $IM_1(G)$, $DP(G)$ and $D(G)$, while the sensitivities when applying it to other matrices are close to zero; see Table 9. This can be explained by the fact that the eigenvalues are uniformly distributed for most of the matrices. In contrast to the matrices $IM_1(G)$, $IM_2(G)$, $DP(G)$ and $D(G)$, the eigenvalues are not uniformly distributed and there is a large zero-free region.

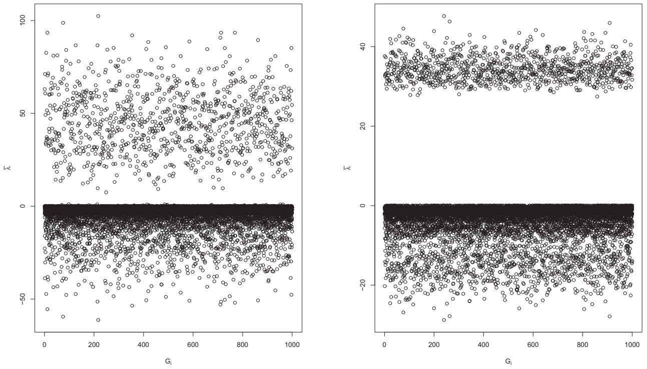
We underpin the just made statement by considering Figures 5, 6 and 7. For this, we randomly choose a set of 1000 graphs from MS2265 and C12Ring1 and plot the distribution of the eigenvalues for $IM_1(G)$, $D(G)$ and $A(G)$. Then, we observe that for the weighted structure function matrices shown in Figure 7, there is exactly one large eigenvalue equal to the number of atoms (vertices) in the structure and the remaining eigenvalues are closely distributed around zero. When considering the results using the adjacency matrix in Figure 5, we rediscover independent of the size of the graph, the eigenvalues are always less than three and are symmetrically distributed around zero. In contrast, the eigenvalue distribution of the distance matrix depicted by Figure 6 shows a large zero-free region above zero and negative eigenvalues occur more frequently.



(a) MS2265

(b) C12Ring1

Figure 5: Distribution of eigenvalues for the Adjacency Matrix $A(G)$

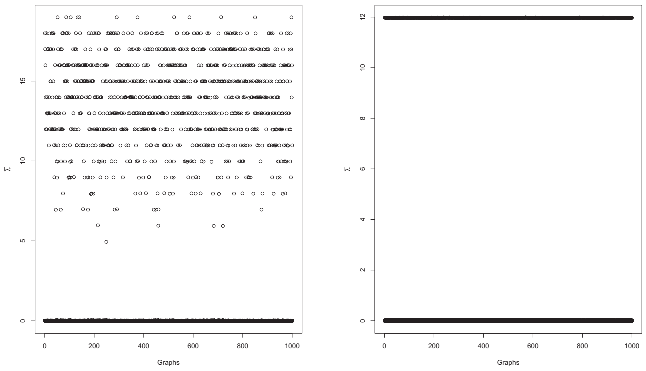


(a) MS2265

(b) C12Ring1

Figure 6: Distribution of eigenvalues for the Distance Matrix $D(G)$

A similar conclusion can be derived when plotting the distribution of the eigenvalues by using other matrices and databases. Studying the distribution of eigenvalues of molecular



(a) MS2265 (b) C12Ring1

Figure 7: Distribution of eigenvalues for the weighted structure function matrix $IM_1(G)$.

matrices is an interesting and novel problem itself and might lead to deeper insights when investigating or designing molecular structure descriptors.

7 Summary and Conclusion

In this paper, we analyzed the novel spectra-based molecular descriptors based on some molecular matrices. Despite the fact that eigenvalue-based measures have been investigated in structural chemistry and related fields, using the eigenvalues for discriminating chemical structures uniquely is a rather new problem. Except the recent work due to Dehmer et al. [14] - to our best knowledge - there is no large scale study to investigate spectra-based measures and particularly their discrimination power. Based on our findings, we conclude that to further study these descriptors might be promising and useful to tackle problems in structural chemistry, chemical theory and related fields.

Let's summarize our findings in brief:

- We considered ten types of measures and applied each to ten different molecular matrices. Then, we evaluated those by using the ten databases containing both real and synthetic chemical structures. Based on the fact that the computational complexity of all defined measures is polynomial, performing such a large scale study (e.g., see APL91075, C13Ring2 and C14Ring1) becomes feasible at all. Among all the measures we have considered in this study, $IP_{M,s}(G)$ gives quite promising results using $IM_1(G)$, $IM_2(G)$ and $L(G)$ (for all the databases). Equally, this also holds for $P_{M,s}(G)$ using $DP(G)$.

- We have found that matrices defined by using distances ($DP(G)$ and $D(G)$) are quite suitable to discriminate the graphs structurally. Interestingly, wherever the performance of $DP(G)$ is not satisfying, the weighted structure function matrices $IM_1(G)$ and $IM_2(G)$ lead to much better results. Again, this holds due to the different distribution of their eigenvalues. The correlation analysis of the descriptors turned out that some of the new measures are uncorrelated and, hence, good candidates to characterize chemical structures. Also, it might be worthwhile to examine the correlation between those and already existing descriptors. This might help when searching for new groups of descriptors.

- The usefulness of such highly discriminative descriptors is not only interesting to tackle problems in structural chemistry and related fields such as drug design and medical chemistry. Such structure descriptors could also be used to solve the still outstanding graph isomorphism problem. As known, there is no complete graph invariant but highly discriminative structure descriptors might help to tackle this problem. Also the efficiency of these measures, e.g., the ones proposed in this paper open a new door to various problems in quantitative graph theory.

- Note that we only considered skeletons of the underlying chemical structures to perform our study. In future, we will extend our mathematical apparatus to process weighted chemical graphs too. In particular, this relates to important chemical properties such as heteroatoms and different bond types.

Acknowledgement
Matthias Dehmer, Lavanya Sivakumar and Kurt Varmuza thank the Austrian Science Funds for supporting this work (project P22029-N13).

References

[1] Molgen isomer generator software, www.molgen.de, Institute of Mathematics II, University of Bayreuth, Germany, 2000.

[2] Asinex, ASINEX platinum collection. `http://www.asinex.com`, 2008.

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

[4] A. T. Balaban, Chemical graphs: Looking back and glimpsing ahead, *J. Chem. Inf. Comput. Sci.* **35** (1995) 339–350.

[5] A. T. Balaban, T. S. Balaban,New vertex invariants and topological indices of chemical graphs based on information on distances, *J. Math. Chem.* **8** (1991) 383–397.

[6] D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures*, Research Studies Press, Chichester, 1983.

[7] D. Bonchev, D. H. Rouvray, *Complexity in Chemistry, Biology, and Ecology*, Springer, New York, 2005.

[8] D. Bonchev, N. Trinajstić, Information theory, distance matrix, and molecular branching, *J. Chem. Phys.* **67** (1977) 4517–4533.

[9] V. Consonni, R. Todeschini, New spectral indices for molecule description, *MATCH Commun. Math. Comput. Chem.* **60** (2008) 3–14.

[10] D. M. Cvetković, M. Doob, H. Sachs, *Spectra of Graphs – Theory and Application*, Academic Press, New York, 1997.

[11] M. Dehmer, A. Mowshowitz, A history of graph entropy measures, *Inf. Sci.* **181** (2011) 57–78.

[12] M. Dehmer, Information processing in complex networks: graph entropy and information functionals, *Appl. Math. Comput.* **201** (2008) 82–94.

[13] M. Dehmer, N. Barbarini, K. Varmuza, A. Graber, A large scale analysis of information-theoretic network complexity measures using chemical structures, *PLoS ONE* **4** (2009) 1–13.

[14] M. Dehmer, L. A. J. Mueller, A. Graber, New polynomial-based molecular descriptors with low degeneracy, *PLoS ONE* **5** (2010) 1–6.

[15] J. Devillers, A. T. Balaban, *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon & Breach, Amsterdam, 1999.

[16] M. V. Diudea, *QSPR / QSAR Studies by Molecular Descriptors*, Nova, New York, 2001.

[17] M. V. Diudea, A. Ilić, K. Varmuza, M. Dehmer, Network analysis using a novel highly discriminating topological index, *Complexity*, in press.

[18] M. V. Diudea, Walk numbers ${}^{e}w_M$: Wiener-type numbers of higher rank, *J. Chem. Inf. Comput. Sci.* **36** (1996) 535–540.

[19] M. V. Diudea, Wiener and hyper-Wiener numbers in a single matrix, *J. Chem. Inf. Comput. Sci.* **36** (1996) 833–836.

[20] F. Emmert-Streib, M. Dehmer, *Analysis of Microarray Data: A Network–Based Approach*, Wiley–VCH, Weinheim, Germany, 2008.

[21] F. Emmert-Streib, M. Dehmer, Networks for systems biology: Conceptual connection of data and function, *IET Sys. Biol.*, accepted.

[22] E. Estrada, Characterization of the folding degree of proteins, *Bioinformatics* **18** (2002) 697–704.

[23] E. Estrada, Topological structural classes of complex networks, *Phys. Rev. E* **75** (2007) 0161031–01610312.

[24] I. Gutman, The energy of a graph, *Ber. Math. Stat. Sekt. Forschungszentrum Graz* **103** (1978) 1–22.

[25] I. Gutman, Polynomials in graph theory, in: D. Bonchev, D. H. Rouvray (Eds.), *Chemical Graph Theory: Introduction and Fundamentals*, Gordon & Breach, New York, 1991, pp. 133–176.

[26] I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer–Verlag, Berlin, 1986.

[27] K. Hansen, S. Mika, T. Schroeter, A. Sutter, A. Ter Laak, T. Steger-Hartmann, N. Heinrich, K. R. Müller, A benchmark data set for in silico prediction of ames mutagenicity, *J. Chem. Inf. Model.* **49** (2009) 2077–2081.

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

[29] O. Ivanciuc, T. Ivanciuc, A. T. Balaban, The graph description of chemical structures, in: J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPAR*, Gordon & Breach, Amsterdam, 1999, pp. 59–167.

[30] O. Ivanciuc, T. Ivanciuc, A. T. Balaban, Vertex- and edge-weighted molecular graphs and derived molecular descriptors, in: J. Devillers, A. T. Balaban (Eds.), *Topological Indices and Related Descriptors in QSAR and QSPAR*, Gordon & Breach Science, Amsterdam, 1999, pp. 169–220.

[31] O. Ivanciuc, T. Ivanciuc, M. V. Diudea, Polynomials and spectra of molecular graphs, *Rouman. Chem. Quart. Rev.* **7** (1999) 41–67.

[32] D. Janežič, A. Miličevič, S. Nikolić, N. Trinajstić, *Graph Theoretical Matrices in Chemistry*, Univ. Kragujevac, Kragujevac, 2007.

[33] D. J. Klein, J. L. Palacios, M. Randić, N. Trinajstić, Random walks and chemical graph theory, *J. Chem. Inf. Comput. Sci.* **44** (2004) 1521–1525.

[34] 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.

[35] E. V. Konstantinova, On some applications of information indices in chemical graph theory, in: R. Ahlswede, L. Bäumer, N. Cai, H. Aydinian, V. Blinovsky, C. Deppe, H. Mashurian (Eds.), *General Theory of Information Transfer and Combinatorics*, Springer– Verlag, Berlin, 2006, pp. 831–852.

[36] E. V. Konstantinova, A. A. Paleev, Sensitivity of topological indices of polycyclic graphs, *Vychisl. Sis.* **136** (1990) 38–48, (in Russian).

[37] L. Lovász, J. Pelikán, On the eigenvalues of trees, *Per. Math. Hung.* **3** (1973) 175–182.

[38] A. Mehler, P. Weiß, A. Lücking, A network model of interpersonal alignment, *Entropy* **12** (2010) 1440–1483.

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

[40] M. Randić, W. R. Müller, J. V. Knop, N. Trinajstić, The characteristic polynomial as a structure discriminator, *J. Chem. Inf. Comput. Sci.* **37** (1997) 1072–1077.

[41] M. Randić, D. Plavšić, Characterization of molecular complexity, *Int. J. Quant. Chem.* **91** (2002) 20–31.

[42] M. Randić, D. Plavšić, On the concept of molecular complexity, *Croat. Chem. Acta* **75** (2002) 107–116.

[43] M. Randić, M. Vračko, M. Novič, Eigenvalues as molecular descriptors, in M. V. Diudea (Ed.), *QSPR / QSAR Studies by Molecular Descriptors*, Nova, New York, 2001, pp. 93–120.

[44] M. Randić, Similarity based on extended basis descriptors, *J. Chem. Inf. Comput. Sci.* **32** (1992) 686–692.

[45] M. Randić, M. Vračko, On the similarity of dna primary sequences, *J. Chem. Inf. Comput Sci.* **40** (2000) 599–606.

[46] C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, S. C. Basak, Discrimination of isomeric structures using information theoretic topological indices, *J. Comput. Chem.* **5** (1984) 581–588.

[47] C. E. Shannon, A mathematical theory of communication, *Bell Sys. Tech. J.* **27** (1948) 379–423 and 623–656.

[48] C. E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, Univ. Illinois, Urbana, 1997.

[49] V. A. Skorobogatov, A. A. Dobrynin, Metrical analysis of graphs, *MATCH Commun. Math. Comput. Chem.* **23** (1988) 105–155.

[50] S. E Stein, NIST, Mass spectral database 98, `www.nist.gov/srd/nist1a.htm`, National Institute of Standards and Technology, Gaithersburg, MD, USA, 1998.

[51] R. Todeschini, R. Cazar, E. Collina, The chemical meaning of topological indices, *Chemom. Intell. Lab. Sys.* **15** (1992) 51–59.

[52] R. Todeschini, V. Consonni, R. Mannhold, *Handbook of Molecular Descriptors*, Wiley–VCH, Weinheim, 2002.

[53] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics: Alphabetical Listing*, Wiley-VCH, Weinheim, 2009.

[54] N. Trinajstić, *Chemical Graph Theory*, CRC Press, Boca Raton, 1992.

[55] R. E. Ulanowicz, Quantitative methods for ecological network analysis, *Comput. Biol. Chem.* **28** (2004) 321–339.

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

[57] Y. Q. Yang, L. Xu, C. Y. Hu, Extended adjacency matrix indices and their applications, *J. Chem. Inf. Comput. Sci.* **34** (1994) 1140–1145.