Handbook of Chemoinformatics
Wiley–VCH
2003

Topological Indices

Ovidiu Ivanciuc

Sealy Center for Structural Biology and Molecular Biophysics
Department of Biochemistry and Molecular Biology
University of Texas Medical Branch
301 University Boulevard
Galveston, Texas 77555-0857, USA
Email: ovidiu_ivanciuc@yahoo.com
Email: iejmd@yahoo.com
URL: http://ivanciuc.org/
URL: http://biochempress.com/

O. Ivanciuc, Topological Indices. In: *Handbook of Chemoinformatics*, Ed.: J. Gasteiger, Wiley–VCH, 2003, pp. 981–1003.

Biographical notes for Ovidiu Ivanciuc are given at the top of Chapter II, Section 4.

1
Topological Indices

Ovidiu Ivanciuc

1.1
Introduction

The structure of organic compounds is represented in a numerical form with a large variety of structural descriptors. Among these, those computed from the molecular graph are widely used in modeling physical, chemical, or biological properties. By removing all hydrogen atoms from the structure diagram of a compound containing covalent bonds one obtains the hydrogen-depleted (or hydrogen-suppressed) molecular graph of that compound, whose vertices correspond to non-hydrogen atoms and whose edges correspond to covalent bonds [1–5]. These chemical graphs can be represented in algebraic form as molecular matrixes. This numerical description of the structure of chemical compounds is the basis for the computation of various polynomials, spectra, spectral moments, other structural descriptors, and topological indices. Topological indices (TI) are a convenient method for expressing in a numerical form the chemical structure encoded in molecular graphs. The topological description of a molecule contains information on the atom–atom connectivity in the molecule, and encodes the size, shape, branching, heteroatoms and the presence of multiple bonds [6–13]. This graph description of molecules neglects information on bond lengths, bond angles, and torsion angles, but is able to encode in a numerical form the important atom connectivity information that determine a wide range of physical, chemical, and biological properties. Topological indices are widely used as structural descriptors in quantitative structure–property relationships (QSPR) and quantitative structure–activity relationships (QSAR) models. Recent applications of structural descriptors derived from the molecular graph have been made in the design of chemical libraries, in virtual screening of combinatorial libraries, and in large-scale evaluation of the molecular similarity and diversity of chemical databases. In the above chemical database generation and mining applications the molecular structure is translated into a numerical form with the aid of various structural descriptors derived from the molecular graph, many of them traditionally used in QSPR and QSAR. To be efficient, the in silico compound screening uses descriptors that require small computational resources, such as counts of atom types, counts of

functional groups, fingerprints, constitutional descriptors, graph invariants and topological indices. In this section we present an overview of the main topological indices used in QSPR, QSAR, and virtual screening of chemical libraries. The theory of topological indices uses notions of graph theory, molecular matrixes, vertex- and edge-weighted (VEW) molecular graphs [14, 15] that represent molecules containing heteroatoms and double bonds, all presented in the section on graph theory (Chapter II, Section 4).

1.2 Topological Indices

1.2.1 The Wiener Index and Related Topological Indices

The Wiener index W, defined in 1947, is widely used in QSPR and QSAR models, and it still represents an important source of inspiration for defining new topological indices. Initially, its definition was not formulated with graph concepts [16]: ''The path number W is defined as the sum of the distances between any two carbon atoms in the molecule, in terms of carbon–carbon bonds. Brief method of calculation: Multiply the number of carbon atoms on one side of any bond by those on the other side; W is the sum of those values for all bonds''. This definition can be applied only to acyclic compounds, and Wiener used it to compute W only for alkanes. We present a translation of this definition into graph terms. Consider an acyclic graph G and denote with N_i and N_j the number of vertices situated on both sides of the edge e_{ij}. The vertex v_i is added to N_i while vertex v_j is added to N_j. For acyclic graphs the Wiener index $W(G)$ of a graph G is (Eq. (1)):

$$W(G) = \sum_{e_{ij} \in E(G)} N_i N_j \tag{1}$$

where the summation goes over all edges from the edge set $E(G)$, $e_{ij} \in E(G)$.

Hosoya extended the application of the Wiener index by defining it from the distance matrix as ''the half sum of the off-diagonal elements of a distance matrix **D** whose element d_{ij} is the number of bonds for the shortest path between atoms i and j'' [17] (Eq. (2)):

$$W(G) = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} [\mathbf{D}(G)]_{ij} \tag{2}$$

Because the formula proposed by Hosoya in Eq. (2) does not consider vertex- and edge-weighted molecular graphs, the current definition of the Wiener index uses a slightly modified equation [15] (Eq. (3)):

$$W(w, G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{D}(w, G)]_{ij} \tag{3}$$

where the graph G has N vertices and the distance matrix is computed with the weighting scheme w.

The success of the Wiener index in QSPR and QSAR models encouraged the development of several related topological indices. The quasi-Wiener index $W^* = W^*(G)$ of a graph G with N vertices is computed from the spectrum of the Laplacian matrix $\mathbf{L}(G)$ [18] (Eq. (4)):

$$W^*(G) = N \sum_{i=1}^{N-1} \frac{1}{\mathbf{Sp}(\mathbf{L}, G)_i} \tag{4}$$

where $\mathbf{Sp}(\mathbf{L})_i,\ i = 1, 2, 3, \ldots, N-1$ denote the positive eigenvalues of the Laplacian matrix (in the Laplacian matrix $\mathbf{Sp}(\mathbf{L})_N = 0$). For acyclic molecular graphs (trees), W^* is identical with the Wiener index W, $W^*(G) = W(G)$, while for cycle-containing graphs the two indices are different.

The resistance distance matrix $\mathbf{\Omega}$ was used to define a Wiener-type topological index, the Kirchhoff index $Kf = Kf(G)$ [19] (Eq. (5)):

$$Kf(G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{\Omega}(G)]_{ij} \tag{5}$$

For acyclic graphs the Wiener index W and the Kirchhoff index Kf coincide, but their values are different for cyclic graphs. Gutman and Mohar demonstrated that for any graph $W^*(G) = Kf(G)$ [20].

The reciprocal distance matrix **RD** is the source of another Wiener-like topological index, $RDSUM$ [21] (Eq. (6)):

$$RDSUM(G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{RD}(G)]_{ij} \tag{6}$$

A TI related to W is the hyper-Wiener index WW [22, 23] (Eq. (7)):

$$WW(G) = \frac{1}{2} \sum_{i<j} ([\mathbf{D}(G)]_{ij}^{\ 2} + [\mathbf{D}(G)]_{ij}) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{D}_\mathbf{p}(G)]_{ij} \tag{7}$$

where $\mathbf{D}_\mathbf{p}$ is the distance-path matrix defined by Diudea.

1.2.2 The Szeged Index

Gutman gave the following definition for the Szeged index of a simple, non-weighted molecular graph. Let e_{ij} be an edge of the molecular graph G, connecting the vertices v_i and v_j from G, $v_i, v_j \in V(G)$ and denote with td_{ij} the topological distance between vertices v_i and v_j representing the minimum number of bonds

between vertices v_i and v_j. Let n_i be the number of vertices v_k of the molecular graph G, having the property $td_{ki} < td_{kj}$ and let n_j be the number of vertices v_k of the molecular graph G, having the property $td_{kj} < td_{ki}$. When a vertex v_k is situated at the same topological distance from vertices v_i and v_j, i.e. $td_{ki} = td_{kj}$, the vertex is not counted neither in n_i nor in n_j. For the two vertices that form the edge e_{ij}, n_i gives the number of vertices closer to vertex v_i and n_j gives the number of vertices closer to vertex v_j. A formal definition of n_i and n_j is offered below (Eqs. (8) and (9)):

$$n_i = |\{v_k : v_i, v_j, v_k \in V(G), e_{ij} \in E(G), td_{ki} < td_{kj}\}| \tag{8}$$

$$n_j = |\{v_k : v_i, v_j, v_k \in V(G), e_{ij} \in E(G), td_{kj} < td_{ki}\}| \tag{9}$$

The Szeged index of the molecular graph G is (Eq. (10)):

$$Sz(G) = \sum_{e_{ij} \in E(G)} n_i n_j \tag{10}$$

where the summation goes over all edges e_{ij} from the edge set $E(G)$, $e_{ij} \in E(G)$. In acyclic graphs $N_i = n_i$ and $N_j = n_j$ and the Wiener and Szeged indices coincide.

Wiener-type topological indices can be computed from any symmetric molecular matrix, using formulas similar to Eq. (3). Such topological indices were derived from a large number of molecular matrixes: edge Szeged $\mathbf{Sz_e}$ [26], detour Δ [27], distance-valency **Dval** [28], electrical conductance **EC** [29], distance complement **DC** [30], complementary distance **CD** [31], and reverse Wiener **RW** [32] matrixes.

1.2.3 The Connectivity Indices

The highly successful Randić connectivity index χ [33] was extended by Kier and Hall for connected subgraphs [6, 9] (Eq. (11)):

$${}^{m}\chi_t^{\mathrm{v}} = \sum_{j=1}^{s} \prod_{i=1}^{n} (\delta_i^{\mathrm{v}})^{-1/2} \tag{11}$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and δ_j^{v} is the valence atomic connectivity computed with the formula (Eq. (12)):

$$\delta_i^{\mathrm{v}} = \frac{Z_i^{\mathrm{v}} - H_i}{Z_i - Z_i^{\mathrm{v}} - 1} \tag{12}$$

where Z_i^{v} is the number of valence electrons of atom i, Z_i is the count of all electrons of atom i, and H_i is the number of hydrogen atoms attached to this atom.

1.2.4 The Electrotopological State

The electrotopological state indices are local (atomic) invariants that encode, for each atom in a molecule, information about the topological environment and the electronic interactions due to all other atoms in the molecule. Each vertex v_i from a graph G with N vertices, representing the atom i from the corresponding organic compound, is assigned an intrinsic state value I_i [11] (Eq. (13)):

$$I_i = \frac{(2/Q_i)^2(Z_i^{\mathrm{v}} - H_i) + 1}{\mathbf{deg}_i} \tag{13}$$

where Q_i is the principal quantum number for the valence shell of atom i, Z_i^{v} is the number of valence electrons of atom i, H_i is the number of hydrogen atoms attached to this atom, and $\mathbf{deg}_i$ is the degree of atom i.

The second contribution to the electrotopological state index comes from the interactions between an atom i and all other atoms in the molecular graph. The perturbation on the intrinsic state value I of atom i, due to the presence of the remaining atoms in the molecule, is a function of the difference between the corresponding intrinsic state values and decreases when the interatomic distance increases [11] (Eq. (14)):

$$\Delta I_i = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{I_i - I_j}{(td_{ij} + 1)^2} \tag{14}$$

where td_{ij} is the topological distance between atoms i and j, equal to the minimum topological length of the paths connecting the two atoms, i.e. the minimum number of bonds between atoms i and j. The electrotopological state index S of atom i is an atomic invariant [11] (Eq. (15)):

$$S_i = I_i + \Delta I_i \tag{15}$$

The atomic indices S corresponding to an atom type j are summed together to give $S(j)$. The atom type indices $S(j)$ are used as structural descriptors in QSPR and QSAR models.

1.2.5 The Hosoya Index

The Hosoya index $Z = Z(G)$ of a graph G with N vertices is given by [17] (Eq. (16)):

$$Z(G) = \sum_{k=0}^{L} m(G,k) = \sum_{k=0}^{L} |a_{2k}| \tag{16}$$

where $m(G,k)$ is the number of k-matchings of G, i.e. the number of selections of k mutually non-adjacent edges in G, $L = [N/2]$ is the smallest integer not exceeding $N/2$, and a_k is the kth coefficient of the acyclic (matching) polynomial.

1.2.6
The Balaban Index *J*

The average distance sum connectivity index J of the molecular graph G is defined by the formula [34–36] (Eq. (17)):

$$J = \frac{M}{\mu + 1} \sum_{E(G)} (s_i \mathbf{DS}_i s_j \mathbf{DS}_j)^{-1/2} \tag{17}$$

where $\mathbf{DS}_i$ and $\mathbf{DS}_j$ denote the distance sums of the vertices v_i and v_j of an edge e_{ij} in the molecular graph G, M is the number of edges in the molecular graph, μ is the cyclomatic number, s_i is the weight of the vertex v_i, s_j is the weight of the vertex v_j, and the summation goes over all edges in the molecular graph, $E(G)$. The vertex weights are computed from the electronegativity and covalent radii, respectively, of the corresponding atom [36]. The distance sum of the vertex v_i, $\mathbf{DS}_i$, is defined as the sum of the topological distances between the vertex v_i and every vertex in the molecular graph, i.e. the sum over row i or column i in the $\mathbf{D}$ matrix (Eq. (18)):

$$\mathbf{DS}_i = \sum_{j=1}^{N} [\mathbf{D}(G)]_{ij} = \sum_{j=1}^{N} [\mathbf{D}(G)]_{ji} \tag{18}$$

The distance matrix used to compute the Balaban index J is obtained from an edge-weighted adjacency matrix, which does not consider vertex weights for heteroatoms. In this adjacency matrix the single, double, triple, and aromatic bonds have the weights 1, 1/2, 1/3, and 1/1.5, respectively.

1.2.7
The Information-Theory Indices I_D^E, $\bar{I}_D^E$, I_D^W, and $\bar{I}_D^W$

Information theory provides a simple quantitative measure of the information content of a system. The information content for various partitions of the molecular graph was numerically characterized with various equations [8]. The information-theory indices I_D^E and $\bar{I}_D^E$, representing the total and mean information on distances in a molecular graph G with N vertices, are [37] (Eqs. (19) and (20)):

$$I_D^E(G) = \frac{N(N-1)}{2} \log_2 \frac{N(N-1)}{2} - \sum_{k=1}^{l} d(G,k) \log_2 d(G,k) \tag{19}$$

$$\bar{I}_D^E(G) = \frac{2I_D^E}{N(N-1)} = -\sum_{k=1}^{l} \frac{2d(G,k)}{N(N-1)} \log_2 \frac{2d(G,k)}{N(N-1)} \tag{20}$$

where $d(G,k)$ represents the number of pairs of vertices in G that are separated by k edges, and l is the largest element of the distance matrix **D**, or the diameter of G. The information-theory indices I_D^W and $\bar{I}_D^W$, representing the total and mean information on the distribution of the distances in a molecular graph G, are [37] (Eqs. (21) and (22)):

$$I_D^W(G) = W \log_2 W - \sum_{k=1}^{l} d(G,k) k \log_2 k \tag{21}$$

$$\bar{I}_D^W(G) = \frac{I_D^W}{W} = -\sum_{k=1}^{l} d(G,k) \frac{k}{W} \log_2 \frac{k}{W} \tag{22}$$

where W represents the Wiener index of the graph G, and l is the diameter of G.

1.2.8
The Information on Distances Indices *U*, *V*, *X*, and *Y*

By transforming the distances of a vertex into a local structural descriptor, Balaban proposed new vertex invariants based on distance vectors and information theory [38, 39]. The operation of global distance summation used in the computation of the J index is replaced by a more refined approach, leading to local and global indices with lower degeneracy. On applying Shannon's formula to the elements of the distance matrix **D** that correspond to a vertex v_i one obtains the mean local information on the magnitude of distances (Eq. (23)):

$$u.inf_i = -\sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{[\mathbf{D}]_{ij}}{\mathbf{DS}_i} \log_2 \frac{[\mathbf{D}]_{ij}}{\mathbf{DS}_i} \tag{23}$$

and the local information on the magnitude of distances (Eq. (24)):

$$v.inf_i = \mathbf{DS}_i \log_2 \mathbf{DS}_i - u.inf_i \tag{24}$$

where $\mathbf{DS}_i$ represents the distance sum of the vertex v_i.

Two related vertex invariants were proposed, namely the extended local information on distance magnitude (Eq. (25)):

$$x.inf_i = \mathbf{DS}_i \log_2 \mathbf{DS}_i - y.inf_i \tag{25}$$

and the mean extended local information on distance magnitude (Eq. (26)):

$$y.inf_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} [\mathbf{D}]_{ij} \log_2 [\mathbf{D}]_{ij} \tag{26}$$

In Eqs (23) and (26) the summation is done for all non-zero elements in the i-th row of the distance matrix **D**. By analogy with the Randić connectivity index four new topological indices were defined [38, 39] on the basis of the four local graph invariants $u.inf_i$, $v.inf_i$, $x.inf_i$, and $y.inf_i$: (Eqs. (27)–(30)):

$$U(G) = \frac{M}{\mu + 1} \sum_{E(G)} (u.inf_i u.inf_j)^{-1/2} \tag{27}$$

$$V(G) = \frac{M}{\mu + 1} \sum_{E(G)} (v.inf_i v.inf_j)^{-1/2} \tag{28}$$

$$X(G) = \frac{M}{\mu + 1} \sum_{E(G)} (x.inf_i x.inf_j)^{-1/2} \tag{29}$$

$$Y(G) = \frac{M}{\mu + 1} \sum_{E(G)} (y.inf_i y.inf_j)^{-1/2} \tag{30}$$

where M is the number of edges in G, μ is the cyclomatic number of G, and the summation goes over all edges from the edge set $E(G)$.

1.2.9 Triplet Topological Indices T.P.R

A new class of vertex structural descriptors was defined as the solution of the following system of equations [40] (Eq. (31)):

$$\mathrm{Q} \cdot \mathrm{S} = \mathrm{R} \tag{31}$$

where **Q** is a matrix derived from a molecular graph matrix, **R** is a column vector, and **S** is the column vector of LOcal Vertex Invariants (LOVIs). The matrix **Q** is obtained from a graph topological matrix **T**, by replacing its diagonal elements $[\mathrm{T}]_{ii}$ with the components $\mathbf{P}_i$ of a nonzero column vector **P** representing a vertex property. The vertex property encoded in the column vectors **P** and **R** can be either topological (the vertex degrees, distance sum, or reciprocal distance sum) or chemical (the atomic number Z, the atomic mass A, electronegativity, polarizability, ionization potential). The system of equations and the type of LOVIs obtained is denoted by a three-element notation **T.P.R**, where **T** denotes the molecular matrix, **P** is the vertex property that replaces the main diagonal of the matrix **T**, and **R** is the descriptor used as free term. For example, the LOVIs obtained using the adjacency matrix **A**, the electronegativity E as the property **P**, and the distance sum **DS** as property **R**, is denoted by **A.E.DS**. Analytical formulas for the **T.P.R** invariants for certain classes of graphs were obtained [41]. The solution of the **T.P.R** system of equations is a vector of vertex descriptors that are combined together into a molecular descriptor. The most simple operation is the sum of the vertex invariants, denoted **Sum**(**T.P.R**).

1.3
Computing Topological Indices with Graph Operators

An inspection of the large number of topological indices defined in the literature shows that many of them are computed with identical mathematical equations, by using different molecular matrices. For example, two recent reviews [42, 43] dedicated to the topological indices related to the Wiener index show that all Wiener-type indices are computed with an identical formula, Eq. (3), applied to different matrices: the distance matrix **D** gives the Wiener index *W*, the reciprocal distance matrix **RD** gives the *RDSUM* index, the distance-path matrix $\mathbf{D_p}$ gives the hyper-Wiener index *WW*, the resistance distance matrix $\mathbf{\Omega}$ gives the Kirchhoff index *Kf*, the edge Szeged matrix $\mathbf{Sz_e}$ gives the Szeged index *Sz*, and the path matrix Szeged $\mathbf{Sz_p}$ gives the hyper-Szeged index. The name of all these indices does not give any indication about their common mathematical formula. Also, each of the above topological indices can be computed with several vertex and edge weighting schemes, but there is no clear way of indicating this in the symbol of the indices. The development of new molecular matrices can multiply the number of topological indices that can be computed with a mathematical formula identical with that used for the Wiener index. The problem of denoting in a simple and unique way such related structural descriptors can be solved by using operators representing a certain mathematical operation or algorithm. An operator defines the mathematical operations performed on a certain molecular matrix or on a graph invariant, together with various parameters and the weighting scheme used to compute a topological index [43, 44]. Therefore, a graph operator is as a simple and flexible notation system that collects together a family of topological indices that are computed with an identical mathematical formula. In this section we review the most important graph operators.

1.3.1
The Vertex Sum Operator

Many topological indices and graph descriptors are defined with the aid of vertex invariants. An operator that computes vertex invariants from molecular matrices is the vertex sum **VS**. Consider the vertex v_i from the graph *G* with *N* vertices and the symmetric graph matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$, where *w* is the weighting scheme used to compute the vertex and edge parameters. The vertex sum of the vertex v_i, $\mathbf{VS}(\mathbf{M}, w)_I = \mathbf{VS}(\mathbf{M}, w, G)_i$, is defined as the sum of the elements in the column *i*, or row *i*, of the molecular matrix **M** (Eq. (32)):

$$\mathbf{VS}(\mathbf{M}, w, G)_i = \sum_{j=1}^{N} [\mathbf{M}(w, G)]_{ij} = \sum_{j=1}^{N} [\mathbf{M}(w, G)]_{ji} \tag{32}$$

The **VS** operator is identical to the degree vector **Deg** if **M** is the adjacency matrix **A**, to the distance sum **DS** if **M** is the distance matrix **D**, and to the reciprocal dis-

tance sum **RDS** if **M** is the reciprocal distance matrix **RD**. This vertex invariant was used to define the Ivanciuc–Balaban operator [45].

1.3.2 The Connectivity Chi Operator

The **Chi** operator was derived from the Kier and Hall connectivity indices [6, 9] by replacing the local invariant δ^V with any other vertex invariant. Consider a vertex structural descriptor $\mathbf{VSD}(\mathbf{M}, w) = \mathbf{VSD}(\mathbf{M}, w, G)$ that assigns a numerical invariant $\mathbf{VSD}(\mathbf{M}, w)_i$ to each vertex v_i from the VEW molecular graph G. The connectivity Chi operator $\mathbf{Chi}(\mathbf{VSD}, \mathbf{M}, w) = \mathbf{Chi}(\mathbf{VSD}, \mathbf{M}, w, G)$ of the graph G is [44] (Eq. (33)):

$$^{m}\mathbf{Chi}(\mathbf{VSD}, \mathbf{M}, w, G)_t = \sum_{i=1}^{s} \prod_{j=1}^{n} (\mathbf{VSD}(\mathbf{M}, w, G)_j)^{-1/2} \tag{33}$$

where s is the number of connected subgraphs of type t with m edges, n is the number of vertices of the subgraph, and w is the weighting scheme. For hydrocarbons and when **VSD** is the degree, the **Chi** operator gives the Kier and Hall connectivity indices.

1.3.3 The Wiener Operator

Consider the VEW molecular graph G with N vertices and its symmetric molecular matrix $\mathrm{M}(w) = \mathrm{M}(w, G)$ computed with the weighting scheme w. The Wiener operator $\mathrm{Wi}(\mathrm{M}, w) = \mathrm{Wi}(\mathrm{M}, w, G)$ is [15, 46] (Eq. (34)):

$$\mathbf{Wi}(\mathbf{M}, w, G) = \sum_{i=1}^{N} \sum_{j=i}^{N} [\mathbf{M}(w, G)]_{ij} \tag{34}$$

1.3.4 The Hyper-Wiener Operator

The extension of the hyper-Wiener index WW [22, 23] to other molecular matrices was done with the hyper-Wiener operator **HyWi** that can be computed both for simple and weighted molecular graphs. Consider the vertex- and edge-weighted graph G with N vertices and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w. The hyper-Wiener operator $\mathbf{HyWi}(\mathbf{M}, w) = \mathbf{HyWi}(\mathbf{M}, w, G)$ of the VEW graph G is [28] (Eq. (35)):

$$\mathbf{HyWi}(\mathbf{M}, w, G) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=i}^{N} ([\mathbf{M}(w, G)]_{ij}^{\ 2} + [\mathbf{M}(w, G)]_{ij}) \tag{35}$$

The hyper-Wiener operator can be applied to any symmetric molecular matrix, such as the adjacency, distance, reciprocal distance, resistance distance, detour, and distance-valency matrixes. If **M** is the distance matrix, the **HyWi** operator is identical with the hyper-Wiener index *WW*.

1.3.5 The Szeged Operator

In computing the Szeged index, the contribution of a vertex to the numbers n_i and n_j is constant and equal to 1, as can be seen from Eqs. (8) and (9). The Szeged operator considers vertex contributions by applying a function to a subset of vertices from the molecular graph, when each vertex is characterized by a local invariant, such as the degree or distance sum. Consider a vertex structural descriptor **VSD** that assigns a numerical value $\mathbf{VSD}_i$ to each vertex v_i from the molecular graph G. Let e_{ij} be an edge of the molecular graph G, connecting the vertices v_i and v_j from G, $v_i, v_j \in V(G)$. Let s_i be the value of the function $f = f(\mathbf{VSD}_k)$ computed from the $\mathbf{VSD}_k$ values for all vertices v_k of the graph G having the property $td_{ki} < td_{kj}$ and let s_j be the value of the function $f = f(\mathbf{VSD}_k)$ computed from the $\mathbf{VSD}_k$ values for all vertices v_k of the graph G having the property $td_{kj} < td_{ki}$. When a vertex v_k is situated at the same distance from vertices v_i and v_j, i.e. $td_{ki} = d_{kj}$, its $\mathbf{VSD}_k$ value is not considered neither in s_i nor in s_j. A formal definition of s_i and s_j is offered below (Eqs. (36) and (37)):

$$s_i = \{f(\mathbf{VSD}_k, p) : v_i, v_j, v_k \in V(G), e_{ij} \in E(G), td_{ki} < td_{kj}\} \tag{36}$$

$$s_j = \{f(\mathbf{VSD}_k, p) : v_i, v_j, v_k \in V(G), e_{ij} \in E(G), td_{kj} < td_{ki}\} \tag{37}$$

The Szeged operator of the molecular graph G is defined by the equation [47] (Eq. (38)):

$$\mathbf{Sz}(\mathbf{VSD}, f, p, G) = \sum_{e_{ij} \in E(G)} s_i s_j \tag{38}$$

where the summation goes over all edges e_{ij} from the edge set $E(G)$, $e_{ij} \in E(G)$, and p is the constant power used in the function f. When the $\mathbf{VSD}_i$ is equal to 1 for all vertices in the graph G, the Szeged operator is identical to the Szeged index *Sz*. The vertex structural descriptor **VSD** that weights the contribution of each vertex to the numbers s_i and s_j can be the degree **deg**, valency **val**, distance sum **DS**, vertex sum **VS**, or any other vertex descriptor. The following three functions are used to compute the s_i and s_j vertex descriptors from Eqs (36) and (37) (Eqs. (39)–(41)):

$$f_1(\mathbf{VSD}, p)_i = \sum_{td_{ki} < td_{kj}} [\mathbf{VSD}_k]^p \tag{39}$$

$$f_2(\mathbf{VSD}, p)_i = \sum_{td_{ki}<td_{kj}} \left[\frac{\mathbf{VSD}_k}{1 + td_{ki}}\right]^p \tag{40}$$

$$f_3(\mathbf{VSD}, p)_i = \sum_{td_{ki}<td_{kj}} \left[\frac{\mathbf{VSD}_k}{(1 + td_{ki})^2}\right]^p \tag{41}$$

In the above three equations the summation goes over all vertices v_k with the property $td_{ki} < td_{kj}$. The Szeged operator $\mathbf{Sz}(\mathbf{1}, f_1, 1)$, i.e. when the function f_1 with $p = 1$ uses a contribution equal to 1 for each vertex, is identical with the Szeged index *Sz*.

1.3.6
The Characteristic Polynomial Operator

The characteristic polynomial operator $\mathbf{Ch}(\mathbf{M}, w, G, x)$ represents the characteristic polynomial of the matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w for a molecular graph G with N vertices [46] (Eq. (42)):

$$\mathbf{Ch}(\mathbf{M}, w, G, x) = \mathbf{det}(x\mathbf{I} - \mathbf{M}(w, G)) = \sum_{n=0}^{N} c_n x^{N-n} \tag{42}$$

where I is the unit matrix of order N, and c_n is the n-th coefficient of the characteristic polynomial. When M is the adjacency matrix A the characteristic polynomial can be denoted $\mathrm{Ch}(w, G)$.

1.3.7
The Matrix Spectrum Operators

Consider the VEW graph G with N vertices and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w. The matrix spectrum operator $\mathbf{Sp}(\mathbf{M}, w, G) = \{x_i, i = 1, 2, \ldots, N\}$ represents the eigenvalues of the matrix $\mathbf{M}(w)$ or the roots of the polynomial $\mathbf{Ch}(\mathbf{M}, w, G, x)$, $\mathbf{Ch}(\mathbf{M}, w, G, x) = 0$. The $\mathbf{MinSp}(\mathbf{M}, w, \mathrm{G})$ and $\mathbf{MaxSp}(\mathbf{M}, w, G)$ operators are equal to the minimum and maximum values of $\mathbf{Sp}(\mathbf{M}, w, G)$, respectively [44] (Eqs. (43) and (44)):

$$\mathbf{MinSp}(\mathbf{M}, w, G) = \mathbf{min}\{\mathbf{Sp}(\mathbf{M}, w, G)\} \tag{43}$$

$$\mathbf{MaxSp}(\mathbf{M}, w, G) = \mathbf{max}\{\mathbf{Sp}(\mathbf{M}, w, G)\} \tag{44}$$

Lovász and Pelikán demonstrated that the largest eigenvalue (spectral radius) $\mathbf{MaxSp}(\mathbf{A}, G)$ of the adjacency matrix of the graph G reflects the graph branching [48]. Their work had a considerable significance in measuring the structure and branching of molecular graphs, and encouraged Burden to use the smallest eigenvalues of the Burden matrix **B** as a numerical measure of molecular complexity [49]. In the original definition, the diagonal elements of the Burden matrix **B** were the atomic number Z of the atoms from the molecular graph. Pearlman

considered that the atomic number is not relevant for intermolecular interactions, and proposed four modified **B** matrices by putting other atomic properties on the diagonal, namely atomic charges, polarizabilities, hydrogen-bond donor- and acceptor-abilities, corresponding to the electrostatic, dispersion, and hydrogen-bonding modes of biomolecular interaction [50]. Using the modified **B** matrices, Pearlman defined the BCUT set of topological indices, representing the lowest and highest eigenvalues of a Burden molecular matrix; based on Burden's (B) modified adjacency matrix validated by the Chemical Abstracts Service (C) as a similarity searching method and extended by Pearlman at the University of Texas (UT), the BCUT descriptors are used in combinatorial chemistry, virtual screening, diversity measure, and QSAR models [50, 51]. The $\mathbf{MinSp}(\mathbf{M}, w, G)$ and $\mathbf{MaxSp}(\mathbf{M}, w, G)$ operators are therefore a generalization of the Lovász and Pelikán index and BCUT descriptors.

The spectral diameter of the molecular matrix $\mathbf{M}(w)$ is equal to the difference between the maximum and minimum eigenvalue of **M** [44] (Eq. (45)):

$$\mathbf{SpDiam}(\mathbf{M}, w, G) = \mathbf{MaxSp}(\mathbf{M}, w, G) - \mathbf{MinSp}(\mathbf{M}, w, G) \tag{45}$$

The sum of the values of the spectrum $\mathrm{Sp}(\mathrm{M}, w, G)$ is [44] (Eq. (46)):

$$\mathbf{SpSum}(p, \mathbf{M}, w, G) = \sum_{i=1}^{N} |\mathbf{Sp}(\mathbf{M}, w, G)_i|^p \tag{46}$$

where p is a parameter, and w is the weighting scheme used to compute the matrix $\mathrm{M}(w)$. The sum of the positive values of the spectrum $\mathrm{Sp}(\mathrm{M}, w)$ is [44] (Eq. (47)):

$$\mathbf{SpSum}(+, p, \mathbf{M}, w, G) = \sum_{i} |Sp(+, \mathbf{M}, w, G)_i|^p \tag{47}$$

where $\mathbf{Sp}(+, \mathbf{M}, w, G)_i$ is the Ith positive value in the spectrum of the molecular matrix $\mathbf{M}(w, G)$. The sum of the negative values of the spectrum $\mathbf{Sp}(\mathbf{M}, w)$ is [44] (Eq. (48)):

$$\mathbf{SpSum}(-, p, \mathbf{M}, w, G) = \sum_{i} |\mathbf{Sp}(-, \mathbf{M}, w, G)_i|^p \tag{48}$$

where $\mathbf{Sp}(-, \mathbf{M}, w, G)_i$ is the ith negative value in the spectrum of the molecular matrix $\mathbf{M}(w, G)$.

1.3.8
The Spectral Moment Operator

The spectral moment operator of order k, $\mathbf{SM}(\mathbf{M}, w, G)_k$, of the matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w for the molecular graph G with N

vertices is defined as [46] (Eq. (49)):

$$\mathbf{SM}(\mathbf{M}, w, G)_k = \sum_{i=1}^{N} x_i^k = \mathbf{Tr}\ \mathbf{M}(w, G)^k = \sum_{i=1}^{N} [\mathbf{M}(w, G)^k]_{ii} \tag{49}$$

where x_i is the Ith matrix eigenvalue (or spectrum element), and $\mathbf{Tr}\ \mathbf{M}(w)^k$ is the trace of the kth power of the molecular matrix $\mathbf{M}(w)$.

1.3.9
The Hosoya Operator

Let $\mathrm{M}(w) = \mathrm{M}(w, G)$ be the molecular matrix computed with the weighting scheme w of a VEW graph G with N vertices. The Hosoya operator $\mathrm{Ho}(\mathrm{M}, w) = \mathrm{Ho}(\mathrm{M}, w, G)$ is defined as the sum of the absolute values for the coefficients c_n of the characteristic polynomial of the matrix M [46] (Eq. (50)):

$$\mathbf{Ho}(\mathbf{M}, w, G) = \sum_{n=0}^{N} |c_n| \tag{50}$$

For acyclic graphs and if $\mathbf{M}$ is the adjacency matrix $\mathbf{A}$, the $\mathbf{Ho}(\mathbf{M}, w)$ index is identical to the Hosoya index $Z(w)$ computed with the same weighting scheme w. Also, if $\mathbf{M}$ is the distance matrix $\mathbf{D}$, the Hosoya operator gives the Z' index, $\mathbf{Ho}(\mathbf{D}, w) = Z'(w)$.

1.3.10
The Ivanciuc–Balaban Operator

The Balaban index J [34–36], initially defined with the distance sum **DS**, was extended for the **VS** local invariant. Consider the VEW graph G with N vertices and its molecular matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w. The Ivanciuc–Balaban operator $\mathbf{IB}(\mathbf{M}, w) = \mathbf{IB}(\mathbf{M}, w, G)$ of the matrix $\mathbf{M}$ is [45] (Eq. (51)):

$$\mathbf{IB}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} [\mathbf{VS}(\mathbf{M}, w, G)_i \mathbf{VS}(\mathbf{M}, w, G)_j]^{-1/2} \tag{51}$$

where M is the number of edges in G, μ is the cyclomatic number of G (the number of cycles in G), and the summation goes over all edges from the edge set $E(G)$. The **IB** operator was recently extended by adding a new parameter, a variable exponent p, which is optimized according to the investigated physical, chemical or biological property [52] (Eq. (52)):

$$\mathbf{IB}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} [\mathbf{VS}(\mathbf{M}, w, G)_i \mathbf{VS}(\mathbf{M}, w, G)_j]^{p} \tag{52}$$

Several QSPR studies indicate that for certain properties, the optimum value of the exponent p is very different from its usual value of $-1/2$.

1.3.11 The Information-theory Operators U(M), V(M), X(M), and Y(M)

The indices U, V, X, and Y for information on distances are computed from the elements of the distance matrix of the molecular graph, and these TIs provided good results both for structure discrimination and in structure–property models [38, 39]. We present here information-theory operators that can be applied to any molecular matrix. The graph vertex operators $\mathbf{VUinf}(\mathbf{M}, w, G)$, $\mathbf{VVinf}(\mathbf{M}, w, G)$, $\mathbf{VXinf}(\mathbf{M}, w, G)$, and $\mathbf{VYinf}(\mathbf{M}, w, G)$ apply the information theory equations to the absolute values of the elements of the molecular matrix $\mathbf{M}(w, G)$. With certain weighting schemes w, the vertex and edge parameters Vw and Ew can have negative values. Because the logarithm is defined only for positive arguments, the four graph vertex operators are computed from the elements of a positive matrix $\mathbf{P}(w) = \mathbf{P}(w, G)$ whose element $[\mathbf{P}(w)]_{ij}$ is equal to the absolute value of the corresponding element from the $\mathbf{M}(w)$ matrix, element $[\mathbf{P}(w)]_{ij} = |[\mathbf{M}(w)]_{ij}|$. The graph vertex operators are defined by Eqs. (53)–(55):

$$\mathbf{VUinf}(\mathbf{M}, w, G)_i = -\sum_{j=1}^{N} \frac{[\mathbf{P}(w)]_{ij}}{\mathbf{VS}(\mathbf{P}, w)_i} \log_2 \frac{[\mathbf{P}(w)]_{ij}}{\mathbf{VS}(\mathbf{P}, w)_i} \tag{53}$$

$$\mathbf{VVinf}(\mathbf{M}, w, G)_i = \mathbf{VS}(\mathbf{P}, w)_i \log_2 \mathbf{VS}(\mathbf{P}, w)_i - \mathbf{VUinf}(\mathbf{M}, w)_i \tag{54}$$

$$\mathbf{VXinf}(\mathbf{M}, w, G)_i = \mathbf{VS}(\mathbf{P}, w)_i \log_2 \mathbf{VS}(\mathbf{P}, w)_i - \mathbf{VYinf}(\mathbf{M}, w)_i \tag{55}$$

$$\mathbf{VYinf}(\mathbf{M}, w, G)_i = \sum_{j=1}^{N} [\mathbf{P}(w)]_{ij} \log_2 [\mathbf{P}(w)]_{ij} \tag{56}$$

where $\mathbf{VS}(\mathbf{P}, w)_i$ is the vertex sum of the vertex v_i computed from the matrix $\mathbf{P}$, w is the weighting scheme, and the summations in Eqs. (53) and (56) are done for the absolute values of the non-zero elements of the molecular matrix $\mathbf{P}$, $[\mathbf{P}(w)]_{ij} \neq 0$. For the notation of the four graph vertex operators $\mathbf{VUinf}(\mathbf{M}, w, G)$, $\mathbf{VVinf}(\mathbf{M}, w, G)$, $\mathbf{VXinf}(\mathbf{M}, w, G)$, and $\mathbf{VYinf}(\mathbf{M}, w, G)$ we have maintained the molecular matrix $\mathbf{M}$ to indicate the source of the invariants.

For a general molecular graph matrix $\mathbf{M}$, the matrix elements $[\mathbf{M}]_{ij}$ may have values lower than 1, giving negative terms for certain vertex structural descriptors computed with the graph vertex operators $\mathbf{VUinf}(\mathbf{M}, w, G)$, $\mathbf{VVinf}(\mathbf{M}, w, G)$, $\mathbf{VXinf}(\mathbf{M}, w, G)$, and $\mathbf{VYinf}(\mathbf{M}, w, G)$. The Randić-like formula used in the case of the indices U, V, X, and Y is therefore replaced by Eq. (57):

$$f(x, y) = \begin{cases} (xy)^{-1/2} & \text{if } xy > 0 \\ -(|xy|)^{-1/2} & \text{if } xy < 0 \end{cases} \tag{57}$$

The information on matrix elements operators $\mathbf{U}(\mathbf{M}, w)$, $\mathbf{V}(\mathbf{M}, w)$, $\mathbf{X}(\mathbf{M}, w)$, and $\mathbf{Y}(\mathbf{M}, w)$ are computed with Eqs. (58)–(61):

$$\mathbf{U}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} f(\mathbf{VUinf}(\mathbf{M}, w)_i, \mathbf{VUinf}(\mathbf{M}, w)_j) \tag{58}$$

$$\mathbf{V}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} f(\mathbf{VVinf}(\mathbf{M}, w)_i, \mathbf{VVinf}(\mathbf{M}, w)_j) \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{59}$$

$$\mathbf{X}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} f(\mathbf{VXinf}(\mathbf{M}, w)_i, \mathbf{VXinf}(\mathbf{M}, w)_j) \tag{60}$$

$$\mathbf{Y}(\mathbf{M}, w, G) = \frac{M}{\mu + 1} \sum_{E(G)} f(\mathbf{VYinf}(\mathbf{M}, w)_i, \mathbf{VYinf}(\mathbf{M}, w)_j) \tag{61}$$

1.3.12 Information Theory Indices

Using information theory one can develop effective structural indices by considering the molecular matrixes and the descriptors derived from them as structures which can be partitioned into subsets of elements that are equivalent according to certain rules [37–39, 53–55]. Obviously, the partitioning of a molecular graph into equivalence classes depends on the particular graph descriptor and the equivalence rules. Several topological indices were derived using information theory both from graph matrixes and vertex invariants. The use of some of the information indices is restricted to alkanes [37], thus greatly limiting their application in QSPR and QSAR studies. We present here the main equations that are used to measure the information content of a system S composed a set of n elements. Using a set of equivalence rules, one establishes a partition P of the n elements into k subsets $n_1, n_2, \ldots, n_k$, with the property that $n_1 + n_2 + \cdots + n_k = n$. In the partition P, denoted with $P = n\{n_1, n_2, \ldots, n_k\}$, the equivalence class i contains n_i elements that have a common property according to the equivalence rules. The information content $I = I(S)$ of the partition P of a system S is [56] (Eq. (62)):

$$I = n \log_2 n - \sum_{i=1}^{k} n_i \log_2 n_i \tag{62}$$

where the logarithm is taken at basis 2 for measuring the information content in bits. Using the partition P, one can define a probability distribution $p_i = n_i/n$, representing the probability for a randomly chosen element to belong to class i. The mean information content $H = H(S)$ of the partition P of a system S is [56] (Eq. (63)):

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i \tag{63}$$

This formula is known as the Shannon equation. Onicescu defined the information energy content $E = E(S)$ of the partition P of a system S [57] (Eq. (64)):

$$E = \sum_{i=1}^{k} p_i^2 \tag{64}$$

Using the above three equations, we present here several information theory operators that can generate structural descriptors from a large variety of equivalence classes derived from molecular matrixes.

1.3.13
Information Theory Operators Derived from the Equality of Matrix Elements

Consider the VEW graph G with N vertices and its symmetric molecular matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w. Since the matrix $\mathbf{M}$ is symmetric, we consider only the elements from the upper-triangle part of the matrix and from the main diagonal, a total of $N(N+1)/2$ elements. The elements of the matrix $\mathbf{M}$ are partitioned in k classes by placing in class i all α_i elements with identical values. The partition P_1, $P_1 = N(N+1)/2\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$, is used to define three information theory operators. These operators can be applied to dense symmetric molecular matrixes, i.e. matrixes with non-diagonal elements different from zero. For some weighting schemes the vertex weight for a carbon atom is zero; in such cases, all diagonal zero values are collected together into an equivalence class.

The total information content derived from the equality of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (65)):

$$\mathbf{TIC}(\mathrm{e}, \mathbf{M}, w) = \frac{N(N+1)}{2} \log_2 \frac{N(N+1)}{2} - \sum_{i=1}^{k} \alpha_i \log_2 \alpha_i \tag{65}$$

The mean information content computed from the equality of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (66)):

$$\mathbf{MIC}(\mathrm{e}, \mathbf{M}, w) = \frac{2\mathbf{TIC}(\mathrm{e}, \mathbf{M}, w)}{N(N+1)} = -\sum_{i=1}^{k} \frac{2\alpha_i}{N(N+1)} \log_2 \frac{2\alpha_i}{N(N+1)} \tag{66}$$

The information energy content obtained from the equality of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (67)):

$$\mathbf{E}(\mathrm{e}, \mathbf{M}, w) = \sum_{i=1}^{k} \left(\frac{2\alpha_i}{N(N+1)} \right)^2 \tag{67}$$

Using a different partition, similar operators can be defined for sparse matrixes, i.e. matrixes that have some zero non-diagonal elements, for example the adjacency matrix. The elements of the matrix **M** are partitioned in k classes by placing in class i all β_i elements with identical values; all zero non-diagonal elements are ignored. The total number of elements of the partition is $\beta = \beta_1 + \beta_2 + \cdots + \beta_k$. As explained above, some diagonal elements representing carbon atoms can be zero for certain weighting schemes; in such cases, an equivalence class is formed by collecting there all diagonal zero values. The partition P_2, $P_2 = \beta\{\beta_1, \beta_2, \ldots, \beta_k\}$, is used to define the total information content derived from the equality of the elements of the matrix $\mathbf{M}(w)$ [44] (Eq. (68)):

$$\mathbf{TIC}(\mathrm{e}, \mathbf{M}, w) = \beta \log_2 \beta - \sum_{i=1}^{k} \beta_i \log_2 \beta_i \tag{68}$$

The mean information content computed from the equality of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (69)):

$$\mathbf{MIC}(\mathrm{e}, \mathbf{M}, w) = \frac{\mathbf{TIC}(\mathrm{e}, \mathbf{M}, w)}{\beta} = -\sum_{i=1}^{k} \frac{\beta_i}{\beta} \log_2 \frac{\beta_i}{\beta} \tag{69}$$

The information energy content obtained from the equality of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (70)):

$$\mathbf{E}(\mathrm{e}, \mathbf{M}, w) = \sum_{i=1}^{k} \left(\frac{\beta_i}{\beta}\right)^2 \tag{70}$$

1.3.14
Information Theory Operators Derived from the Size of Matrix Elements

The size of the distance matrix elements was used to define two information indices for alkanes [37]. We present here three graph operators that can derive such indices for any symmetric molecular matrix. Consider the VEW graph G with N vertices and its symmetric molecular matrix $\mathbf{M}(w) = \mathbf{M}(w, G)$ computed with the weighting scheme w, and denote with m the number of non-zero matrix elements situated on its upper-triangle part and from its main diagonal. For some weighting schemes, certain vertex and edge weights have negative values, and in such cases some elements of the matrix **M** have negative values. Because for negative numbers the logarithm is not defined, we introduce the absolute matrix **M**, denoted **AM**, whose elements are the absolute values of the corresponding elements in the matrix **M**, $[\mathbf{AM}(w)]_{ij} = |[\mathbf{M}(w)]_{ij}|$. The sum of these m matrix elements is the Wiener index of the matrix **AM**, $\mathbf{Wi}(\mathbf{AM}, w)$. The non-zero elements from the matrix **AM** form a list with m elements. The total information content computed from the size of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (71)):

$$\mathbf{TIC}(\mathrm{s},\mathbf{M},w)=\mathbf{Wi}(\mathbf{AM},w)\log_2\mathbf{Wi}(\mathbf{AM},w)-\sum_{k=1}^{m}\mathbf{AM}(w)_k\log_2\mathbf{AM}(w)_k \tag{71}$$

where $\mathbf{AM}(w)_k$ represents the kth element from the list of m non-zero **AM** elements.

The mean information content obtained from the size of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (72)):

$$\mathbf{MIC}(\mathrm{s},\mathbf{M},w)=\frac{\mathbf{TIC}(\mathrm{s},\mathbf{M},w)}{\mathbf{Wi}(\mathbf{AM},w)}=-\sum_{k=1}^{m}\frac{\mathbf{AM}(w)_k}{\mathbf{Wi}(\mathbf{AM},w)}\log_2\frac{\mathbf{AM}(w)_k}{\mathbf{Wi}(\mathbf{AM},w)} \tag{72}$$

The information energy content derived from the size of the elements of the matrix $\mathbf{M}(w)$ is [44] (Eq. (73)):

$$\mathbf{E}(\mathrm{s},\mathbf{M},w)=\sum_{k=1}^{m}\left(\frac{\mathbf{AM}(w)_k}{\mathbf{Wi}(\mathbf{AM},w)}\right)^2 \tag{73}$$

1.3.15
Information Theory Operators Derived from the Equality of Vertex Invariants

Consider a vertex structural descriptor $\mathbf{VSD}(\mathrm{M},w)=\mathbf{VSD}(\mathrm{M},w,G)$ that assigns a numerical invariant $\mathbf{VSD}(\mathrm{M},w)_i$ to each vertex v_i from the VEW molecular graph G with N vertices. The elements of the vector **VSD** are partitioned into k classes by placing in the class i all γ_i elements with identical values, giving a total number of elements of the partition $N=\gamma_1+\gamma_2+\cdots+\gamma_k$. The partition P_3, $P_3=N\{\gamma_1,\gamma_2,\ldots,\gamma_k\}$, is used to define the total information content derived from the equality of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M},w)$ [44] (Eq. (74)):

$$\mathbf{TIC}(\mathrm{e},\mathbf{VSD},\mathbf{M},w)=N\log_2 N-\sum_{i=1}^{k}\gamma_i\log_2\gamma_i \tag{74}$$

The mean information content computed from the equality of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M},w)$ is [44] (Eq. (75)):

$$\mathbf{MIC}(\mathrm{e},\mathbf{VSD},\mathbf{M},w)=\frac{\mathbf{TIC}(\mathrm{e},\mathbf{VSD},\mathbf{M},w)}{N}=-\sum_{i=1}^{k}\frac{\gamma_i}{N}\log_2\frac{\gamma_i}{N} \tag{75}$$

The information energy content obtained from the equality of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M},w)$ is [44] (Eq. (76)):

$$\mathbf{E}(\mathrm{e},\mathbf{VSD},\mathbf{M},w)=\sum_{i=1}^{k}\left(\frac{\gamma_i}{N}\right)^2 \tag{76}$$

1.3.16
Information Theory Operators Derived from the Size of Vertex Invariants

Consider a vertex structural descriptor $\mathbf{VSD}(\mathrm{M}, w) = \mathbf{VSD}(\mathrm{M}, w, G)$ that assigns a numerical invariant $\mathbf{VSD}(\mathrm{M}, w)_i$ to each vertex v_i from the VEW molecular graph G with N vertices. As already pointed, for some weighting schemes, certain vertex and edge weights have negative values. In such cases some elements of the matrix M have negative values and for certain molecules one can obtain negative values for a given vertex structural descriptor **VSD**. Because the logarithm is not defined for negative numbers, we use the absolute values for the respective **VSD** invariants. The sum of the absolute **VSD** values for all vertices in the molecular graph G is denoted with SA**VSD**(M, w) [44] (Eq. (77)):

$$\mathbf{SAVSD}(\mathbf{M}, w) = \sum_{i=1}^{N} |\mathbf{VSD}(\mathbf{M}, w)_i| \tag{77}$$

The total information content computed from the size of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M}, w)$ is [44] (Eq. (78)):

$$\mathbf{TIC}(\mathrm{s}, \mathbf{VSD}, \mathbf{M}, w) = \mathbf{SAVSD}(\mathbf{M}, w) \log_2 \mathbf{SAVSD}(\mathbf{M}, w) - \sum_{i=1}^{N} |\mathbf{VSD}(\mathbf{M}, w)_i| \log_2 |\mathbf{VSD}(\mathbf{M}, w)_i| \tag{78}$$

The mean information content computed from the size of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M}, w)$ is [44] (Eq. (79)):

$$\mathbf{MIC}(\mathrm{s}, \mathbf{VSD}, \mathbf{M}, w) = -\sum_{i=1}^{N} \frac{|\mathbf{VSD}(\mathbf{M}, w)_i|}{\mathbf{SAVSD}(\mathbf{M}, w)} \log_2 \frac{|\mathbf{VSD}(\mathbf{M}, w)_i|}{\mathbf{SAVSD}(\mathbf{M}, w)} \tag{79}$$

The information energy content obtained from the size of the elements of the vertex structural descriptor $\mathbf{VSD}(\mathrm{M}, w)$ is [44] (Eq. (80)):

$$\mathbf{E}(\mathrm{s}, \mathbf{VSD}, \mathbf{M}, w) = \sum_{i=1}^{N} \left(\frac{|\mathbf{VSD}(\mathbf{M}, w)_i|}{\mathbf{SAVSD}(\mathbf{M}, w)} \right)^2 \tag{80}$$

1.4
Topological Indices for Combinatorial Chemistry

In the drug discovery process combinatorial libraries (CL) and high–throughput screening (HTS) are efficiently used to identify biologically active molecules more rapidly than with the conventional approaches. An efficient way to reduce the

number of compounds that enter the HTS process is the in silico screening of CL, a process applied both to diverse and focused libraries with the aim to select for HTS the compounds with potential 'drug–like' characteristics and sufficient diversity. The process of virtual screening of combinatorial libraries (VSCL) starts from a wide selection of reactants that are used to generate in silico a huge number of chemical compounds. Then, the structural descriptors relevant for the investigated biological target are identified and computed for all compounds in the virtual library. Finally, the compounds for chemical synthesis and HTS are selected with a statistical algorithm that implements a similarity, diversity, or drug–like paradigm.

In VSCL the chemical structure is translated into a numerical form with the aid of various structural descriptors, many of them traditionally used in QSPR and QSAR. To be efficient, the in silico compound screening uses descriptors that require small computational resources, such as counts of atom types, counts of functional groups, fingerprints, constitutional descriptors, graph invariants and topological indices. A recent VSCL method proposes to compute the topological indices of reaction products without actually assembling the molecules from the building blocks [58–60]. Using various graph decomposition equations, several algorithms were developed for computing the Wiener-type indices, Wiener polynomial and Ivanciuc–Balaban indices for large combinatorial libraries. The TI of the reaction product is obtained from the topological indices of the reactants, thus greatly increasing the efficiency of generating the structural descriptors in the VSCL process.

1.5 Conclusions

Molecular graph descriptors are widely used in modeling physical, chemical, or biological properties, in similarity and diversity assessment, database mining, and in the virtual screening of combinatorial libraries. In 1947 Wiener introduced a graph invariant related to graph distances, the now called Wiener index *W*. Hosoya extended the *W* index to cyclic graphs, a development that was essential for the widely acceptance of this topological index and for its applications in QSPR and QSAR models. Among the large number of molecular graph descriptors proposed in the literature, Wiener-type, connectivity, and electrotopological indices are the most widely used topological indices in QSPR, QSAR, similarity, diversity, and virtual screening of combinatorial libraries. Another important contribution to the theory of topological indices was their definition for vertex- and edge-weighted graphs representing molecules with heteroatoms and multiple bonds. Recent publications in the domain of topological indices indicate several directions were interesting developments are expected to occur: development of novel molecular matrixes, that reflect in a numerical form new features of the molecular graph; definition of new graph operators that can generate families of topological indices; QSPR and QSAR applications for large databases of chemical compounds and for new physical, chemical, or biological properties; use of topological indices to mea-

sure the similarity, diversity, and drug–like character of chemical libraries; new algorithms for fast calculation of topological indices for large scale VSCL.

References

1 F. Harary, *Graph Theory*, Addison–Wesley, Reading, MA, **1971**.

2 A. T. Balaban (ed.), *Chemical Applications of Graph Theory*; Academic Press, London, **1976**.

3 I. Gutman, O. E. Polansky, *Mathematical Concepts in Organic Chemistry*, Springer, Berlin, **1986**.

4 N. Trinajstić, *Chemical Graph Theory, 2nd ed.*, CRC Press, Boca Raton, **1992**.

5 A. T. Balaban (ed.), *From Chemical Topology to Three-Dimensional Geometry*, Plenum, New York, **1997**.

6 L. B. Kier, L. H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, **1976**.

7 A. T. Balaban, A. Chiriac, I. Motoc, Z. Simon, *Steric Fit in Quantitative Structure–Activity Relations, Lect. Notes Chem. Vol. 15*, Springer, Berlin, **1980**.

8 D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structure, Research Studies Press – Wiley, Chichester, UK, **1983**.

9 L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press, Letchworth, **1986**.

10 N. Voiculetz, A. T. Balaban, I. Niculescu-Duvaz, Z. Simon, *Modeling of Cancer Genesis and Prevention*, CRC Press, Boca Raton, **1990**.

11 L. B. Kier, L. H. Hall, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, **1999**.

12 J. Devillers, A. T. Balaban (eds.), *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach Science Publishers, The Netherlands, **1999**.

13 M. V. Diudea (ed.), *QSPR/QSAR Studies by Molecular Descriptors*, Nova Science, Huntington, N.Y., **2001**.

14 M. Barysz, G. Jashari, R. S. Lall, V. K. Srivastava, N. Trinajstić, *On the Distance Matrix of Molecules Containing Heteroatoms*. In: *Chemical Applications of Topology and Graph Theory*, Ed.: R. B. King, Elsevier, Amsterdam, **1983**, pp. 222–227.

15 O. Ivanciuc, *Rev. Roum. Chim.* **2000**, *45*, 289–301.

16 H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

17 H. Hosoya, *Bull. Chem. Soc. Japan* **1971**, *44*, 2332–2339.

18 B. Mohar, D. Babić, N. Trinajstić, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 153–154.

19 D. J. Klein, M. Randić, *J. Math. Chem.* **1993**, *12*, 81–95.

20 I. Gutman, B. Mohar, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 982–985.

21 O. Ivanciuc, T.-S. Balaban, A. T. Balaban, *J. Math. Chem.* **1993**, *12*, 309–318.

22 M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 535–540.

23 M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 833–836.

24 I. Gutman, *Graph Theory Notes New York* **1994**, *27*, 9–15.

25 I. Gutman, A. A. Dobrynin, *Graph Theory New York* **1998**, *34*, 37–44.

26 M. V. Diudea, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 292–299.

27 O. Ivanciuc, A. T. Balaban, *MATCH (Commun. Math. Chem.)* **1994**, *30*, 141–152.

28 O. Ivanciuc, *Rev. Roum. Chim.* **2000**, *45*, 587–596.

29 O. Ivanciuc, *ACH – Models Chem.* **2000**, *137*, 607–631.

30 M. Randić, *New J. Chem.* **1997**, *21*, 945–951.

31 O. Ivanciuc, T. Ivanciuc, A. T. Balaban, *Internet Electron. J. Mol. Des.* **2002**, *1*, 467–487, *http://www.biochempress.com*

32 A. T. Balaban, D. Mills, O. Ivanciuc, S. C. Basak, *Croat. Chem. Acta* **2000**, *73*, 923–941.

33 M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

34 A. T. Balaban, *Chem. Phys. Lett.* **1982**, *89*, 399–404.

35 A. T. Balaban, *Pure Appl. Chem.* **1983**, *55*, 199–206.

36 A. T. Balaban, *MATCH (Commun. Math. Chem.)* **1986**, *21*, 115–122.

37 D. Bonchev, N. Trinajstić, *J. Chem. Phys.* **1977**, *67*, 4517–4533.

38 A. T. Balaban, T.-S. Balaban, *J. Math. Chem.* **1991**, *8*, 383–397.

39 A. T. Balaban, V. Feroiu, *Rep. Mol. Theor.* **1990**, *1*, 133–139.

40 P. A. Filip, T.-S. Balaban, A. T. Balaban, *J. Math. Chem.* **1987**, *1*, 61–83.

41 O. Ivanciuc, T.-S. Balaban, P. Filip, A. T. Balaban, *MATCH (Commun. Math. Chem.)* **1992**, *28*, 151–164.

42 M. V. Diudea, I. Gutman, *Croat. Chem. Acta* **1998**, *71*, 21–51.

43 O. Ivanciuc, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412–1422.

44 O. Ivanciuc, *Rev. Roum. Chim.* **2001**, *46*, 243–253.

45 O. Ivanciuc, T. Ivanciuc, A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 395–401.

46 O. Ivanciuc, *Rev. Roum. Chim.* **1999**, *44*, 519–528.

47 O. Ivanciuc, *Rev. Roum. Chim.* **2000**, *45*, 1105–1114.

48 L. Lovász, J. Pelikán, *Period. Math. Hung.* **1973**, *3*, 175–182.

49 F. R. Burden, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.

50 R. S. Pearlman, K. M. Smith, *Perspect. Drug Discovery Design* **1998**, *9/10/11*, 339–353.

51 R. S. Pearlman, K. M. Smith, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.

52 O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A. T. Balaban, *Internet Electron. J. Mol. Des.* **2002**, *1*, 319–331, *http://www.biochempress.com*

53 O. Ivanciuc, A. T. Balaban, *Rev. Roum. Chim.* **1999**, *44*, 479–489.

54 O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 631–643.

55 O. Ivanciuc, T. Ivanciuc, D. Cabrol-Bass, A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 732–743.

56 C. Shannon, W. Weaver, *Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, **1949**.

57 O. Onicescu, *C.R. Acad. Sci. Paris Ser. A*, **1966**, *263*, 841–842.

58 O. Ivanciuc, D. J. Klein, *Croat. Chem. Acta* **2002**, *75*, 577–601.

59 O. Ivanciuc, D. J. Klein, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 8–22.

60 O. Ivanciuc, *Internet Electron. J. Mol. Des.* **2002**, *1*, 1–9, *http://www.biochempress.com*