

# M

- **MACC descriptors**  $\equiv$  *Maximum Auto-Cross-Correlation descriptors*  $\rightarrow$  autocorrelation descriptors
- **MACC-2 transform**  $\rightarrow$  grid-based QSAR techniques ( $\odot$  GRIND descriptors)
- **MACCS keys**  $\rightarrow$  substructure descriptors ( $\odot$  structural keys)
- **macromolecular graph**  $\rightarrow$  biodescriptors ( $\odot$  peptide sequences)
- **Madan chemical degree**  $\rightarrow$  vertex degree
- **magnetic permittivity**  $\rightarrow$  physico-chemical properties ( $\odot$  magnetic susceptibility)
- **magnetic susceptibility**  $\rightarrow$  physico-chemical properties
- **Mahalanobis distance**  $\rightarrow$  similarity/diversity ( $\odot$  Table S7)
- **Main Distance-Dependent Matrix**  $\rightarrow$  4D Molecular Similarity Analysis
- **Mallows  $C_p$**   $\rightarrow$  regression parameters
- **Manhattan distance**  $\rightarrow$  similarity/diversity ( $\odot$  Table S7)
- **map connectivity matrices**  $\rightarrow$  biodescriptors ( $\odot$  proteomics maps)
- **MaP descriptors**  $\rightarrow$  substructure descriptors ( $\odot$  pharmacophore-based descriptors)
- **map invariants**  $\rightarrow$  biodescriptors ( $\odot$  proteomics maps)

## ■ MARCH-INSIDE descriptors

The MARCH-INSIDE (*MARKovian CHemicals "IN Silico" DDesign*) method uses the concepts of Markov's Chain Theory to codify information about the molecular structure [González Díaz, Olazabal *et al.*, 2002; González Díaz, Gia *et al.*, 2003; González Díaz, Torres-Gómez *et al.*, 2005]. This procedure considers as the Markovian states the Pauling's electronegativities of the external electron layers (valence electrons) of any atom core in the molecule. The basic idea underpinning the MARCH-INSIDE approach is that a series of atoms interact to form a molecule at an arbitrary initial time  $t_0$ . Then, after this initial hypothetical situation, electrons start to distribute around cores in discrete intervals of time  $t_k$ .

MARCH-INSIDE descriptors are derived from the different  $k$ th powers of the **electron-transition stochastic matrix**, denoted as  ${}^1\Pi$ , which is a  $\rightarrow$  *stochastic matrix* of dimension  $A \times A$  derived from the  $\rightarrow$  *electronegativity-weighted adjacency matrix*  ${}^{\chi}A$ , modified by a 3D central chirality factor  $\omega$  [González Díaz, Sánchez *et al.*, 2003], as

$$[{}^{\chi}A(\omega)]_{ij} = \begin{cases} \chi_j \cdot e^{\omega_j} & \text{if } (i,j) \in E(G) \\ \chi_i \cdot e^{\omega_i} & \text{if } i=j \\ 0 & \text{if } (i,j) \notin E(G) \end{cases} \quad \omega = 0, \pm 1$$

where  $\chi$  are the atomic Pauling's electronegativities,  $\mathcal{E}(\mathcal{G})$  is the set of edges in the molecular graph  $\mathcal{G}$ , and the variable  $\omega$  accounts for the spatial configuration of every atom in the molecule:  $\omega = +1$ , if the atom has R- or axial configuration or E-isomerism,  $\omega = 0$ , if the atom does not have a specific spatial configuration, and  $\omega = -1$  if the atom has S- or equatorial configuration or Z-isomerism. If atom chirality is not taken into account ( $\omega = 0$ ), this matrix coincides with the electronegativity-weighted adjacency matrix  ${}^{\chi}\mathbf{A}$ .

The electron-transition stochastic matrix  ${}^1\Pi$  is then defined as:

$$[{}^1\Pi(\omega)]_{ij} = \begin{cases} \frac{\chi_j e^{\omega_j}}{VS_i({}^{\chi}\mathbf{A}(\omega))} & \text{if } (i,j) \in \mathcal{E}(\mathcal{G}) \\ \frac{\chi_i e^{\omega_i}}{VS_i({}^{\chi}\mathbf{A}(\omega))} & \text{if } i = j \\ 0 & \text{if } (i,j) \notin \mathcal{E}(\mathcal{G}) \end{cases} \quad \omega = 0, \pm 1$$

where  $VS$  is the  $\rightarrow$  vertex sum operator that returns for the  $i$ th atom the sum of the electronegativity values for all the atoms bonded to the  $i$ th atom, including the  $i$ th atom itself by the following:

$$VS_i({}^{\chi}\mathbf{A}(\omega)) = \sum_{j=1}^A [{}^{\chi}\mathbf{A}(\omega)]_{ij} = \chi_i \cdot e^{\omega_i} + \sum_{j=1}^A a_{ij} \cdot (\chi_j \cdot e^{\omega_j})$$

where  $A$  is the number of the molecule atoms and  $a_{ij}$  the elements of the  $\rightarrow$  adjacency matrix, equal to one for pairs of bonded atoms, and zero otherwise.

#### Example M1

Calculation of the electron-transition stochastic matrix  ${}^1\Pi$  for the molecule shown below.  $VS_i$  and  $CS_j$  indicate the matrix row and column sums, respectively; Pauling's electronegativities are  $\chi_C = 2.5$ ,  $\chi_N = 3.0$ ,  $\chi_O = 3.5$ , and  $\chi_F = 4.0$ .

N#CC(=O)F

${}^1\Pi =$

Atom	O	F	C <sub>1</sub>	C <sub>2</sub>	N	$VS_i$
O	0.583	0	0.417	0	0	1
F	0	0.615	0.385	0	0	1
C <sub>1</sub>	0.280	0.320	0.200	0.200	0	1
C <sub>2</sub>	0	0	0.313	0.313	0.375	1
N	0	0	0	0.455	0.545	1
$CS_j$	0.863	0.935	1.315	0.968	0.920	5

${}^1\Pi =$

Atom	O	F	C <sub>1</sub>	C <sub>2</sub>	N
O	$\frac{O}{O+C_1}$	0	$\frac{C_1}{O+C_1}$	0	0
F	0	$\frac{F}{F+C_1}$	$\frac{C_1}{F+C_1}$	0	0
C <sub>1</sub>	$\frac{O}{C_1+O+F+C_2}$	$\frac{F}{C_1+O+F+C_2}$	$\frac{C_1}{C_1+O+F+C_2}$	$\frac{C_2}{C_1+O+F+C_2}$	0
C <sub>2</sub>	0	0	$\frac{C_1}{C_2+C_1+N}$	$\frac{C_2}{C_2+C_1+N}$	$\frac{N}{C_2+C_1+N}$
N	0	0	0	$\frac{C_2}{N+C_2}$	$\frac{N}{N+C_2}$

The  $k$ th order electron-transition stochastic matrix  ${}^k\Pi(\omega)$  is calculated as the  $k$ th power of  ${}^1\Pi(\omega)$ :

$${}^k\Pi(\omega) = ({}^1\Pi(\omega))^k$$

The elements of this matrix are interpreted as the transition probabilities of electrons of going from the  $i$ th atom to the  $j$ th atom at different time intervals. The diagonal elements are called **self-return probabilities** by analogy with the  $\rightarrow$  *self-returning walks*.

Finally, the MARCH-INSIDE total molecular descriptors are the  $\rightarrow$  *stochastic spectral moments* of the  $k$ th powers of the electron-transition stochastic matrix defined as

$${}^{\text{SR}}\pi_k(\omega) = \text{tr}[{}^k\Pi(\omega)] = \sum_{i=1}^A [{}^k\Pi(\omega)]_{ii}$$

MARCH-INSIDE atom-type descriptors are analogously calculated, but unlike considering the contributions of all atoms in the molecule, only the diagonal elements of each  $k$ th order matrix corresponding to the atom of a given type (e.g., halogens, carbons in aliphatic chains, and so on) are summed up to give the molecular descriptor.

Moreover, to model proteins and peptides, MARCH-INSIDE biodescriptors were derived from the  $k$ th powers of an electron-transition stochastic matrix based on the  $\rightarrow$  *Electronic Charge Index* used in place of the electronegativity [Ramos de Armas, González Díaz *et al.*, 2005].

**MEDNE descriptors** (*Markovian Electron Delocalization NEgentropies*) are a set of molecular descriptors that, like the MARCH-INSIDE descriptors, are calculated from the different  $k$ th powers of the electron-transition stochastic matrix  ${}^k\Pi$  [González Díaz, Marrero *et al.*, 2003; Cruz-Montegudo, González Díaz *et al.*, 2008]. Each  $k$ th order matrix is transformed into an  $A$ -dimensional vector  ${}^A\pi_k$  as the following:

$${}^A\pi_k = {}^A\pi_0^T \cdot {}^k\Pi$$

where  $T$  is the transpose and  ${}^A\pi_0$  an  $A$ -dimensional column vector whose elements  ${}^A\pi_0(i)$  are the normalized electronegativities defined as the following:

$${}^A\pi_0(i) = \frac{\chi_i \cdot e^{\omega_i}}{\sum_{j=1}^A \chi_j \cdot e^{\omega_j}}$$

where  $\omega$  is the chirality factor previously defined. It must be noted that the electronegativity of each  $i$ th atom is here normalized by using the sum of the electronegativities of all the atoms in the molecule and not only the values of the bonded atoms as in the MARCH-INSIDE descriptors.

The elements of the  $A$ -dimensional vector  ${}^A\pi_k$  of the  $k$ th order are called *absolute probabilities*  ${}^A\pi_k(i)$ , and they codify the attraction of each  $i$ th atom over any electron in the molecule at any time  $t_k$  after traveling along the different walks of length smaller than  $k$ .

The  $k$ th order MEDNE descriptor  $\Theta_k$  (called **electronic delocalization entropy** [Ramos de Armas, González Díaz *et al.*, 2004]), which is the sum over all atoms of the entropy  $\Theta_k(i)$  involved in the attraction of electrons at least  $k$  bonds away from any  $i$ th atom in the molecule, and defined as

$$\Theta_k = \sum_{i=1}^A \Theta_k(i) = - \sum_{i=1}^A [{}^A\pi_k(i) \cdot \ln {}^A\pi_k(i)]$$

📖 [González Díaz and Uriarte, 2005; González Díaz, Bonet *et al.*, 2007]

- **mass spectral features** → spectra descriptors
- **matching polynomial** → Hosoya Z-index
- **mathematical representation of molecular descriptors** → molecular descriptors

### ■ matrices of molecules

Matrices are the most common mathematical tool to encode structural information of molecules. They usually are the starting point for the calculation of many → *molecular descriptors* and → *graph invariants*; moreover, they constitute the mathematical form used as the molecule input in the majority of software packages for calculation of molecular descriptors.

Important kinds of matrices are the → *molecular matrix*, which collects atom spatial coordinates, and all the matrices related to → *molecular geometry*, such as the → *geometry matrix*, whose 3D molecular descriptors are derived from, and computational chemistry approaches are based on, → *WHIM weighted covariances matrices* and the → *molecular influence matrix*. Moreover, derived from computational chemistry, the → *charge density matrix* is another fundamental matrix able to give a deep quantum mechanical description of the molecule.

Other important and very popular matrices are the **graph-theoretical matrices**, a huge number of which were proposed in the last decades to derive topological indices and describe molecules from a topological point of view. Graph-theoretical matrices are matrices derived from a molecular graph  $G$ , often from a → *H-depleted molecular graph*. However, in a less restrictive sense, graph-theoretical matrices are all the matrices derived from a molecular graph, even if they encode additional contributions from the molecular geometry or other nontopological quantities. A comprehensive collection of graph-theoretical matrices is reported by Janežič *et al.* [Janežič, Miličević *et al.*, 2007] and extended overviews in [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc and Ivanciuc, 1999].

Graph-theoretical matrices can be either **vertex matrices**, if both rows and columns refer to graph vertices (atoms) and matrix elements encode some property of pairs of vertices, or **edge matrices**, if both rows and columns refer to graph edges (bonds) and matrix elements encode some property of pairs of edges. Vertex matrices are square matrices of dimension  $A \times A$ ,  $A$  being the number of graph vertices, whereas edge matrices are square matrices of dimension  $B \times B$ ,  $B$  being the number of graph edges. In the book, a vertex matrix is usually referred to by omitting the word “vertex,” whereas for edge matrices, the prefix “edge” is always specified in the matrix name and the superscript E is used in the matrix symbol as  $^E\mathbf{M}$ .

Vertex matrices are undoubtedly the graph-theoretical matrices most frequently used for characterizing a molecular graph. The matrix entries encode some information about pairs of vertices, such as their connectivities, topological distances, sums of the weights of the vertices along the connecting paths; the diagonal entries can encode chemical information about the vertices. The most important vertex matrices are the → *adjacency matrix*  $\mathbf{A}$ , which encodes information about vertex connectivities and the → *distance matrix*  $\mathbf{D}$ , which also encodes information about relative locations of graph vertices.

From vertex matrices a huge number of topological indices were proposed. Edge matrices have been less used to characterize a molecular graph and derive molecular descriptors.

The most important edge matrices are the → *edge adjacency matrix*  $^E\mathbf{A}$  and → *the edge distance matrix*  $^E\mathbf{D}$ . Edge matrices of a molecular graph  $G$  are usually calculated from the → *line*

graph of the actual molecular graph  $G$ , whose vertices represent edges of  $G$  [Gutman and Estrada, 1996]; therefore, an edge matrix of  $G$  is simply the corresponding vertex matrix of the line graph of  $G$ . For instance, the edge adjacency matrix of  $G$  is the adjacency matrix  $A$  of the line graph of  $G$ . Following this approach, a number of edge matrices were proposed.

It is noteworthy to point out that in literature vertex matrices are sometimes referred to as “edge-matrices,” denoted by  $M_e$  (or sometimes as  $^eM$ ), when only matrix elements corresponding to the pairs of adjacent vertices are different from zero, or “path-matrices,” denoted by  $M_p$  (or sometimes as  $^pM$ ), when all off-diagonal elements can be different from zero, meaning that all pairs of vertices in the graph (i.e., paths) are accounted for and not only the pairs of adjacent vertices (i.e., edges). The  $\rightarrow$  adjacency matrix  $A$ , which is one of the fundamental vertex matrices, is an example of “edge-matrix,” whereas the  $\rightarrow$  distance matrix  $D$  is the corresponding “path-matrix.” Edge- and path-matrices are usually related, and, specifically, edge-matrices are derived from path-matrices by the following:

$$M_e = M_p \otimes A$$

where  $A$  is the adjacency matrix and the symbol  $\otimes$  indicates the  $\rightarrow$  Hadamard matrix product. Examples of path-matrices and related edge-matrices are  $\rightarrow$  path-Cluj matrices and  $\rightarrow$  edge-Cluj matrices,  $\rightarrow$  path-Wiener matrix and  $\rightarrow$  edge-Wiener matrix,  $\rightarrow$  path-Szeged matrix and  $\rightarrow$  edge-Szeged matrix.

Obviously, the prefix “edge-” in the matrix name is misleading because it can refer to two different kinds of graph-theoretical matrices; therefore, care must be taken when dealing with this terminology. In this book, the original matrix names with the prefix “edge” were retained, although they can be sometimes ambiguous, but with the following artifice: edge matrices ( $B \times B$ ) whose rows and columns refer to graph edges are called edge matrices without the hyphen in the matrix name and denoted as  $^eM$ , whereas edge-matrices ( $A \times A$ ) that are vertex matrices containing information only about pairs of adjacent vertices (i.e., edges) are referred to by using the hyphen in the matrix name and denoted as  $M_e$ .

Together with vertex matrices and edge matrices,  $\rightarrow$  incidence matrices are other important graph-theoretical matrices used to describe a molecular graph. These are matrices whose rows can represent either vertices or edges and whose columns represent some subgraphs, such as edges, paths, or cycles.

Moreover, matrices can be derived from unweighted or vertex- and/or edge-weighted molecular graphs; in the latter case, several  $\rightarrow$  weighted matrices can be obtained depending on the  $\rightarrow$  weighting scheme.

Most of the graph-theoretical matrices are symmetrical, whereas some of them are unsymmetrical. Examples of unsymmetrical matrices are  $\rightarrow$  Szeged matrices,  $\rightarrow$  Cluj matrices,  $\rightarrow$  random walk Markov matrix,  $\rightarrow$  combined matrices such as the topological distance–detour distance combined matrix, and some weighted adjacency and distance matrices.

From unsymmetrical matrices  $UM$ , the corresponding symmetric matrices  $SM$  can be obtained as

$$SM = UM^T \otimes UM$$

where  $\otimes$  is the  $\rightarrow$  Hadamard matrix product, or, alternatively, by adding to the unsymmetrical matrix its transposed matrix as

$$SM = UM + UM^T$$

Given a matrix  $\mathbf{M}$ , some general classes of matrices can be derived by applying algebraic transformations. They are reported below.

Given a matrix  $\mathbf{M}$ , **power matrices**, denoted by  $\mathbf{M}^k$ , are defined as

$$\mathbf{M}^k = \mathbf{M} \cdot \mathbf{M}^{k-1}$$

where  $k$  is the matrix power [Ivanciuc, 2000e]. An easy way to calculate a  $k$ th power of a matrix is passing through the matrix eigenvalue/eigenvector decomposition as

$$\mathbf{M}^k = \mathbf{L}^T \cdot \mathbf{\Lambda}^k \cdot \mathbf{L}$$

where  $\mathbf{L}$  is the matrix collecting the eigenvectors of  $\mathbf{M}$  and  $\mathbf{\Lambda}^k$  is a diagonal matrix whose elements are the eigenvalues  $\lambda^k$  of  $\mathbf{M}$  raised to the  $k$ th power.

Several molecular descriptors are calculated from matrices raised to different powers; for example,  $\rightarrow$  walk counts,  $\rightarrow$  self-returning walk counts, and  $\rightarrow$  spectral moments are derived from the different powers of the adjacency matrix  $\mathbf{A}$  (i.e.,  $\mathbf{M} = \mathbf{A}$ ),  $\rightarrow$   $k\alpha$  descriptors from the powers of the  $\rightarrow$  path- $\chi$  matrix,  $\rightarrow$  random walk counts from the powers of the  $\rightarrow$  random walk Markov matrix,  $\rightarrow$  spectral moments of the edge adjacency matrix and  $\rightarrow$  TOMOCOMD descriptors from the powers of the  $\rightarrow$  edge adjacency matrix.

**Generalized matrices**, denoted by  $\mathbf{M}^\lambda$ , are obtained by using the Hadamard matrix product or, alternatively, by raising to different powers the elements of the matrix  $\mathbf{M}$ :

$$[\mathbf{M}^\lambda]_{ij} = [\mathbf{M}]_{ij}^\lambda$$

where  $\lambda$  is a real parameter.

Examples of molecular descriptors derived from generalized matrices are  $\rightarrow$  distance distribution moments and  $\rightarrow$   $W_\lambda$  indices from the  $\rightarrow$  distance matrix and  $\rightarrow$  molecular profiles from the  $\rightarrow$  geometry matrix. Moreover,  $\rightarrow$  vertex Zagreb matrix ( $\lambda = 2$ ) and  $\rightarrow$  modified vertex Zagreb matrix ( $\lambda = -2$ ) are a generalization of the  $\rightarrow$  vertex degree matrix, and the  $\rightarrow$  generalized molecular-graph matrix is a generalization of the distance matrix based on variable parameters.

**Generalized reciprocal matrices** are a class of generalized matrices obtained by raising the matrix elements to some negative exponent:

$$[\mathbf{M}^{-\lambda}]_{ij} = \begin{cases} [\mathbf{M}]_{ij}^{-\lambda} & \text{if } i \neq j \\ [\mathbf{M}]_{ii} & \text{if } i = j \end{cases}$$

where  $\lambda$  is usually an integer positive parameter. Note that the reciprocal is not applied to diagonal elements. In effect, diagonal elements are equal to zero for simple graphs and for vertex-weighted molecular graphs, they are real values representing atomic properties.

The most popular **reciprocal matrices** are obtained for  $\lambda = 1$ , such as the  $\rightarrow$  Harary matrix,  $\rightarrow$  reciprocal geometry matrix,  $\rightarrow$  reciprocal detour matrix,  $\rightarrow$  reciprocal Szeged matrix,  $\rightarrow$  reciprocal Cluj matrix. The  $\rightarrow$  reciprocal square distance matrix is derived from the distance matrix by setting  $\lambda = 2$ .

Other important classes of graph-theoretical matrices are neighborhood matrices and matrices derived by a combination of pairs of matrices, such as the sum matrices, augmented matrices, difference matrices, complement matrices, quotient matrices, combined matrices, and expanded matrices (Table M1).

**Table M1** Notations of some molecular matrices.

Matrix	Notation	Matrix	Notation
Power matrix	$M^k$	Quotient matrix	$M_1/M_2$
Complement matrix	$^cM$	Combined matrix	$M_1 \wedge M_2$
Neighborhood matrix	$^N M$	Expanded matrix	$M_1 \_ M_2$
Sum matrix	$M_1 M_2 \Sigma$	Difference matrix	$M_1 M_2 \Delta$

Given two equal-sized graph-theoretical matrices  $M_1$  and  $M_2$ , **sum matrices**, denoted as  $M_1 M_2 \Sigma$ , are obtained by summing corresponding elements of matrices  $M_1$  and  $M_2$ :

$$M_1 M_2 \Sigma = M_1 + M_2$$

The most popular sum matrices are the  $\rightarrow$  *adjacency-plus-distance matrix* obtained by summing the vertex  $\rightarrow$  *adjacency matrix*  $A$  and the vertex  $\rightarrow$  *distance matrix*  $D$  and the  $\rightarrow$  *edge-adjacency-plus-edge-distance matrix* obtained by summing the  $\rightarrow$  *edge adjacency matrix*  $^E A$  and the  $\rightarrow$  *edge distance matrix*  $^E D$ . From the adjacency-plus-distance matrix the  $\rightarrow$  *Schultz molecular topological index* is derived, whereas from the edge-adjacency-plus-edge-distance matrix is derived the  $\rightarrow$  *edge-Schultz index*. Moreover, the determinant of the adjacency-plus-distance matrix, that is, the  $\rightarrow$  *det|A + D| index*, was also proposed as a molecular descriptor.

The sum matrix resulting from the  $\rightarrow$  *geometry matrix*  $G$  and the  $\rightarrow$  *bond length-weighted adjacency matrix*  $^b A$ , where elements corresponding to the pairs of adjacent vertices are  $\rightarrow$  *bond distances*, was defined by Mihalić *et al.* [Mihalić, Nikolić *et al.*, 1992]. Its determinant, that is, the  $\rightarrow$  *det| $^b A + G$ | index*, and  $\rightarrow$  *3D-Schultz index* were proposed and used in QSAR modeling as the molecular descriptors.

**Augmented matrices**, denoted as  $^a M$ , are a special case of sum matrices, resulting from the sum of a matrix  $M$  plus a diagonal matrix whose diagonal elements are some atomic properties:

$$^a M = M + \alpha \cdot I \quad \text{and} \quad ^a M = M + p \cdot I$$

where  $I$  is the  $\rightarrow$  *identity matrix*,  $\alpha$  a constant, and  $p$  a  $A$ -dimensional vector containing the considered atomic property. Typical augmented matrices are derived from adjacency and distance matrices, whose diagonal elements equal to zero are replaced with any nonzero value.

Augmented matrices  $^a M$  were defined for the calculation of local vertex invariants by  $\rightarrow$  *MPR approach*. Moreover, Randić [Randić, 1991b] proposed the  $\rightarrow$  *augmented adjacency matrix* by replacing the zero diagonal entries of the adjacency matrix with specific values empirically obtained and by characterizing different atom types in the molecule. The augmented distance matrix was defined by analogy with the augmented adjacency matrix. The  $\rightarrow$  *Laplacian matrix* is another augmented matrix, obtained by the combination of the  $\rightarrow$  *vertex degree matrix* and the adjacency matrix.

Given two equal-sized graph-theoretical matrices  $M_1$  and  $M_2$ , **difference matrices**, denoted as  $M_1 M_2 \Delta$ , are obtained by subtracting the corresponding elements of matrices  $M_1$  and  $M_2$ :

$$M_1 M_2 \Delta = M_1 - M_2$$

The most popular difference matrix is the  $\rightarrow$  *Laplacian matrix* defined as the difference between the  $\rightarrow$  *vertex degree matrix*  $\mathbf{V}$  and the  $\rightarrow$  *adjacency matrix*  $\mathbf{A}$ ; other difference matrices are the  $\rightarrow$  *delta matrix*,  $\rightarrow$  *detour-delta matrix*,  $\rightarrow$  *Wiener difference matrix*,  $\rightarrow$  *Szeged difference matrix*, and  $\rightarrow$  *Cluj difference matrix*.

**Complement matrices** are a special case of difference matrices, resulting from the difference between a matrix with all the off-diagonal elements equal to a constant and a matrix  $\mathbf{M}$ ; they are denoted by  ${}^c\mathbf{M}$  and defined as

$$[{}^c\mathbf{M}]_{ij} = \begin{cases} K - [\mathbf{M}]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where  $K$  is a constant assumed to be greater than all the  $\mathbf{M}$  matrix elements.

Examples of complement matrices are the  $\rightarrow$  *distance complement matrix*,  $\rightarrow$  *complementary distance matrix*,  $\rightarrow$  *reverse Wiener matrix*,  $\rightarrow$  *complement Barysz distance matrix*, and  $\rightarrow$  *detour complement matrix*.

It must be noted that the value of the elements of the reciprocal and complement matrices decreases when the corresponding value in  $\mathbf{M}$  increases; therefore, molecular descriptors calculated from reciprocal or complement matrices have numerical behavior and meaning opposite to molecular descriptors derived from the corresponding original matrix  $\mathbf{M}$ . For instance, the row sums of the  $\rightarrow$  *distance matrix*  $\mathbf{D}$  (i.e., distance degrees) are greater for outer vertices than for the core vertices; thus, the value of the  $\rightarrow$  *Balaban distance connectivity index*, which is based on the inverse of the distance degrees, is much more determined by the core vertices than the outer ones. On the contrary, the row sums of the reciprocal or complement distance matrix are greater for the core vertices and, therefore,  $\rightarrow$  *Balaban-like indices* are much more determined by the outer vertices.

**Neighborhood matrices**, denoted by  ${}^N\mathbf{M}$ , are sparse matrices whose entries are the elements of  $\mathbf{M}$ , which have values smaller than or equal to a predefined threshold  $t$  and zero otherwise:

$$[{}^N\mathbf{M}]_{ij} = \begin{cases} [\mathbf{M}]_{ij} & \text{if } [\mathbf{M}]_{ij} \leq t \\ 0 & \text{if } [\mathbf{M}]_{ij} > t \end{cases}$$

The adjacency matrix is an example of neighborhood matrix trivially obtained by applying a threshold  $t=1$  on the vertex distance matrix; applying a threshold  $t=2$  on the same matrix, only topological distances of 1 and 2 are retained. A threshold less than 1 applied on the  $\rightarrow$  *distance/detour quotient matrix* results in a matrix whose elements different from zero are, to some extent, related to cyclic substructures. Applied on the  $\rightarrow$  *geometry matrix*, a threshold  $t$  equal to a predefined geometric distance produces a sparse geometry matrix where only the pairs of atoms not too far from each other are considered, thus accounting for the most relevant interactions.

Derived from a geometry matrix collecting the Euclidean distances between spots of a proteomics map, the  $\rightarrow$  *neighborhood geometry matrix* was originally proposed to calculate descriptors of  $\rightarrow$  *proteomics maps* by the additional constraint that the matrix element  $i-j$  is set at zero also for nonconnected protein spots [Bajzer, Randić *et al.*, 2003].

Given two graph-theoretical matrices  $\mathbf{M}_1$  and  $\mathbf{M}_2$  of the same size, **quotient matrices** are matrices, denoted by  $\mathbf{M}_1/\mathbf{M}_2$ , whose elements are given by the ratio of the off-diagonal elements



of  $M_1$  over the corresponding elements of  $M_2$  [Randić, Kleiner *et al.*, 1994]:

$$[M_1/M_2]_{ij} = \begin{cases} \frac{[M_1]_{ij}}{[M_2]_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The most popular quotient matrices are the  $\rightarrow$  distance/distance matrices obtained by the ratio of two different measures of the separation between molecule atoms. These separation measures can be either 2D if derived from the  $\rightarrow$  molecular graph or 3D if derived from the  $\rightarrow$  molecular geometry. The quotient matrices proposed as first are the  $\rightarrow$  topological distance/detour distance quotient matrix, based on 2D distance measures, and the  $\rightarrow$  geometric distance/topological distance quotient matrix (or alternatively the  $\rightarrow$  topographic distance/topological distance quotient matrix), based on a 3D and a 2D distance measure.

Note that if both distances are measured through bonds, then the resulting quotient matrix is not meaningful for acyclic graphs, since all the off-diagonal matrix elements are equal to one. Moreover, matrices obtained by the ratio of a 3D distance (geometric or topographic) over a 2D distance (topological, detour, or resistance) are the same for acyclic structures independently of the chosen 2D distance measure, these 2D measures being all the same. In Table M2, a collection of quotient matrices is reported.

**Table M2** List of quotient matrices and their suggested symbols.

ID	$M_1$	$M_2$	$M_1/M_2$	Matrix name	*
1	G	D	G/D	Geometric distance/topological distance quotient matrix	
2	G	$\Delta$	G/ $\Delta$	Geometric distance/detour distance quotient matrix	
3	G	$\Omega$	G/ $\Omega$	Geometric distance/resistance distance quotient matrix	
4	T	D	T/D	Topographic distance/topological distance quotient matrix	
5	T	$\Delta$	T/ $\Delta$	Topographic distance/detour distance quotient matrix	
6	T	$\Omega$	T/ $\Omega$	Topographic distance/resistance distance quotient matrix	
7	D	$\Delta$	D/ $\Delta$	Topological distance/detour distance quotient matrix	*
8	D	$\Omega$	D/ $\Omega$	Topological distance/resistance distance quotient matrix	*
9	$\Delta$	$\Omega$	$\Delta/\Omega$	Detour distance/resistance distance quotient matrix	*
10	D	DC	D/DC	Distance/distance complement quotient matrix	
11	D	G	D/G	Topological distance/geometric distance quotient matrix	
12	$\Delta$	G	$\Delta/G$	Detour distance/geometric distance quotient matrix	
13	$\Omega$	G	$\Omega/G$	Resistance distance/geometric distance quotient matrix	
14	D	T	D/T	Topological distance/topographic distance quotient matrix	
15	$\Delta$	T	$\Delta/T$	Detour distance/topographic distance quotient matrix	
16	$\Omega$	T	$\Omega/T$	Resistance distance/topographic distance quotient matrix	
17	$\Delta$	D	$\Delta/D$	Detour distance/topological distance quotient matrix	*
18	$\Omega$	D	$\Omega/D$	Resistance distance/topological distance quotient matrix	*
19	$\Omega$	$\Delta$	$\Omega/\Delta$	Resistance distance/detour distance quotient matrix	*
20	DC	D	DC/D	Distance complement/distance quotient matrix	

“\*” indicates all the matrices that should be calculated only for cycle-containing structures.

The reciprocal of a quotient matrix is still a quotient matrix, obtained by reversing the role of  $M_2$  and  $M_1$  matrices. In the lower part of Table M2, reciprocal matrices (11–20) of the quotient matrices defined above (1–10) are reported.

Given two graph-theoretical matrices  $\mathbf{M}_1$  and  $\mathbf{M}_2$  of the same size, **combined matrices**, denoted by  $\mathbf{M}_1 \wedge \mathbf{M}_2$ , are unsymmetrical matrices whose upper matrix elements are the elements of  $\mathbf{M}_1$  and lower matrix elements are those of  $\mathbf{M}_2$ :

$$[\mathbf{M}_1 \wedge \mathbf{M}_2]_{ij} = \begin{cases} [\mathbf{M}_1]_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ [\mathbf{M}_2]_{ij} & \text{if } i > j \end{cases}$$

Also, the transpose of this matrix can be defined as

$$[\mathbf{M}_2 \wedge \mathbf{M}_1]_{ij} = \begin{cases} [\mathbf{M}_2]_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ [\mathbf{M}_1]_{ij} & \text{if } i > j \end{cases}$$

but note that eigenvalues and the sum of all the matrix elements are the same for both combined matrices and their transpose and, accordingly,  $\rightarrow$  *spectral indices* and  $\rightarrow$  *Wiener-type indices*. However, row sums of combined matrices and the corresponding transposed matrices are different and thus all the related graph invariants. The most popular combined matrices are listed in Table M3.

**Table M3** List of combined matrices and their suggested symbols.

ID	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1 \wedge \mathbf{M}_2$	Matrix name
1	G	D	$G \wedge D$	Geometric distance–topological distance combined matrix
2	G	$\Delta$	$G \wedge \Delta$	Geometric distance–detour distance combined matrix
3	G	$\Omega$	$G \wedge \Omega$	Geometric distance–resistance distance combined matrix
4	T	D	$T \wedge D$	Topographic distance–topological distance combined matrix
5	T	$\Delta$	$T \wedge \Delta$	Topographic distance–detour distance combined matrix
6	T	$\Omega$	$T \wedge \Omega$	Topographic distance–resistance distance combined matrix
7	D	$\Delta$	$D \wedge \Delta$	Topological distance–detour distance combined matrix
8	D	$\Omega$	$D \wedge \Omega$	Topological distance–resistance distance combined matrix
9	$\Delta$	$\Omega$	$\Delta \wedge \Omega$	Detour distance–resistance distance combined matrix
10	D	G	$D \wedge G$	Topological distance–geometric distance combined matrix
11	$\Delta$	G	$\Delta \wedge G$	Detour distance–geometric distance combined matrix
12	$\Omega$	G	$\Omega \wedge G$	Resistance distance–geometric distance combined matrix
13	D	T	$D \wedge T$	Topological distance–topographic distance combined matrix
14	$\Delta$	T	$\Delta \wedge T$	Detour distance–topographic distance combined matrix
15	$\Omega$	T	$\Omega \wedge T$	Resistance distance–topographic distance combined matrix
16	$\Delta$	D	$\Delta \wedge D$	Detour distance–topological distance combined matrix
17	$\Omega$	D	$\Omega \wedge D$	Resistance distance–topological distance combined matrix
18	$\Omega$	$\Delta$	$\Omega \wedge \Delta$	Resistance distance–detour distance combined matrix

As for quotient matrices, in the case of acyclic graphs, combining two different D distance measures does not make sense because it results into a combined matrix coincident with the  $\rightarrow$  *distance matrix D*.

**Expanded matrices**, denoted as  $\mathbf{M}_1 \_ \mathbf{M}_2$ , constitute another class of graph-theoretical matrices resulting from the  $\rightarrow$  *Hadamard matrix product* of two equal-sized matrices  $\mathbf{M}_1$  and  $\mathbf{M}_2$ :

$$\mathbf{M}_1 \_ \mathbf{M}_2 = \mathbf{M}_1 \otimes \mathbf{M}_2$$

The most popular expanded matrices are  $\rightarrow$  *expanded distance matrices*,  $\mathbf{D}_M$ , derived as the Hadamard product between the  $\rightarrow$  *distance matrix*  $\mathbf{D}$  and some different graph-theoretical matrix  $\mathbf{M}$ , such as the  $\rightarrow$  *Wiener matrix*,  $\rightarrow$  *Cluj matrices*,  $\rightarrow$  *Szeged matrix*, and  $\rightarrow$  *walk matrices*. Moreover,  $\rightarrow$  *expanded reciprocal distance matrices*,  $\mathbf{D}^{-1}_M$ , were defined by analogy with the expanded distance matrices by using the  $\rightarrow$  *reciprocal distance matrix*  $\mathbf{D}^{-1}$  instead of the distance matrix in the Hadamard product. Finally,  $\rightarrow$  *expanded geometric distance matrices*,  $\mathbf{G}_M$ , and  $\rightarrow$  *expanded reciprocal geometric distance matrices*,  $\mathbf{G}^{-1}_M$ , were also proposed based on the  $\rightarrow$  *geometry matrix*  $\mathbf{G}$  and its reciprocal matrix, respectively.

**Combinatorial matrices**, denoted by  $\mathbf{M}_B$ , are defined in terms of the binomial coefficient of the elements of a graph-theoretical matrix  $\mathbf{M}$ . Each entry  $i$ - $j$  of the combinatorial matrix is calculated as the following [Diudea, 1996a]:

$$[\mathbf{M}_B]_{ij} = \binom{[\mathbf{M}]_{ij} + k}{2} \quad [\mathbf{M}_B]_{ij} = \frac{[\mathbf{M}]_{ij}^2 + [\mathbf{M}]_{ij}}{2} \quad k = 1$$

where  $k$  is a constant, usually equal to zero or one.

The most common combinatorial matrices are derived from the  $\rightarrow$  *distance matrix* and  $\rightarrow$  *detour matrix*; these are the  $\rightarrow$  *distance-path matrix*,  $\rightarrow$  *detour-path matrix*,  $\rightarrow$  *delta matrix*, and  $\rightarrow$  *detour-delta matrix*.

Other classes of graph-theoretical matrices are  $\rightarrow$  *walk matrices*,  $\rightarrow$  *layer matrices*,  $\rightarrow$  *distance-degree matrices*, and matrices from which  $\rightarrow$  *Schultz-type indices* are derived.

The most important graph-theoretical matrices or classes of graph-theoretical matrices are listed in Table M4.

**Table M4** Some matrices used as representation of the molecular structure: symbol, current name, number of rows and columns ( $A$ , number of vertices;  $B$ , number of edges;  $C^+$ , cyclicity;  $D$ , topological diameter; and  $K$ , maximum walk length), and type ( $S$ , symmetric matrix;  $U$ , unsymmetrical matrix;  $B$ , symmetric or unsymmetrical matrix; and  $D$ , diagonal matrix).

Symbol	Matrix	Rows	Columns	Types
$A$	Adjacency matrix	$A$	$A$	$S$
$\Omega^{AP}$	All-path matrix	$A$	$A$	$S$
$C$	Atom connectivity matrix	$A$	$A$	$S$
${}^ZD$	Barysz distance matrix	$A$	$A$	$S$
${}^aE(r)$	Bond distance-weighted edge adjacency matrix	$B$	$B$	$S$
${}^rE$	Bond order-weighted edge adjacency matrix	$B$	$B$	$S$
$CT$	Charge term matrix	$A$	$A$	$U$
$CJ$	Cluj matrices	$A$	$A$	$B$
$CJ\Delta$	Cluj-detour matrix	$A$	$A$	$B$
$CJD$	Cluj-distance matrix	$A$	$A$	$B$
$D_\Delta$	Delta matrix	$A$	$A$	$S$
$\Delta$	Detour matrix	$A$	$A$	$S$
$\Delta/D$	Detour/distance quotient matrix	$A$	$A$	$U$
$\Delta_P$	Detour-path matrix	$A$	$A$	$S$
$D$	Distance matrix	$A$	$A$	$S$
$D/D$	Distance/distance matrix	$A$	$A$	$S$

(Continued)

Table M4 (Continued)

Symbol	Matrix	Rows	Columns	Types
$D/\Delta$	Distance/detour quotient matrix	$A$	$A$	S
$D_P$	Distance-path matrix	$A$	$A$	S
${}^E A \equiv E$	Edge adjacency matrix	$B$	$B$	S
${}^{EC} I$	Edge-cycle incidence matrix	$B$	$C^+$	U
${}^b A$	Bond length-weighted adjacency matrix	$A$	$A$	S
${}^E D$	Edge distance matrix	$B$	$B$	S
$\tilde{D}$	Expanded distance matrix	$A$	$A$	S
$D\_M$	Expanded distance matrices	$A$	$A$	S
$G\_M$	Expanded geometric distance matrices	$A$	$A$	S
$H\_M$	Expanded reciprocal distance matrices	$A$	$A$	S
$H\_G\_M$	Expanded reciprocal geometric distance matrices	$A$	$A$	S
$EA$	Extended adjacency matrix	$A$	$A$	S
$M$	Galvez matrix	$A$	$A$	U
$V^\lambda$	Generalized vertex degree matrix	$A$	$A$	D
${}^E G$	Geometric edge distance matrix	$B$	$B$	S
$G$	Geometry matrix	$A$	$A$	S
$I$	Incidence matrix	$A$	$B$	U
$IM$	Interaction graph matrices	$A$	$A$	B
$Z$	Hosoya matrix	$A$	$A$	S
$L$	Laplacian matrix	$A$	$A$	S
$LM$	Layer matrices	$A$	$D + 1$	U
$M$	Molecular matrix	$A$	3	U
${}^* D$	Multigraph distance matrix	$A$	$A$	S
${}^N M$	Neighborhood matrices	$A$	$A$	S
$P$	P matrix	$A$	$A$	S
$PM$	Perturbation graph matrices	$A$	$A$	B
$\Delta^{-1}$	Reciprocal detour matrix	$A$	$A$	S
$CJ\Delta^{-1}$	Reciprocal Cluj-detour matrix	$A$	$A$	B
$CJD^{-1}$	Reciprocal Cluj-distance matrix	$A$	$A$	B
$\Delta/D^{-1}$	Reciprocal detour/distance matrix	$A$	$A$	S
$D^{-1}$	Reciprocal distance matrix	$A$	$A$	S
$G^{-1}$	Reciprocal geometry matrix	$A$	$A$	S
$D^{-2}$	Reciprocal square distance matrix	$A$	$A$	B
$SZ^{-1}$	Reciprocal Szeged matrix	$A$	$A$	B
$W^{-1}$	Reciprocal Wiener matrix	$A$	$A$	S
$\Omega^{-1}$	Conductance matrix	$A$	$A$	S
$\Omega$	Resistance matrix	$A$	$A$	S
$RRW$	Restricted random walk matrix	$A$	$A$	U
$SM$	Sequence matrices	$A$	$K$	U
$SZ$	Szeged matrix	$A$	$A$	B
$SZ_{UP}$	Szeged property matrices	$A$	$A$	U
$T$	Topological state matrix	$A$	$A$	S
${}^{VC} I$	Vertex-cycle incidence matrix	$A$	$C^+$	U
$V$	Vertex degree matrix	$A$	$A$	D
${}^k W_M$	Walk diagonal matrix	$A$	$A$	D
$W_{(M_1, M_2, M_3)}$	Walk matrix	$A$	$A$	U
$W$	Wiener matrix	$A$	$A$	S
$ZM$	Zagreb matrices	$A$	$A$	S
$\chi$	$\chi$ matrix	$A$	$A$	S

📖 [Rouvray, 1976; Kunz, 1989; Randić, Guo *et al.*, 1993; Ivanciuc, Ivanciuc *et al.*, 1997]

- **matrix method for canonical ordering** → canonical numbering
- **matrix spectrum operators** → spectral indices
- **matrix sum indices** → Wiener-type indices
- **Matthews correlation index**  $\equiv$  *Pearson similarity coefficient* → classification parameters
- **maximal binding energy** → scoring functions ( $\odot$  average binding energy)
- **maximal information content** → information content
- **maximal R indices** → GETAWAY descriptors
- **maximal R total index** → GETAWAY descriptors
- **Maximum Auto-Cross-Correlation descriptors** → autocorrelation descriptors
- **maximum bond length** → resonance descriptors ( $\odot$  RC index)

### ■ maximum common substructure (MCS)

The maximum common substructure (often “maximal common substructure”) of two compounds is the largest possible substructure that is present in both structures. The recognition of a maximum common substructure depends on the defined matching conditions; for example, two substructures are considered to be identical if all atoms and all bonds (single, double, triple, aromatic) can be matched. A further restriction can be applied concerning the number of hydrogen atoms: two nonhydrogen atoms are considered to be identical only if the number of hydrogens bonded to them is equal.

The MCS is a measure and a description of the similarity of two structures whose numerical value  $MCS$  is the number of common elements provided by the matching conditions, that is, a measure of the size of the maximum common substructure. It is commonly used in → *similarity searching* [Scsibraný and Varmuza, 1992a].

The MCS of a set of  $N$  compounds, however, may be very small or may not even exist if an exotic structure is contained in the set. Therefore, the common structural characteristics of a set of structures are better described by a set of MCSs, each of them being the MCS of a pair of structures. Such a set is obtained by determining the MCS for all the  $N(N-1)/2$  pairs of compounds; then the number  $N_i$  of occurrences of each different MCS is counted in the set. Finally, an ordered set of MCSs is obtained by a ranking function, which considers both frequency and size of the MCS:

$$R_i = (1-k) \cdot \frac{N_i}{N} + k \cdot \frac{A_i}{A^{\max}}$$

where  $A_i$  is the number of non-hydrogen atoms in  $MCS_i$  and  $A^{\max}$  is the maximum number of non-hydrogen atoms in all MCSs.  $k$  is a user-adjustable parameter (ranging between 0 and 1), which determines the different influence of the frequency and size of MCS; for  $k=1$ , only size is considered in the ranking, while for  $k=0$ , only the frequencies [Varmuza, Penchev *et al.*, 1998]. The ordered set of MCSs characterizes common and typical structural properties in the investigated set of compounds.

A measure of similarity obtained by the maximum common substructure between two compounds  $s$  and  $t$  is given by

$$SI_{st} = \frac{(A+B)_{MCS}}{(A+B)_s} \cdot \frac{(A+B)_{MCS}}{(A+B)_t}$$

where  $(A + B)$  is the sum of atoms and bonds in the maximum common substructure MCS, in the  $s$ th compound and  $t$ th compound, respectively [Durand, Pasari *et al.*, 1999]. A topological distance between the  $s$ th and  $t$ th compounds is usually defined as

$$d_{st} = (A + B)_s + (A + B)_t - 2 \cdot (A + B)_{\text{MCS}}$$

Usually the considered MCSs are connected graphs, that is, continuous bonded substructures, but disconnected substructures can also be allowed, using a corrected MCS, as for example,

$$\text{MCS}_{st} = A_{\text{MCS}} - k(N_{\text{FRAG}} - 1)$$

where  $N_{\text{FRAG}}$  is the number of disconnected fragments,  $k$  a penalty function between 0 and 1, and  $A_{\text{MCS}}$  the atom number of the MCS between the  $s$ th and  $t$ th compounds.

Therefore, the **highest scoring common substructure (HSCS)** value is a standardized variable proposed to measure the similarity between the two molecules  $s$  and  $t$  [Sheridan and Miller, 1998], defined as

$$\text{HSCS}_{st} = \frac{\text{MCS}_{st} - \text{Mean}(A_s, A_t)}{\text{Std}(A_s, A_t)}$$

where  $\text{MCS}_{st}$  is the score of the actual MCS, and  $\text{Mean}$  and  $\text{Std}$  the mean expected score and standard deviation of the MCS within a large sample of randomly selected molecules of the same size; they are calculated by regression analysis as

$$\begin{aligned}\text{Mean}(A_s, A_t) &= b_0^{\text{mean}} + b_1^{\text{mean}} \cdot \min(A_s, A_t) \\ \text{Std}(A_s, A_t) &= b_0^{\text{std}} + b_1^{\text{std}} \cdot \min(A_s, A_t)\end{aligned}$$

where  $\text{Mean}(A_s, A_t)$  is the number of atoms in the smallest molecule for a pair of randomly selected molecules of the same size as molecules  $s$  and  $t$ .  $\text{HSCS}$  values greater than 4.0 can be considered highly significant.

📖 [Cone, Venkataraghavan *et al.*, 1977; Brint and Willett, 1987a, 1987b; Stahl and Mauser, 2005; Sheridan, Hunt *et al.*, 2006; Gardiner, Gillet *et al.*, 2007]

- **maximum electrophilic superdelocalizability** → quantum-chemical descriptors (⊙ electrophilic superdelocalizability)
- **maximum–minimum path matrix**  $\equiv$  *detour-distance combined matrix* → detour matrix
- **maximum–minimum path sum** → detour matrix
- **maximum negative charge** → charge descriptors

#### ■ maximum nuclear repulsion for C–H bond index

A molecular descriptor accounting for nuclear repulsion energy between bonded carbon and hydrogen nuclei. It is defined as

$$E_{nm}^{\text{CH}} = \max_k \left( \frac{Z_{\text{C}} \cdot Z_{\text{H}}}{r_{\text{CH}}} \right)_k$$

where  $Z$  are the atomic numbers,  $r_{\text{CH}}$  the C–H bond length, and  $k$  refers to a pair of bonded carbon and hydrogen atoms [Katritzky, Sild *et al.*, 1998a]. It possibly encodes the information

about the hybridization state of carbon atoms because the C–H bond length depends on the carbon hybridization state.

- **maximum nucleophilic superdelocalizability** → quantum-chemical descriptors (⊙ nucleophilic superdelocalizability)
- **maximum path degree sequence** → detour matrix
- **maximum path frequency sequence** → detour matrix
- **maximum path matrix**  $\equiv$  *detour matrix*
- **maximum path sum** → detour matrix
- **maximum positive charge** → charge descriptors
- **MCASE**  $\equiv$  *MULTICASE* → lipophilicity descriptors (⊙ Klopman hydrophobic models)
- **MCB index** → multiple bond descriptors
- **McClelland resonance energy** → delocalization degree indices
- **McConnaughey similarity coefficient** → similarity/diversity (Table S9)
- **McFarland model** → Hansch analysis
- **Mc Gowan's characteristic volume** → volume descriptors
- **MDDM**  $\equiv$  *Main Distance-Dependent Matrices* → 4D-Molecular Similarity Analysis
- **MDE vector**  $\equiv$  *molecular distance-edge vector*
- **MDL keys**  $\equiv$  *MACCS keys* → substructure descriptors (⊙ structural keys)
- **mean absolute deviation** → statistical indices (⊙ indices of dispersion)
- **mean difference** → statistical indices (⊙ indices of dispersion)
- **mean distance degree deviation** → distance matrix
- **mean extended local information on distances** → topological information indices
- **mean information content** → information content
- **mean information content on the adjacency equality** → topological information indices
- **mean information content on the adjacency magnitude** → topological information indices
- **mean information content on the distance degree equality** → topological information indices
- **mean information content on the distance degree magnitude** → topological information indices
- **mean information content on the distance equality** → topological information indices
- **mean information content on the distance magnitude** → topological information indices
- **mean information content on the edge adjacency equality** → topological information indices
- **mean information content on the edge adjacency magnitude** → topological information indices
- **mean information content on the edge-cycle matrix elements equality** → topological information indices
- **mean information content on the edge-cycle matrix elements magnitude** → topological information indices
- **mean information content on the edge cyclic degree equality** → topological information indices
- **mean information content on the edge cyclic degree magnitude** → topological information indices
- **mean information content on the edge degree equality** → topological information indices

- **mean information content on the edge degree magnitude** → topological information indices
- **mean information content on the edge distance degree equality** → topological information indices
- **mean information content on the edge distance degree magnitude** → topological information indices
- **mean information content on the edge distance equality** → topological information indices
- **mean information content on the edge distance magnitude** → topological information indices
- **mean information content on the edge equality** → information connectivity indices
- **mean information content on the edge magnitude** → information connectivity indices
- **mean information content on the incidence matrix** → incidence matrices (⊙ vertex-edge incidence matrix)
- **mean information content on the leverage magnitude** → GETAWAY descriptors
- **mean information content on the vertex-cycle matrix elements equality** → topological information indices
- **mean information content on the vertex-cycle matrix elements magnitude** → topological information indices
- **mean information content on the vertex cyclic degree equality** → topological information indices
- **mean information content on the vertex cyclic degree magnitude** → topological information indices
- **mean information content on the vertex degree equality** → topological information indices
- **mean information content on the vertex degree magnitude** → topological information indices
- **mean information index on atomic composition** → atomic composition indices
- **mean information index on molecular conformations** → information index on molecular conformations
- **mean local information on distances**  $\equiv$  *vertex distance complexity* → topological information indices
- **mean overcrossing number** → polymer descriptors
- **mean polarizability** → electric polarization descriptors
- **mean Randić branching index**  $\equiv$  *mean Randić connectivity index* → connectivity indices
- **mean Randić connectivity index** → connectivity indices
- **mean square distance index** → distance matrix
- **mean square error** → regression parameters
- **mean topological charge index** → topological charge indices
- **mean Wiener index** → Wiener index
- **median** → statistical indices (⊙ indices of central tendency)
- **median effective dose** → biological activity indices (⊙ pharmacological indices)
- **median inhibitory concentration** → biological activity indices (⊙ toxicological indices)
- **median lethal dose** → biological activity indices (⊙ toxicological indices)
- **MEDNE descriptors** → MARCH-INSIDE descriptors



### ■ MEDV-13 descriptor

MEDV, namely, the **Molecular Electronegativity Distance Vector**, is a vectorial molecular descriptor comprising of 91 terms encoding information about relative electronegativities, represented by modified  $\rightarrow$  *E-state indices*, and topological distances between all the possible pairs of 13 atom types (MEDV-13) [Liu, Cai *et al.*, 2000; Liu, Yin *et al.*, 2001b, 2002a, 2002b; Sun, Zhou *et al.*, 2004].

To generate MEDV descriptors, first the atoms in the molecule are assigned to one of the 13 defined atom types (Table M5). These are distinguished on the basis of the chemical element of the most occurring atoms in organic molecules, the number of bonded non-hydrogen atoms, that is, the  $\rightarrow$  *vertex degree*, which reflects the local topological environment of an atom, and the number of valence electrons of the atom, which is used to distinguish atoms with the same vertex degree but different chemical element.

**Table M5** MEDV-13 atom types.  $\delta$  is the atom vertex degree.

Type	Chemical element	$\delta$	Type	Chemical element	$\delta$	Type	Chemical element	$\delta$
1	C	1	6	N, P	2	10	O, S	2
2	C	2	7	N, P	3	11	S	3
3	C	3	8	P	4	12	S	4
4	C	4	9	O, S	1	13	F, Cl, Br, I	1
5	N, P	1						

Moreover, for each *i*th atom in the molecule, a modified  $\rightarrow$  *E-state index*, used as a measure of relative electronegativity, is calculated as

$$S_i^* = I_i^* + \Delta I_i^* = I_i^* + \sum_{j=1}^A \frac{I_i^* - I_j^*}{d_{ij}^2}$$

where  $d_{ij}$  is the  $\rightarrow$  *topological distance* between *i*th and *j*th atoms and  $I^*$  is a modified  $\rightarrow$  *intrinsic state* defined as

$$I_i^* = \sqrt{\frac{Z_i^v}{4}} \cdot \frac{(2/L_i)^2 \cdot \delta_i^b + 1}{\delta_i}$$

where  $Z_i^v$  is the number of valence electrons and the  $\rightarrow$  *valence vertex degree*  $\delta_i^v$  is replaced by the  $\rightarrow$  *bond vertex degree*  $\delta_i^b$ , which accounts for atom connectedness and bond multiplicity.

On the basis of the values of the vertex degree  $\delta$ , the bond vertex degree  $\delta^b$ , and the modified intrinsic state  $I^*$ , atoms can be classified into 43 types, called atomic attributes, which are proposed by analogy with the *E-state* atom types of Kier and Hall. In defining the atomic attributes, a conjugated system indicator (CSI) is used instead of the aromatic system indicator to distinguish atoms located at different positions of a conjugated system because they have different effects on the molecule (Table M6).

**Table M6** MEDV-13 atomic attributes:  $\delta^b$ , bond vertex degree;  $\delta$ , vertex degree;  $I^*$ , modified intrinsic state.

No.	Atom type	$\delta^b$	$\delta$	$I^*$	No.	Atom type	$\delta^b$	$\delta$	$I^*$
1	—CH <sub>3</sub>	1	1	2.000	23	>N—	3	3	1.491
2	—CH <sub>2</sub> —	2	2	1.500	24	=NH	2	1	3.354
3	>CH—	3	3	1.333	25	=N—	3	2	2.236
4	>C<	4	4	1.250	26	≡N	3	1	4.472
5	=CH <sub>2</sub>	2	1	3.000	27	aNH	1.5	1	2.795
6	=CH—	3	2	2.000	28	aN—	2.5	2	1.957
7	=C<	4	3	1.667	29	aNa	3	2	2.236
8	=C=	4	2	2.500	30	=N—(=)	5	3	2.236
9	≡CH	3	1	4.000	31	—SH	1	1	1.769
10	≡C—	4	2	2.500	32	—S—	2	2	1.157
11	aCH <sub>2</sub>	1.5	1	2.500	33	=S	2	1	2.313
12	aCH—	2.5	2	1.750	34	>S=	4	3	1.134
13	aC<	3.5	3	1.500	35	≥S≤	6	4	1.123
14	aCHa	3	2	2.000	36	—F	1	1	2.646
15	aCa—	4	3	1.667	37	—Cl	1	1	1.911
16	aaCa	4.5	3	1.833	38	—Br	1	1	1.654
17	—OH	1	1	2.449	39	—I	1	1	1.534
18	=O	2	1	3.674	40	—PH <sub>2</sub>	1	1	1.615
19	—O—	2	2	1.837	41	—PH—	2	2	1.056
20	aO	1.5	1	3.062	42	>P—	3	3	0.870
21	—NH <sub>2</sub>	1	1	2.236	43	≥P<	5	4	0.901
22	—NH—	2	2	1.677					

The symbol “a” refers to a bond in any conjugated system, including aromatic systems. Data from Liu, Yin et al. [Liu, Yin et al., 2002b].

Finally, for each combination of two atom types ( $u, v$ ), a single molecular descriptor  $h(u, v)$  is calculated as

$$h(u, v) = \sum_{i \in u} \sum_{j \in v} \frac{S_i^* \cdot S_j^*}{d_{ij}^2} \quad u, v = 1, 2, \dots, 13$$

where the first sum runs over all the atoms of type  $u$  and the second sum on the atoms of type  $v$ ,  $S^*$  is the modified  $E$ -state index, and  $d_{ij}$  the topological distance between vertices  $v_i$  and  $v_j$ .

Since 13 different atom types are considered, a total of 91 ((13 × 14)/2) different molecular descriptors result, which constitute the final MEDV-13 vector.

An extension of MEDV-13 descriptor, called **Molecular Holographic Distance Vector** (MHDV) was further proposed to describe the structure of molecules containing several heteroatoms and multiple bonds as well as → *peptide sequences* [Liu, Yin et al., 2001a].

➤ **melting point** → physico-chemical properties

### ■ Membrane Interaction QSAR analysis (MI-QSAR)

MI-QSAR is a methodology that combines classic molecular descriptors with membrane–solute intermolecular properties of compounds to model chemically and structurally diverse compounds interacting with cellular membranes [Kulkarni, 1999; Kulkarni and Hopfinger, 1999; Iyer, Mishra et al., 2002]. MI-QSAR aims at providing insight into the mechanism of skin

penetration, capturing features of cellular lateral transverse transport involved in the overall skin penetration process of organic compounds.

The MI-QSAR method is receptor based, in effect the assumption made is that the phospholipid regions of a cellular membrane constitute the “receptor.” The receptor is usually constructed as a monolayer from the phospholipids that comprise the cell membrane of the system of interest; for instance, a single dimyristoylphosphatidylcholine (MDPC) molecule is selected as the model phospholipid and an assembly of 25 DMPC molecules ( $5 \times 5 \times 1$ ) in ( $x, y, z$ ) directions, respectively, is used as the model membrane monolayer.

Molecular dynamics simulations (MDSs) on the model membrane are initially carried out to allow for structural relaxation and distribution of the kinetic energy over the monolayer. Then, each molecule is inserted at three different positions in the monolayer, with the most polar group of the molecule “facing” toward the headgroup region of the monolayer. The three positions are (a) headgroup region, (b) center region of the aliphatic chains, and (c) tail region of the aliphatic chains. Separate MDSs are performed for each compound in each of the three trial positions, and the most favorable orientation and location of the compound in the monolayer is determined. Then, the MI-QSAR descriptors are calculated. These are divided into (a) general **intramolecular solute descriptors**, (b) **intermolecular solute–membrane interaction descriptors**, and (c) **solute aqueous dissolution and solvation descriptors** [Iyer, Tseng *et al.*, 2007]. The first set of solute properties are the molecular descriptors calculated by the program Cerius and examples are given in Table M7. The intermolecular solute–membrane interaction descriptors (set b) are derived from MDS trajectories; these particular intermolecular descriptors are calculated using the most stable solute–membrane geometry realized from MDS sampling of the three initial positions for each compound (Table M8). Solute aqueous dissolution and solvation descriptors (set c), although computed by intramolecular computational methods, are intermolecular properties, the first three relating to solute solvation and the last three to solute dissolution (Table M9).

**Table M7** MI-QSAR general intramolecular solute descriptors (set a).


Symbol	General intramolecular solute descriptors
HOMO	Highest occupied molecular orbital energy
LUMO	Lowest unoccupied molecular orbital energy
$\mu$	Dipole moment
$V_m$	Molecular volume
SA	Molecular surface area
$\rho$	Density
MW	Molecular weight
MR	Molecular refractivity
HBA	Number of hydrogen-bond acceptors
HBD	Number of hydrogen-bond donors
RBN	Number of rotatable bonds
CPSA	Charged partial surface area descriptors
$^m\chi$ and $^m\kappa$	Kier and Hall connectivity and shape descriptors
$R_G$	Radius of gyration
$I$	Principal moment of inertia
PSA	Polar surface area
$S_{\text{conf}}$	Conformational entropy
$q_x$	Partial atomic charge densities

**Table M8** MI-QSAR intermolecular solute–membrane interaction descriptors (set b).

Symbol	Intermolecular solute–membrane interaction descriptors
$\langle F(\text{total}) \rangle$	Average total free energy of interaction of the solute and membrane
$\langle E(\text{total}) \rangle$	Average total interaction energy of the solute and membrane
$E_{\text{INTER}}(\text{total})$	Interaction energy between the solute and the membrane at the total intermolecular system minimum potential energy
$E_{\text{XY}}(Z)_E$	$Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, torsion, and combinations thereof energies at the total intermolecular system minimum potential energy. X, Y can be the solute, S, and/or membrane, M, and if $E = \text{free}$ , then $X = Y = S$ and the energies are for the solute not in the membrane, but isolated by itself.
$\Delta E_{\text{XY}}(Z)$	Change in the $Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, torsion, and combinations thereof energies due to the uptake of the solute to the total intermolecular system minimum potential energy
$E_{\text{TT}}(Z)$	$1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen bonding, torsion, and combinations thereof energies of the total (solute and membrane model) intermolecular minimum potential energy
$\Delta E_{\text{TT}}(Z)$	Change in the $Z = 1,4$ -nonbonded, general van der Waals, electrostatic, hydrogen-bonding, and combinations thereof of the total (solute and membrane model) intermolecular minimum potential energy
$\Delta S$	Change in entropy of the membrane due to the uptake of the solute
$S$	Absolute entropy of the solute–membrane system
$\Delta \rho$	Change in density of the model membrane due to the permeating solute
$\langle d \rangle$	Average depth of the solute molecule from the membrane surface

**Table M9** MI-QSAR solute aqueous dissolution and solvation descriptors (set c)

Symbol	Solute aqueous dissolution and solvation descriptors
$F(\text{H}_2\text{O})$	Aqueous solvation free energy
$F(\text{oct})$	1-Octanol solvation free energy
$\log P$	1-Octanol/water partition coefficient
$E(\text{coh})$	Cohesive packing energy of the solute molecules
$T_{\text{M}}$	Hypothetical crystal-melt transition temperature of the solute
$T_{\text{G}}$	Hypothetical glass transition temperature of the solute

 [Kulkarni, Han *et al.*, 2002; Santos-Filho, Hopfinger *et al.*, 2004; Li, Liu *et al.*, 2005]

- **MEP**  $\equiv$  *Molecular Electrostatic Potential*  $\rightarrow$  quantum-chemical descriptors
- **Merrifield–Simmons bond order**  $\rightarrow$  symmetry descriptors ( $\odot$  Merrifield–Simmons index)
- **Merrifield–Simmons index**  $\rightarrow$  symmetry descriptors
- **mesomeric effect**  $\rightarrow$  electronic substituent constants
- **Method of Ideal Symmetry**  $\rightarrow$  geometry matrix
- **Meyer anchor sphere volume**  $\rightarrow$  size descriptors
- **Meyer–Richards similarity index**  $\rightarrow$  quantum similarity
- **Meyer visual descriptor of globularity**  $\rightarrow$  shape descriptors

➤ **Meylan–Howard hydrophobic model**  $\equiv$  KOWWIN  $\rightarrow$  lipophilicity descriptors

### ■ Mezey 3D shape analysis

This is an approach to shape analysis and comparison of molecules based on algebraic topological methods suitable for algorithmic, nonvisual analysis, and coding of molecular shapes by  $\rightarrow$  *computational chemistry* [Mezey, 1985, 1993c]. Several topological methods were proposed for the analysis and coding of molecular shapes, most of them based on the concept of  $\rightarrow$  *molecular surface*. In particular, the **Shape Group Method** (SGM) is a topological shape analysis technique of any, almost everywhere twice continuously, differentiable 3D functions (e.g.,  $\rightarrow$  *electron density*). A detailed description of these methods is given in Mezey [Mezey, 1990c].

In general, topological methods are based on subdividing the molecular surface into domains, according to physical or geometrical conditions [Mezey, 1991b].

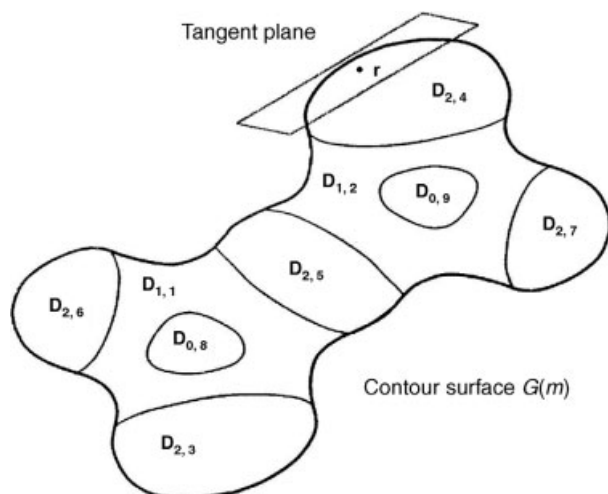
For example, if two molecular surfaces of the same molecule are considered to be based on two different physical properties such as the  $\rightarrow$  *electron isodensity contour surface*  $G_1(m_1)$  and the  $\rightarrow$  *molecular electrostatic potential contour surface*  $G_2(m_2)$ , then the former can be subdivided into domains according to electrostatic potential values. The interpenetration of the two surfaces provides several closed loops on the isodensity contour surface; these loops are sets of points of  $G_1$  with equal value  $m_2$  of  $\rightarrow$  *molecular electrostatic potential* (MEP), and define the boundaries of the surface  $G_1$  regions that are characterized by MEP values either greater or lower than the threshold value  $m_2$  for all the points in the region. Using different threshold values of MEP, several different electrostatic potential ranges can be mapped on the isodensity surface; these ranges define a subdivision of  $G_1(m_1)$  into domains whose topological interrelations can be characterized by a numerical code, which provides a shape characterization of the molecule.

Applying the same approach to several molecules, the similarity of their shape can be searched for by comparing the topological relations among the corresponding domains on the molecular surfaces.

An alternative approach to domain subdivisions of the molecular surface is based on local curvature properties. It is applicable only to differentiable molecular surfaces such as contour surfaces, for example, the electron isodensity contour surface  $G(m)$ ,  $m$  being the threshold value defining the contour surface.

The local curvature properties of the surface  $G(m)$  in each point  $\mathbf{r}$  of the surface are given by the eigenvalues of the local Hessian matrix. Moreover, for a defined reference curvature  $b$ , the number  $\mu(\mathbf{r}, b)$  is defined as the number of local canonical curvatures (Hessian matrix eigenvalues) that are less than  $b$ . Usually  $b$  is chosen equal to zero and therefore the number  $\mu(\mathbf{r}, 0)$  can take values 0, 1, or 2 indicating that at the point  $\mathbf{r}$  the molecular surface is locally concave, saddle type, or convex, respectively. The three disjoint subsets  $A_0$ ,  $A_1$ , and  $A_2$  are the collections of the surface points at which the molecular surface is locally concave, saddle type, or convex, respectively; the maximum connected components of these subsets  $A_0$ ,  $A_1$ , and  $A_2$  are the surface domains denoted by  $D_{0,k}$ ,  $D_{1,k}$ , and  $D_{2,k}$  where the index  $k$  refers to an ordering of these domains, usually according to decreasing surface size (Figure M1).

The mutual arrangement of the domains along the molecular surface can be represented by the topological neighborhood relation, two domains being neighbors if they have a common boundary line. Therefore, a symmetric square *shape matrix* can be built where the rows and columns represent the surface domains; the off-diagonal entries of the shape matrix can be equal to 1 if the considered domains are adjacent and 0 otherwise, the diagonal entries are equal to the number  $\mu$  (usually 0, 1, or 2) depending on the domain type. If the surface domains are



**Figure M1** A subdivision of a molecular contour surface  $G(m)$ , based on local curvature properties. The contour surface is subdivided into locally concave ( $D_{0,k}$ ), saddle-type ( $D_{1,k}$ ), and locally convex ( $D_{2,k}$ ) domains with respect to the local tangent plane in each point  $r$ . The second index  $k$  refers to an ordering of these domains according to decreasing surface area.

listed in increasing order according to the size of their surface areas, then the shape matrix encodes both shape and size information. Moreover, from the shape matrix, the corresponding *shape graph* can be derived as additional tool for shape characterization of the molecular surface with respect to the reference curvature  $b$ ; in fact, shape matrix and shape graph are topological descriptors of molecular surface shape regarded as 3D topological shape codes. They are particularly useful in quantifying the similarity of shapes of the different molecules; the comparison of molecular shape is reduced to a comparison of shape matrices or graphs.

It is worth noting that the topological relations among the domains are invariant within a given range of different molecular conformations. In effect, change in molecular conformation can lead to change in size, location, and even in domain existence, but for certain conformational changes, the existence and the mutual neighborhood relations of the domains remain invariant.

Considering a finite number of threshold values  $m$ , a set of contour surfaces  $G(m)$  is studied for each molecule combined with a set of reference curvature values  $b$ . Therefore, for each pair  $(m, b)$  of parameters, the curvature domains  $D_0(m, b)$ ,  $D_1(m, b)$ , and  $D_2(m, b)$  are computed and the truncation of contour surfaces  $G(m)$  is performed by removing all curvature domains  $D_\mu$  of specified type  $\mu$  (in most applications  $\mu = 2$ ) from the contour surface, thus obtaining a truncated surface  $G(m, \mu)$  for each  $(m, b)$  pair. For most small changes of the parameter values, the truncated surfaces remain topologically equivalent and only a finite number of equivalence classes are obtained for the entire range of  $a$  and  $b$  values.

In the next step, the *shape groups* of the molecular surface, that is, the zero-, one-, and two-dimensional algebraic homology groups of the truncated surfaces, are computed. The zero-, one-, and two-dimensional **Betti numbers**  $b_\mu^0(m, b)$ ,  $b_\mu^1(m, b)$ , and  $b_\mu^2(m, b)$  are the ranks of these zero-, one-, and two-dimensional homology groups, that is, the shape groups. They are a list of

topologically invariant numbers and represent a numerical shape code of the molecule, providing a detailed shape characterization of the distribution of the property used to define the molecule surface. In practice, the Betti numbers of 1D shape groups give the most important chemical information and the analysis is often restricted to this class of Betti numbers.

📖 [Mezey, 1987a, 1987b, 1988a, 1988b, 1988c, 1989, 1990b, 1991c, 1992, 1993a, 1993b, 1993d, 1994, 1996, 1997a, 1999; Arteca and Mezey, 1987, 1988a, 1988b, 1989; Arteca, Jammal *et al.*, 1988a, 1988b; Walker, Maggiora *et al.*, 1995]

- **ME-MFP descriptors** → substructure descriptors (⊙ structural keys)
- **MFP descriptors** ≡ *Mini-FingerPrints* → substructure descriptors (⊙ structural keys)
- **MHDV** ≡ *molecular holographic distance vector* → MEDV-13 descriptor
- **micelle–water partition coefficient** → physico-chemical properties (⊙ partition coefficients)
- **MIC index** ≡ *Modified Information Content index* → indices of neighborhood symmetry
- **Migration Index** → chromatographic descriptors (⊙ capacity factor)
- **MINBID** → ID numbers
- **MINCID** → ID numbers
- **Mini-FingerPrints** → substructure descriptors (⊙ structural keys)
- **minimal spanning tree** → graph
- **minimal steric difference** → minimal topological difference

### ■ Minimal Topological Difference (MTD)

Among the → *hyperstructure-based QSAR techniques*, the MTD method is based on the approximate atom-per-atom superimposition of the  $n$  molecules of a → *data set* to build a → *hypermolecule* (i.e., 3D → *hyperstructure*): hydrogen atoms, small differences in atomic positions, bond lengths and bond angles are neglected. The  $S$  vertices of the hypermolecule correspond to the positions of the data set molecule atoms [Simon, Dragomir *et al.*, 1973; Simon and Szabadai, 1973b; Simon, Chiriac *et al.*, 1984].

The basic idea is that the geometry of the hypermolecule is related to the geometry of the receptor binding site, and molecule steric affinity to the binding site is obtained by comparing geometry of the molecule and that of the hypermolecule. Moreover, to represent the active regions of the hypermolecule, vertices within the → *binding site cavity* (*cavity vertices*) are labeled with a parameter  $\epsilon = -1$ , vertices in the cavity walls (*wall vertices*) with  $\epsilon = +1$ , and vertices in the external part of the cavity, that is, in the aqueous solution, with  $\epsilon = 0$ .

In this way, **MTD descriptors** measuring the → *steric misfit* between the binding site cavity and the considered molecules are calculated as

$$MTD_i = c + \sum_{s=1}^S \epsilon_s \cdot I_{is}$$

where the subscript  $i$  refers to the considered molecule;  $c$  is the total number of cavity vertices of the hypermolecule and should be a measure of the volume of the binding site cavity;  $S$  is the total number of hypermolecule vertices;  $I_{is}$  is a binary variable for the  $s$ th hypermolecule vertex equal to 1 if the  $i$ th molecule occupies the  $s$ th vertex with an atom, otherwise it equals zero.

The  $MTD_i$  descriptor is a measure of steric misfit of the  $i$ th molecule with respect to the receptor cavity and is equal to the number of unoccupied cavity vertices plus the number of

occupied wall vertices; it can be considered both a  $\rightarrow$  *steric descriptor* and a  $\rightarrow$  *differential descriptor*.

For a molecule coincident with the active region of the hypermolecule, it can be observed that  $c$  atoms are located at the hypermolecule cavity vertices ( $\epsilon = -1$ ), and no atoms coincide with the hypermolecule wall vertices, that is, there is no steric misfit and  $MTD = 0$ .

The **MTD model** is obtained by including  $MTD$  descriptors in the  $\rightarrow$  *Hansch model*:

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} - b' \cdot MTD_i$$

where  $\Phi$  are selected  $\rightarrow$  *physico-chemical properties* of a Hansch-type model and the sign minus of the  $MTD$  coefficient indicates the detrimental contribution to the activity due to steric misfit.

The final optimal set of  $\epsilon_1, \epsilon_2, \dots, \epsilon_S$  values (*receptor map*) is searched for according to an optimization procedure, starting from random or arbitrary assignment of hypermolecule vertices to cavity, wall, or external regions. To each set of  $\epsilon_s$  values corresponds a set of  $MTD$  descriptors and a set of calculated responses. The optimization procedure is based on the maximization of the correlation of the biological responses calculated by the  $MTD$  model with the experimental responses.

The **MTD-MC method** is a modified version of the  $MTD$  method, accounting for the existence of several low-energy conformations of molecules used to derive the hypermolecule by overlap. Each molecule is described by a vector of binary variables  $I_{i(k)s}$  equal to one if the  $s$ th vertex of the hypermolecule is occupied by the  $i$ th molecule in the  $k$ th conformation. If more than one low-energy conformation is allowed for a molecule, the conformations considered will be the one that best fits the binding site cavity, that is, the one with the lowest  $MTD$  value:

$$MTD_i = \min_k \left( c + \sum_{s=1}^S \epsilon_s \cdot I_{i(k)s} \right)$$

where the minimum is chosen over all of the considered conformations of the  $i$ th molecule.

A modification of the  $MTD$  approach that also acts on biological responses, called the **MTD-ADJ method**, was proposed to improve the performance of the modeling power of the method, accounting for the relative contribution of the active conformation to the activity of each compound [Sulea, Kurunczi *et al.*, 1998].

Let  $C_A$  be the concentration of active conformation A, the following relationships hold

$$C_A = \alpha_A \cdot C \quad y^{\text{exp}} = -\log C \quad y^{\text{adj}} = -\log(\alpha_A C) = y^{\text{exp}} - \log \alpha_A$$

where  $C$  is the total concentration,  $y^{\text{exp}}$  and  $y^{\text{adj}}$  the experimental and adjusted biological responses. The factor  $\alpha_A$  for each conformation is calculated by the Boltzmann distribution:

$$\alpha_A = \frac{g_A \cdot \exp^{-E_A/RT}}{\sum_{k=1}^N g_k \cdot \exp^{-E_k/RT}}$$

where  $E$  are the calculated total conformational energies,  $g$  the degeneration degree of the conformational energy levels, and  $R$  and  $T$  the gas constant and the absolute temperature, respectively.



For each conformation of each compound, the corresponding *MTD* value and adjusted biological response are calculated. By using the optimization and  $\rightarrow$  *validation techniques* for *MTD* model, each compound's conformation that best fits the adjusted response is retained and should be considered the active conformation of the compound. If more than one conformation is selected for the same compound, all these conformers have the same *MTD* values, while the adjusted response is calculated considering the conformer contributions to the total population, that is,  $\Sigma\alpha$ .

The *MTD-ADJ* method provides additional information concerning the active conformations of the compounds.

The **Minimal Steric Difference** (*MSD*) method is the first version of the *MTD* approach, based on the comparison of each molecule with the molecule with the highest biological activity in the data set, taken as the  $\rightarrow$  *reference structure* instead of the hypermolecule. The assumption is that the most active molecule fits into the binding site best [Simon and Szabadai, 1973b].

$MSD_i$  is a descriptor of steric misfit defined as the number of nonoverlapping non-hydrogen atoms for the maximal superimposition of the  $i$ th molecule to the reference molecule. *MSD* coincides with *MTD* when simple minimal steric differences are calculated with respect to the most active compound, no external atoms are considered in the hypermolecule, and the number of hypermolecule cavity vertices corresponds to the atoms of the reference molecule.

The **Monte Carlo version of MTD** (*MCD*) is a modification of the *MSD* approach, where the  $MCD_i$  descriptor is calculated as the nonoverlapping volume ( $NOV_i$ ) of the  $i$ th molecule with respect to the reference molecule and the reference molecule with respect to the  $i$ th molecule [Motoc, Holban *et al.*, 1977].

The two superimposed molecules are included within a cube with volume  $V$  and a large number of points  $N$  is randomly dispersed within the cube. The **nonoverlapping volume**  $NOV_i$  of the  $i$ th molecule is calculated as

$$NOV_i = V \cdot \left( \frac{N_{REF} + N_i}{N} \right)$$

where  $N_{REF}$  is the number of points falling into the  $\rightarrow$  *van der Waals molecular surface* of the reference molecule but not into that of the  $i$ th molecule, and  $N_i$  is the number of points falling into the van der Waals envelope of the  $i$ th molecule but not into that of the reference molecule;  $N$  is the total number of points randomly distributed throughout volume  $V$  of the cube.

A further modification of the *MTD* approach is called **Steric Interactions in Biological Systems** (*SIBIS*), where **attractive steric effects** (*SMDC*) and **repulsive steric effects** (*SMDW*) are considered separately [Motoc and Dragomir, 1981; Motoc, 1984a, 1984b]. Moreover, the optimization procedure searching for the receptor map, that is, the optimal set of  $\epsilon_s$  values, is modified by the introduction of connectivity restrictions, where all the cavity vertices have to form a single topological connected network, that is, the receptor cavity is not fragmented into several subcavities.

The two steric contributions are defined as

$$SMDC_i = \sum_{s=1}^{S_{cav}} b'_{is} \cdot I_{is} \quad SMDW_i = \sum_{s=1}^{S_{wall}} b''_{is} \cdot I_{is}$$

where  $b_{is}$  are correction factors, accounting for the size of the atom of the  $i$ th molecule in the  $s$ th position of the hypermolecule;  $I_{is}$  is a binary variable for the  $s$ th hypermolecule vertex equal to 1 if the  $i$ th molecule occupies the  $s$ th vertex with an atom, and equal to zero otherwise. The first descriptor  $SMDC$  is a sum running over all hypermolecule cavity positions  $S_{cav}$  ( $\epsilon_s = -1$ ) and the second descriptor  $SMDW$  a sum over all the hypermolecule cavity wall positions  $S_{wall}$  ( $\epsilon_s = +1$ ).

Therefore, the **SIBS model** is defined as

$$\hat{y}_i = b_0 + \sum_{j=1}^J b_j \cdot \Phi_{ij} + b' \cdot SMDC_i - b'' \cdot SMDW_i$$

where  $\Phi$  are selected physico-chemical properties of the Hansch model; the plus sign of the  $SMDC$  coefficient indicates a favorable contribution (attractive steric effects) to activity and sign minus of the  $SMDW$  coefficient a detrimental contribution (repulsive steric effects).

Moreover, similar to the  $\rightarrow$  *Molecular Field Topology Analysis* (MFTA), a further variant of the  $MTD$  approach was proposed with the name **MTD-PLS method** [Oprea, Kurunczi *et al.*, 2001; Kurunczi, Olah *et al.*, 2002], intended as a simple (topologically based)  $\rightarrow$  *scoring function* that could be useful in the absence of relevant receptor information. In particular, a “chemically intuitive” function is obtained forcing  $MTD$ - $PLS$  coefficients to assume only negative (or zero) values for fragmental volume descriptors and positive (or zero) values for fragmental hydrophobicity descriptors.

📖 [Simon and Szabadai, 1973a; Simon, 1974, 1993; Simon, Holban *et al.*, 1976; Simon, Chiriac *et al.*, 1976; Simon, Badilescu *et al.*, 1977; Popoviciu, Holban *et al.*, 1978; Balaban, Chiriac *et al.*, 1980; Motoc, 1983b; Balaban *et al.*, 1985; Balaban, Niculescu-Duvaz *et al.*, 1987; Magee, 1991; Simon and Bohl, 1992; Ciubotariu, Deretey *et al.*, 1993; Oprea, Ciubotariu *et al.*, 1993; Fabian, Timofei *et al.*, 1995; Muresan, Bologa *et al.*, 1995; Sulea, Kurunczi *et al.*, 1995; Mracec, Mracec *et al.*, 1996; Timofei, Kurunczi *et al.*, 1996; Oprea, Kurunczi *et al.*, 1997; Polanski, 1997; Hadaruga, Muresan *et al.*, 1999; Ciubotariu, Grozav *et al.*, 2001; Ciubotariu, Gogonea *et al.*, 2001; Minailiuc and Diudea, 2001; Timofei, Kurunczi *et al.*, 2001; Thakur, Thakur *et al.*, 2004b; Mracec, Juchel *et al.*, 2006]

- **minimum–maximum path matrix**  $\equiv$  *distance-detour combined matrix*  $\rightarrow$  detour matrix
- **minimum path matrix**  $\equiv$  *distance matrix*
- **Minkowski distance**  $\rightarrow$  similarity/diversity (☉ Table S7)
- **Minoli–Bonchev complexity index**  $\rightarrow$  molecular complexity
- **Minoli complexity index**  $\rightarrow$  molecular complexity
- **MI-QSAR**  $\equiv$  *Membrane Interaction QSAR analysis*
- **misclassification risk**  $\rightarrow$  classification parameters
- **MIS indices**  $\rightarrow$  molecular geometry
- **mixed CoMFA approach**  $\rightarrow$  comparative molecular field analysis
- **mixed CoMFA model**  $\rightarrow$  comparative molecular field analysis
- **MLOGP**  $\rightarrow$  lipophilicity descriptors
- **MLSER**  $\equiv$  *modified LSER*  $\rightarrow$  Linear Solvation Energy Relationships
- **MmPS topological index**  $\equiv$  *detour-Wiener combined index*  $\rightarrow$  detour matrix
- **M/M quotient matrix**  $\rightarrow$  biodescriptors (☉ DNA sequences)

- **MNA descriptors**  $\equiv$  *Multilevel Neighborhoods of Atoms descriptors*  $\rightarrow$  substructure descriptors ( $\odot$  fingerprints)
- **MobyDigs software**  $\rightarrow$  DRAGON descriptors
- **mode**  $\rightarrow$  statistical indices ( $\odot$  indices of central tendency)

### ■ model complexity

Model complexity is an important parameter to compare different QSAR/QSPR models. Moreover, the prediction power of a model is inversely related to its complexity, when complexity is unnecessary increased.

In general, model complexity is related to the number of variables selected for modeling purposes. Let  $\mathbf{I}$  be the vector of length  $p$ , where  $p$  is the total number of variables, constituted of  $p$  binary variables. Each variable takes a value equal to zero ( $I_j = 0$ ) if the  $j$ th variable is not in the model and a value equal to one ( $I_j = 1$ ) if the  $j$ th variable is in the model.

The general problem of searching for the best set of variables ( $\mathbf{I}^*$  vector) can be faced with by two different approaches: methods for  $\rightarrow$  *variable reduction* and methods for  $\rightarrow$  *variable selection*. The first group of methods allows selection of variables by inner relationships among the  $p$  variables  $\mathbf{x}_j$ :

$$\mathbf{I} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$$

while the second group of methods by considering the relationships among the variables  $\mathbf{x}_j$  and the response  $y$  to be modeled:

$$\mathbf{I} = f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p; y)$$

In the first case, attention is paid to excluding variables carrying low or redundant information, in the second, to excluding variables, which are not functionally related to the studied response. In the latter, besides the exclusion of specific variables, one can condense the information from all the original variables into a few significant latent variables (linear combinations) by methods such as *Principal Component Regression* and *Partial Least Squares regression*.

The main measures of model complexity are reported below.

#### • number of terms in the model

The simplest definition of model complexity is based on the number of terms in the model or, in other words, the model complexity is made up by the number of model variables from Ordinary Least Squares regression ( $cp\mathbf{x} = p$ ), the number  $M$  of significant principal components from Principal Component Regression ( $cp\mathbf{x} = M$ ), and the number of significant latent variables from Partial Least Squares regression ( $cp\mathbf{x} = M$ ).

#### • standardized regression coefficients sum

Model complexity is defined as the sum of standardized regression coefficients:

$$cp\mathbf{x} = \sum_{j=1}^p |b'_j| = \sum_{j=1}^p \left| \frac{b_j \cdot s_j}{s_y} \right|$$

where  $b'_j$  is the  $j$ th standardized regression coefficient,  $b_j$  the ordinary regression coefficient,  $s_y$  and  $s_j$  the standard errors of the response and  $j$ th variable, respectively, and  $p$  the total number of model variables.

• **information content ratio**

Model complexity is defined as the ratio between the **multivariate entropy**  $S_X$  of the data matrix  $\mathbf{X}$  ( $n$  objects and  $p$  variables) of the model and  $\rightarrow$  *Shannon's entropy*  $H_Y$  of the response vector  $\mathbf{y}$ , thus also accounting for the information content of the response  $\mathbf{y}$  [Todeschini and Consonni, 2000]:

$$cpx = \frac{S_X}{H_Y} \quad 0 \leq cpx \leq \frac{p \cdot \log_2 n}{H_Y} \leq p$$

where  $H_Y$  and  $S_X$  are defined as

$$H_Y = - \sum_k \frac{n_Y}{n} \log_2 \frac{n_Y}{n} \quad S_X = [1 + (p-1)(1-K)] \cdot \frac{\sum_{j=1}^p H_j}{p}$$

where  $k$  runs on the different equivalence classes for  $\mathbf{y}$  and  $n_Y$  is the number of equal  $y$  values;  $H_j$  is the Shannon entropy of the  $j$ th variable;  $p$  is the total number of variables and  $n$  is the total number of objects.  $K$  is the  $\rightarrow$  *multivariate  $K$  correlation index*.

When all the  $y$  and  $x$  values (for each  $j$ th variable) are different (i.e., the Shannon's entropy of each variable is  $\log_2 n$ ), the model complexity depends only on the total number  $p$  of model variables and the  $K$  correlation in the matrix  $\mathbf{X}$ :

$$cpx = \frac{[1 + (p-1)(1-K)] \cdot \frac{1}{p}(p \log_2 n)}{\log_2 n} = 1 + (p-1)(1-K)$$

Then,  $cpx = p$  indicates the presence in the model of  $p$  perfectly uncorrelated  $x$ -variables, while  $cpx$  values lower than 1 indicate insufficient information in the  $X$ -block to completely model the  $y$  response.

- **model of the frontier steric effect**  $\rightarrow$  steric descriptors ( $\odot$  Taft steric constant)
- **model sum of squares**  $\rightarrow$  regression parameters

■ **Mode of Action ( $\equiv$  MOA)**

In toxicology, it is accepted that reliable QSARs can be attained only if toxicants are considered separately, depending on their mechanism of action (MOA), that is, only those chemicals showing a similar mode of action can be used together to search for a QSAR [Bradbury and Lipnick, 1990; Bradbury, 1994; Schüürmann, Segner *et al.*, 1997; Freidig and Hermens, 2000; Nendza and Müller, 2000].

For chemicals with the same MOA, similar structural features can be searched for assuming that they give rise to similar reactivity mechanisms. Then, the basic QSAR strategy provides for identifying critical structural elements responsible for activity via a hypothetical shared mode of action and then constructing QSAR models able to classify different modes of action.

Moreover, the mode of action is very important for mixtures of chemicals because the biological effect of mixtures can give rise to different kinds of toxicological response (antagonism, less than additive response, additive response, and synergism), depending on the toxicological mode of action and the chemical interactions of the substances involved [Calamari and Vighi, 1992; Gramatica, Vighi *et al.*, 2001]. As a consequence, in several cases,  $\rightarrow$  *biological*

*activity indices* need to be studied for different classes of compounds, each having a different mode of action.

📖 [Lipnick, 1991; Gramatica, Vighi *et al.*, 2001; Musumarra, Condorelli *et al.*, 2001; Aptula, Netzeva *et al.*, 2002; Ren, 2002d; Pino, Giuliani *et al.*, 2003; Ren, 2003f; Schüürmann, Aptula *et al.*, 2003; Öberg, 2004a; Spycher, Nendza *et al.*, 2004; Chakraborty and Devakumar, 2005; Papa, Villa *et al.*, 2005; Spycher, Pellegrini *et al.*, 2005; Basak, Gute *et al.*, 2006]

- **modified edge-weighted Harary index** → weighted matrices (⊙ weighted adjacency matrices)
- **modified edge-weighted Harary matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **modified edge-Zagreb matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **modified Free–Wilson analysis** → Free–Wilson analysis
- **modified Hosoya index** → delocalization degree indices (⊙ Hosoya resonance energy)
- **modified Hosoya index**  $\equiv$  *stability index* → characteristic polynomial-based descriptors
- **modified Hosoya index**  $\equiv$  *Z\* index* → Hosoya Z index
- **Modified Information Content index** → indices of neighborhood symmetry
- **modified LEACH index** → environmental indices (⊙ leaching indices)
- **Modified LSER** → Linear Solvation Energy Relationships
- **modified LUDI energy function** → scoring functions (⊙ LUDI energy function)
- **modified partial equalization of orbital electronegativities** → electronegativity
- **modified Randić index** → connectivity indices
- **modified spectrum-like descriptors** → spectrum-like descriptors
- **modified total adjacency index** → Zagreb indices
- **modified vertex Zagreb matrix** → vertex degree
- **modified weighted Tanimoto coefficient** → similarity/diversity
- **modified Wiener index** → Wiener index
- **modified Wiener indices**  $\equiv$  *generalized Wiener indices* → Wiener index
- **modified Zagreb indices** → Zagreb indices
- **modulo compression algorithm** → substructure descriptors
- **modulo-L descriptors** → spectra descriptors
- **MOE descriptors**  $\equiv$  *QuaSAR descriptors*
- **Mohar indices** → Laplacian matrix
- **molar polarization** → electric polarization descriptors
- **molar refractivity** → physico-chemical properties
- **molar refractivity partition index** → physico-chemical properties (⊙ molar refractivity)
- **molar volume** → volume descriptors
- **MolBlaster descriptors** → substructure descriptors (⊙ structural keys)

### ■ MolConn-Z descriptors

MolConn-Z is a software for the calculation of more than 400 molecular descriptors of different kinds [MolConn-Hall Associates Consulting, 1991; Faulon, Visco *et al.*, 2003].

MolConn-Z descriptors include valence, path, cluster, path/cluster, and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, Wiener and Platt indices, information

indices, counts of different vertices, and counts of paths and edges between different types of vertices, fragment counts, hydrogen-bonding descriptors, as well as a small set of 3D-descriptors.

📖 [Zheng and Tropsha, 2000; Tropsha and Zheng, 2002; Xiao, Xiao *et al.*, 2002; Bergström, Norinder *et al.*, 2003; Balaban, Basak *et al.*, 2004; Basak, Gute *et al.*, 2004; Kovatcheva, Golbraikh *et al.*, 2004; Medina-Franco, Golbraikh *et al.*, 2005; Wang, Li *et al.*, 2005]

- **MolDiA descriptors** → substructure descriptors (⊙ structural keys)
- **molecular branching** → molecular complexity
- **molecular centrality** → molecular complexity

### ■ molecular complexity

The concept of *molecular complexity* was introduced into chemistry only quite recently and is mainly based on the → *information content* of molecules. Several different measures of complexity can be obtained according to the diversity of the considered structural elements such as atom types, bonds, connections, cycles, and so on. The first attempts to quantify molecular complexity were based on the elemental composition of molecules; later other molecular characteristics were considered, such as the symmetry of molecular graphs, molecular branching, molecular cyclicity, and centrality [Bonchev, 1990; Bonchev and Seitz, 1996; Rücker and Rücker, 2000; Randić and Plavšić, 2002; Bonchev and Rouvray, 2003, 2007; Newman, 2004; Rücker, Rücker *et al.*, 2004; Bonchev and Buck, 2005].

A composite hierarchical concept of molecular complexity was proposed by Bonchev–Polansky [Bonchev and Polansky, 1987] according to their *general complexity scheme*, which begins with molecule size and proceeds through topology; molecules of the same size and topology are distinguished by their atom and bond types; moreover, a further discrimination is provided by geometric → *interatomic distances* and molecular symmetry.

In particular, topological complexity is hierarchically defined and the main features are *molecular branching*, *molecular cyclicity*, and *molecular centrality*.

Some reviews about molecular complexity are [Bonchev, 1999, 2003b; Barone and Chanon, 2001; Hann, Leach *et al.*, 2001; Cerruti, 2005; Schuffenhauer, Brown *et al.*, 2006].

### • molecular branching

This is a molecule property comprising several structural variables such as number of branching, valence, distances apart, distances from the → *graph center*, and length of branches [Kirby, 1994]. Given this multifaceted definition of branching, its quantification is not an easy task. However, operational definitions of branching can be given by selected molecular indices, called **branching indices**, which, to some extent, reflect the branching of molecules as intended in an intuitive way [Randić, 1975d; Gutman and Randić, 1977; Bonchev, von Knop *et al.*, 1979; Barysz, von Knop *et al.*, 1985; Bertz, 1988; Rouvray, 1988b; Bonchev, 1995; Klein and Babic, 1997; Randić, 1997b; Gutman and Vidović, 2002b; Perdih, 2003].

The → *Wiener index* was the first proposed index of molecular branching [Bonchev and Trinajstić, 1977]; it is a function, inversely related to branching, of the number, length, and position of branches as well as of the number of atoms. For an isomeric series, it can be considered mainly dependent on molecular branching. Other specific molecular descriptors

proposed as measures of branching are the  $\rightarrow$  *Lovász–Pelikan index*,  $\rightarrow$  *ramification index*,  $\rightarrow$  *Zagreb indices*,  $\rightarrow$   $\lambda\lambda_1$  *branching index*, and  $\rightarrow$  *branching ETA index*. The  $\rightarrow$  *Balaban centric index*, the  $\rightarrow$  *Randić connectivity index*, the  $\rightarrow$  *mean information content of the distance equality* [Bonchev and Trinajstić, 1978], and the  $\rightarrow$  *Merrifield–Simmons index* can also be used to discriminate among isomeric molecules with different branching patterns. Moreover, several branching indices were proposed by specific pairs of  $m$  and  $n$  values of  $\rightarrow$   $v^m d^n$  *matrices* or  $a, b, c$  values of the  $\rightarrow$  *general distance–degree matrix*  $G(a, b, c)$  [Perdih, 2003], by summing all the off-diagonal elements, by calculating the largest eigenvalues and by the product of all the off-diagonal elements. In particular, the largest eigenvalue obtained for  $a = b = -1/4$  and  $c = -1$  was proposed as branching index.

The simplest but effective index of molecular branching is the **Bertz branching index**, defined as [Bertz, 1988; Ivanciuc, Ivanciuc *et al.*, 2000c]

$$BI = \frac{1}{2} \cdot \sum_{i=1}^A \delta_i \cdot (\delta_i - 1)$$

where the sum runs over all the atoms and  $\delta_i$  is the  $\rightarrow$  *vertex degree* of the  $i$ th atom; as it can be noted, contribution to branching of terminal atoms ( $\delta_i = 1$ ) is zeroed. Note that this index can also be calculated from the  $\rightarrow$  *edge adjacency matrix* as the  $\rightarrow$  *connection number*.

#### • molecular cyclicity

It is another important feature of molecular complexity, defined in terms of number of molecule cycles and the manner in which the cycles are connected. It was first characterized by Bonchev [Bonchev, Mekenyan *et al.*, 1980b; Bonchev and Mekenyan, 1983] by a system of rules based on the number of atoms, the number of cycles, the number of atoms in a cycle, the number of cycles having a common edge, and so on. Moreover, a number of molecular descriptors, usually called  $\rightarrow$  *ring descriptors*, were proposed to describe molecular cyclicity, accounting for the presence of cycles in molecules.

The  $\rightarrow$  *Wiener index* was initially chosen as a single-valued molecular descriptor related to the cyclicity degree of isomeric molecules; it decreases for molecules with cyclic structures more complex than in molecules of the same size but with fewer rings. The  $\rightarrow$  *Harary index* and  $\rightarrow$  *Kirchhoff number* were also proposed as discriminating cyclicity indices [Bonchev, Balaban *et al.*, 1994]. Moreover, Randić [Randić, 1997a] defined the cyclicity of a molecule in terms of the cyclicity of the corresponding molecular graph “... as the departure of the cyclic character of the graph from that of the monocyclic graph relative to the departure of the complete graph from the monocyclic graph.” The cyclic character of the graph is given by the  $\rightarrow$   $D/\Delta$  *index* calculated from the  $\rightarrow$  *distance/detour quotient matrix*. In practice, the degree of cyclicity of a molecule with  $A$  atoms is calculated by the comparison of the corresponding molecular graph with the two extreme graphs of the same size (monocycle and  $\rightarrow$  *complete graph*). The  $\rightarrow$  *cyclicity index*  $\gamma$  is a quantitative measure of molecular cyclicity as defined by Randić.

Other cyclicity indices are the  $\rightarrow$  *total edge cyclicity*, the  $\rightarrow$  *global cyclicity indices*, the  $\rightarrow$  *spanning tree number*, the  $\rightarrow$  *ring ID number*, the  $\rightarrow$  *ring degree–distance index*, and all the indices derived from  $\rightarrow$  *quotient matrices* defined in terms of two different graph distance matrices.

- **molecular centrality**

It is considered less important than branching and cyclicity, but it contributes to the quantification of molecular complexity by distinguishing between molecular structures organized differently with respect to their centers [Bonchev, 1997]. → *Centric indices* are topological descriptors related to molecular centrality. Moreover, the → *centric topological index* and → *centrocomplexity topological index* calculated from the → *branching layer matrix* of an → *iterated line graph sequence* were proposed to measure molecular centrality [Diudea, Horvath *et al.*, 1992].

**Molecular complexity indices** are mainly based on → *information indices* defined to account for molecule complexity. These indices may be broadly divided into topological complexity indices and chemical complexity indices [Basak, 1987]. The former are calculated as the → *information content* of molecular graphs where atoms are not distinguished; among these, only those able to account for multiplicity in the graph are used to measure molecular complexity. The latter accounts for the chemical nature of the individual atoms in terms of bonding topology of weighted graphs or through the use of the → *physico-chemical properties* of the atoms in the molecule.

Some complexity indices are defined below [Nikolić, Trinajstić *et al.*, 2003]. Other molecular descriptors that give information about molecular complexity are reported elsewhere; these are the → *total walk count*, → *path counts*, → *spanning tree number* → *indices of neighborhood symmetry*, and the → *total adjacency index*. The latter was proposed by Bonchev and Polansky [Bonchev and Polansky, 1987] as a simple measure of topological complexity, being a measure of the degree of connectedness of molecular graph. Moreover, some among the → *GETAWAY descriptors* have been proposed to account at varying extents for molecular complexity.

- **Bertz complexity index ( $I_{CPX}$ )**

The most popular complexity index was introduced by Bertz (Bertz, 1981, 1983a, 1983b) taking into account both the variety of bond connectivities and atom types of a → *H-depleted molecular graph*.

A general form of a molecular complexity index  $I_{CPX}$  is

$$I_{CPX} = I_{CPB} + I_{CPA}$$

where  $I_{CPB}$  and  $I_{CPA}$  are the → *information contents* related to the bond connectivity and the atom-type diversity, respectively. The term  $I_{CPB}$  was originally defined as

$$I_{CPB} \equiv C(TI) = 2 \cdot TI \cdot \log_2 TI - \sum_g TI_g \cdot \log_2 TI_g$$

where  $TI$  is any → *graph invariant*, and  $TI_g$  is the number of equivalent elements forming the graph invariant  $TI$ . The choice of the graph invariant should be based on the assumption that molecular complexity increases with size, branching, vertex and edge weights, and so on. The → *connection number*  $N_2$  (also called → *Bertz branching index*,  $BI$ ), that is, the number of bond pairs, was proposed by Bertz as a good choice for evaluating molecular complexity as it measures both the size and symmetry of the graph. Therefore, the two terms of the Bertz complexity index are defined as

$$I_{CPB} = 2 \cdot N_2 \cdot \log_2 N_2 - \sum_g (N_2)_g \cdot \log_2 (N_2)_g = N_2 \cdot \log_2 N_2 + {}^{CONN}I_{ORB}$$

$$I_{CPA} \equiv I_{AC} = A \log_2 A - \sum_g A_g \log_2 A_g$$



where  $(N_2)_g$  is the number of symmetrically identical connections of type  $g$ ;  $A$  is the total number of atoms (hydrogen excluded);  $A_g$  is the number of atoms of the  $g$ th element, and  $^{CONN}I_{ORB}$  is the  $\rightarrow$  *total connection orbital information content*. The term  $I_{CPB}$  measures the complexity of a molecule given by the partition of equivalent connections sensitive to branching, rings, and multiple bonds of the molecule; when all the connections are the same, the bond complexity term is equal to  $N_2 \log_2 N_2$  to take into account the size of the molecule, together with its symmetry in terms of bond connectivity. The atom complexity term  $I_{CPA}$  accounts for the presence of heteroatoms in the molecule and corresponds to the  $\rightarrow$  *total information index on atomic composition* calculated for H-depleted molecular graphs.

📖 [Nikolić and Trinajstić, 2000]

- **Rashevsky complexity index** ( $\bar{I}_{RASH}$ )

This is a quantitative measure of graph complexity per vertex based on the sum of a chemical and a topological term:

$$\bar{I}_{RASH} = \bar{I}_{AC} + \bar{I}_{TOP}$$

where the two terms are the  $\rightarrow$  *mean information index on atomic composition*  $\bar{I}_{AC}$  and the  $\rightarrow$  *topological information content*  $\bar{I}_{TOP}$ , respectively [Rashevsky, 1955]. Note that the topological information content proposed by Rashevsky is not based on graph orbits as is the most general topological information content later proposed by Trucco [Trucco, 1956a, 1956b]. In effect, two vertices  $v_i$  and  $v_j$  are considered topologically equivalent if for each  $k$ th neighboring vertex ( $k$  ranging between 1 and the  $\rightarrow$  *atom eccentricity*) of vertex  $v_i$ , there exists a  $k$ th neighboring vertex of the same degree for vertex  $v_j$ .

The Rashevsky complexity index was further developed by Mowshowitz to obtain a measure of relative complexity of undirected and directed graphs [Mowshowitz, 1968a, 1968b, 1968c, 1968d].

- **Dosmorov complexity index** ( $I_{DOSM}$ )

This is a molecular complexity index defined as a linear combination of five single information indices [Dosmorov, 1982]:

$$I_{DOSM} = I_{AC} + I_{at} + I_B + I_{SYM} + I_{CONF}$$

where  $I_{AC}$  is the  $\rightarrow$  *total information index on atomic composition* calculated for H-depleted molecular graphs,  $I_{at}$  the  $\rightarrow$  *atomic information content*,  $I_B$  the  $\rightarrow$  *information bond index*,  $I_{SYM}$  the  $\rightarrow$  *information index on molecular symmetry*, and  $I_{CONF}$  the  $\rightarrow$  *information index on molecular conformations*. This combination of indices was proposed as a general index of molecular complexity accounting for the chemical nature of atoms, molecular size, number, and kind of molecular bonds, symmetry, and conformations. Moreover, by incorporating the atomic information content the Dosmorov index becomes more discriminating than the Bertz index in the case of different substituent atoms of the same valence [Bonchev, 1983].

- **Bonchev complexity information index** ( $I_{BONC}$ )

This is a molecular descriptor defined by analogy with the Dosmorov complexity index, also accounting for electronic properties of molecules [Bonchev, 1983]:

$$I_{BONC} = I_{IC} + I_{NUCL} + I_{EL} + I_{Topology} + I_{SYM} + I_{CONF}$$

where  $I_{IC}$  is the  $\rightarrow$  information index on isotopic composition,  $I_{EL}$  is one of the  $\rightarrow$  electronic information indices;  $I_{Topology}$  can be any topological information index accounting for structural complexity;  $I_{SYM}$  is the  $\rightarrow$  information index on molecular symmetry,  $I_{CONF}$  the  $\rightarrow$  information index on molecular conformations, and  $I_{NUCL}$  the  $\rightarrow$  nuclear information content.

#### • Minoli complexity index

This is a measure of complexity of a molecular graph monotonically increasing on the number of vertices and edges, and reflecting the degree of connectedness of the graph [Minoli, 1976]. It is defined for undirect graphs, with no loops and multiple edges as

$$\chi = \frac{A \cdot B}{A + B} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A P_{ij} = \frac{A \cdot B}{A + B} \cdot \sum_{m=1}^L {}^m P$$

where  $A$  is the number of graph vertices,  $B$  the number of graph edges,  $P_{ij}$  the number of paths of any length between vertices  $v_i$  and  $v_j$ ,  ${}^m P$  the total number of paths of length  $m$  in the graph, and  $L$  the length of the longest path in the graph.  ${}^m P$  are the elements of the  $\rightarrow$  molecular path code.

The Bonchev variant [Bonchev, 1990] of the Minoli index, called **Minoli–Bonchev complexity index**, was defined by replacing the number of paths  ${}^m P$  of length  $m$  with the products of the total number of paths by their length as

$$\chi_{MB} = \frac{A \cdot B}{A + B} \cdot \sum_{m=1}^L {}^m P \cdot m = \frac{A \cdot B}{A + B} \cdot W^{AP}$$

where  $W^{AP}$  is the  $\rightarrow$  all-path Wiener index.

#### • Bertz–Herndon relative complexity index ( $C_{BH}$ )

This is a simple measure of structural complexity of a molecule based on its graph representation  $G$  compared with the parent  $\rightarrow$  complete graph  $K(G)$ , that is, the complete graph with the same number of vertices. It is defined as

$$C_{BH} = \frac{K_G}{K_{K(G)}}$$

where  $K$  is the total number of connected subgraphs in  $G$  and  $K(G)$ , respectively [Bertz and Herndon, 1986; Bonchev, 1997].

#### • Bonchev topological complexity indices ( $\equiv$ topological complexity indices, TC, or overall topological indices, OI)

These are derived from the graph representation of molecules. A general overall topological index is formulated as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$OI = \sum_{k=1}^K \mathcal{D}(G_k) = \sum_{k=1}^K f(\mathcal{L}_i(i \in G_k))$$

where the summation runs over all the connected subgraphs  $G_k$  of the molecular graph  $G$ ,  $K$  being the total number of connected subgraphs of  $G$ , and  $\mathcal{D}(G_k)$  is any  $\rightarrow$  graph invariant derived from each  $k$ th subgraph and defined as some function  $f$  of the  $\rightarrow$  local vertex invariants  $\mathcal{L}_i$  of all vertices belonging to the  $k$ th subgraph.

The overall topological index  $\text{OI}$  is defined in two versions, depending on whether local vertex invariants  $L_i$  are those of the entire graph  $G$  or those of the subgraph  $G_k$ ; for the latter, the symbol  $\text{OI1}$  was suggested instead of  $\text{OI}$ .

The  $m$ th order overall topological index, denoted by  ${}^m\text{OI}$ , is defined as the sum of the invariants  $\mathcal{D}({}^mG_j)$  of all  ${}^mK$  subgraphs  ${}^mG_j$ , which have  $m$  edges:

$${}^m\text{OI} = \sum_{j=1}^{{}^mK} \mathcal{D}({}^mG_j)$$

The overall topological vector  $\text{OI}(G)$  of any graph  $G$  having  $B$  edges is the sequence of all  ${}^m\text{OI}$  indices listed in ascending order:

$$\text{OI}(G) = ({}^0\text{OI}, {}^1\text{OI}, {}^2\text{OI}, \dots, {}^B\text{OI})$$

The  $m$ th order overall topological indices  ${}^m\text{OI}_t$  of the  $t$ th class of subgraphs,  $t$  standing for subgraph type such as path, cluster, path cluster, cycles, and so on, were also defined by limiting the summation to the subgraphs of type  $t$ .

Moreover, the complexity vector  $\mathbf{K}(G)$  of a graph  $G$  was analogously defined as

$$\mathbf{K}(G) = ({}^0K, {}^1K, {}^2K, \dots, {}^BK), \quad K = \sum_{m=0}^B {}^mK$$

where  ${}^mK$  is the number of subgraphs having  $m$  edges.

**Overall connectivity indices** are overall topological indices derived from the  $\rightarrow$  adjacency matrix of the molecular graph. The overall connectivity, denoted by  $TC$ , is defined as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$TC = \sum_{m=0}^B {}^mTC = \sum_{k=1}^K A_V(G_k) = \sum_{k=1}^K \sum_{i=1}^{N_k} \delta_i (i \in G_k)$$

where  $A_V(G_k)$  is the  $\rightarrow$  total adjacency index of the  $k$ th subgraph  $G_k$ ;  $N_k$  is the number of vertices in the  $k$ th connected subgraph, and  $\delta_i$  is the  $\rightarrow$  vertex degree of the  $i$ th vertex of the subgraph. In practice, the vertex degree is assigned to each vertex in the molecular graph, and then for each connected subgraph the degrees of all subgraph vertices are added; this quantity is summed up over all subgraphs of the same order to generate the partial  $m$ th order overall connectivity  ${}^mTC$ , and, finally, all these partial overall connectivities are summed up to give the total overall connectivity  $TC$ . Note that the overall connectivity of zero order coincides with the total adjacency index of the entire molecular graph:

$${}^0TC \equiv A_V(G) = \sum_{k=1}^A \delta_k (k \in G)$$

A different definition of overall connectivities  $TC1$  was given by considering the vertex degrees as they are in isolated subgraphs. The inequality  $TC > TC1$  always holds. Moreover, **valence overall connectivity indices**  $TC^v$  were defined by using the  $\rightarrow$  valence vertex degrees  $\delta^v$  in place of the simple vertex degrees:

$$TC^v = \sum_{k=1}^K \sum_{i=1}^{N_k} \delta_i^v (i \in G_k)$$

Overall connectivity indices were proposed as a meaningful measure of topological complexity of molecules, since they satisfy two fundamental requirements to a complexity measure: to increase with both the number of structural elements and their interconnectedness; the basic idea is that “*The higher the connectivity of molecular graph and its connected subgraphs, the more complex the molecule*” [Bonchev and Trinajstić, 1977].

**Overall Zagreb indices**, denoted by OM2, were defined as an extension of  $\rightarrow$  Zagreb indices as [Bonchev, 1997; Nikolić, Tolić *et al.*, 2000]

$$OM2 = \sum_{m=0}^B {}^mOM2$$

where  ${}^mOM2$  is the  $m$ th order overall Zagreb index, which can be calculated either generally  ${}^mOM2$  or separately for each type of subgraphs  ${}^mOM2_t$ :

$${}^mOM2 = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \delta_i (i \in G_j) \quad {}^mOM2_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \delta_i (i \in G_j)$$

where  $\delta_i$  is the  $\rightarrow$  vertex degree of the  $i$ th vertex belonging to the  $j$ th subgraph, having  $N_j$  vertices;  ${}^mK$  is the total number of connected subgraphs having  $m$  edges, and  ${}^mK_t$  is the total number of connected subgraphs of type  $t$  having  $m$  edges. Note that the overall Zagreb index of zero order is simply the sum of all the vertex degrees, thus resulting equivalent to the overall connectivity of zero order and hence the total adjacency index:

$${}^0OM2 \equiv {}^0TC \equiv A_v(G) = \sum_{k=1}^A \delta_k (k \in G)$$

Moreover, the first order overall Zagreb index coincides with the second Zagreb index  $M_2$ :

$${}^1OM2 \equiv M_2 = \sum_{k=1}^B (\delta_i \cdot \delta_j)_k$$

Other overall indices, similar to overall Zagreb indices, were defined by using a different function that has terms inverse to those of OM2; these indices were denoted by ON and are defined as [Bonchev, 1999, 2000, 2001a, 2001b; Bonchev and Trinajstić, 2001]

$$ON = \sum_{m=0}^B {}^mON \quad {}^mON = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \delta_i^{-1} (i \in G_j) \quad {}^mON_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \delta_i^{-1} (i \in G_j)$$

where the difference from overall Zagreb indices is given by the use of reciprocal vertex degrees. The zero order overall index ON is

$${}^0ON = \sum_{k=1}^A \delta_k^{-1} (k \in G)$$

whereas the overall index of first order is the  $\rightarrow$  *modified Zagreb index*  ${}^mM_2$ :

$${}^1\text{ON} \equiv {}^mM_2 = \sum_{k=1}^B (\delta_i \cdot \delta_j)_k^{-1}$$

The **overall Wiener indices** were analogously defined with the aim of extending the  $\rightarrow$  *Wiener index* to its most complete version [Bonchev, 2001b]. They are defined as

$$\text{OW} = \sum_{m=0}^B {}^m\text{OW} \quad {}^m\text{OW} = \sum_{j=1}^{mK} \prod_{i=1}^{N_j} \sigma_i(i \in G_j) \quad {}^m\text{OW}_t = \sum_{j=1}^{mK_t} \prod_{i=1}^{N_j} \sigma_i(i \in G_j)$$

where  $\sigma_i$  is the  $\rightarrow$  *vertex distance degree*, OW is the total overall Wiener index,  ${}^m\text{OW}$  the  $m$ th order overall Wiener index, and  ${}^m\text{OW}_t$  the  $m$ th order overall Wiener index restricted to those subgraphs having  $m$  edges and of type  $t$ . Note that the overall Wiener index of maximal order is the Wiener index:  ${}^B\text{OW} = W$ ; the zero-order overall Wiener index equals zero:  ${}^0\text{OW} = 0$ ; and the first-order overall Wiener index is equal to the number of graph edges:  ${}^1\text{OW} = B$ .

#### • Bonchev-Trinajstić complexity index (BT)

This is a complexity index defined in terms of  $\rightarrow$  *information content*. The equivalence classes are defined collecting equal topological distances in the  $\rightarrow$  *H-depleted molecular graph* [Bonchev and Trinajstić, 1977]. It is defined as

$$BT = \frac{A \cdot (A-1)}{2} \cdot \log_2 \frac{A \cdot (A-1)}{2} - \sum_{k=1}^D {}^kf \cdot \log_2 {}^kf$$

where  $A$  is the number of graph vertices,  ${}^kf$  is the  $\rightarrow$  *graph distance count* of  $k$ th order, and  $D$  the topological diameter.

#### • Randić–Plavšić complexity index ( $\xi$ )

The Randić–Plavšić complexity index or  **$\xi$  index** is based on the concept of the  $\rightarrow$  *augmented valence*. The augmented valence AV of the  $i$ th vertex is obtained by adding to the  $\rightarrow$  *vertex degree* the vertex degrees of all the other vertices in the graph, each weighted by a quantity that decreases as the distance from the vertex  $v_i$  increases [Randić, 2001b; Randić and Plavšić, 2002, 2003].

The Randić–Plavšić complexity index  $\xi$  is defined as the sum of augmented valences of all mutually nonequivalent vertices in the graph:

$$\xi = \sum_{i=1}^{A'} \text{AV}_i$$

where the sum is restricted to nonsymmetrical atoms  $A'$ .

- **molecular complexity indices**  $\rightarrow$  molecular complexity
- **Molecular Connectivity Indices**  $\rightarrow$  connectivity indices
- **molecular connectivity topochemical index**  $\equiv$  *atomic molecular connectivity index*  $\rightarrow$  connectivity indices
- **molecular cyclicity**  $\rightarrow$  molecular complexity

- **molecular cyclized degree** → ring descriptors
- **Molecular Descriptor Family** → graph invariants
- **molecular descriptor properties** → molecular descriptors

■ **molecular descriptors** ( $\equiv$  *chemical descriptors*)

Molecular descriptors play a fundamental role in chemistry, pharmaceutical sciences, environmental protection policy, and health researches, as well as in quality control, being the way molecules, thought of as real bodies, are *transformed* into numbers, allowing some mathematical treatment of the chemical information contained in the molecule [Todeschini and Consonni, 2000]. Molecular descriptors allow to find → *structure/response correlations* and perform → *similarity searching*, → *substructure searching*, and → *drug design*.

Therefore, molecular descriptors are formally mathematical representations of a molecule obtained by a well-specified algorithm applied to a defined *molecular representation* or a well-specified experimental procedure: *the molecular descriptor is the final result of a logical and mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment*.

The term “useful” has a double meaning: it means that the number can give more insight into the interpretation of the molecular properties and/or is able to take a part in a model for the prediction of some interesting molecular properties. Even if the interpretation of a molecular descriptor can be weak, provisional, or even completely lacking, it could be strongly correlated to some molecular properties to give models with high prediction power. On the other hand, descriptors with poor prediction power can be usefully retained in models when they are well theoretically founded and interpretable due to their ability to encode structural chemical information.

Although several molecular quantities were defined from the beginning of the quantum-chemistry and the graph theory, the term “molecular descriptor” has become popular with the development of structure–property correlation models. The → *Platt number* [Platt, 1947] and → *Wiener index* [Wiener, 1947c] defined in 1947 are sometimes referred to as the first molecular descriptors.

By the definition given above, the molecular descriptors are divided into two main classes: → *experimental measurements*, such as → *log P*, → *molar refractivity*, → *dipole moment*, *polarizability*, and, in general, → *physico-chemical properties*, and **theoretical molecular descriptors**, which are derived from a symbolic representation of the molecule and can be further classified according to the different types of *molecular representation*.

Mathematics and statistics, graph theory, computational chemistry, and molecular modeling techniques enable the definition of a large number of theoretical descriptors characterizing physico-chemical and biological properties, reactivity, shape, steric hindrance, and so on of the whole molecule, molecular fragments, and substituents.

The fundamental difference between theoretical descriptors and experimentally measured ones is that theoretical descriptors contain no statistical error due to experimental noise, which is not the case for experimental measurements.

However, the assumptions needed to facilitate calculation and numerical approximation are themselves associated with an inherent error, although in most cases the direction, but not the magnitude, of the error is known. Moreover, within a series of related compounds, the error term is usually considered to be approximately constant. All kinds of error are absent only for the

most simple theoretical descriptors such as count descriptors or for descriptors directly derived from exact mathematical theories such as  $\rightarrow$  *graph invariants*.

Theoretical descriptors derived from physical and physico-chemical theories show some natural overlap with experimental measurements. Several  $\rightarrow$  *quantum-chemical descriptors*, surface areas, and  $\rightarrow$  *volume descriptors* are examples of such descriptors also having an experimental counterpart.

With respect to the experimental measurements, the greatest recognized advantages of the theoretical descriptors are usually (but not always) in terms of cost, time, and availability.

The **molecular representation** is the way that a molecule, that is, a phenomenological real body, is symbolically represented by a specific formal procedure and conventional rules. The quantity of chemical information, which is transferred to the molecule symbolic representation, depends on the kind of representation [Testa and Kier, 1991; Jurs, Dixon *et al.*, 1995].

The simplest molecular representation is the **chemical formula** (or **molecular formula**), which is the list of the different atom types, each accompanied by a subscript representing the number of occurrences of the atoms in the molecule. For example, the chemical formula of *p*-chlorotoluene is  $C_7H_7Cl$ , indicating the presence in the molecule of  $A=8$  (number of atoms, hydrogen excluded),  $N_C=7$ ,  $N_H=7$ , and  $N_{Cl}=1$  (the subscript "1" is usually omitted in the chemical formula).

This representation is independent of any knowledge concerning the molecular structure, and hence molecular descriptors obtained from the chemical formula can be called **0D descriptors**. Examples are the  $\rightarrow$  *atom number A*,  $\rightarrow$  *molecular weight MW*,  $\rightarrow$  *atom-type count  $N_x$* , and, in general,  $\rightarrow$  *constitutional descriptors* and any function of the  $\rightarrow$  *atomic properties*.

The atomic properties are usually the  $\rightarrow$  *weighting schemes* used to characterize molecule atoms; the most common atomic properties are atomic mass,  $\rightarrow$  *atomic charge*,  $\rightarrow$  *van der Waals radius*,  $\rightarrow$  *atomic polarizability*, and  $\rightarrow$  *atom electronegativity*. Atoms can also be characterized by the  $\rightarrow$  *local vertex invariants* (LOVIs) derived from graph theory.

The **substructure list representation** can be considered as a one-dimensional representation of a molecule and consists of a list of structural fragments of a molecule; the list can be only a partial list of fragments, functional groups, or substituents of interest present in the molecule, thus not requiring a complete knowledge of the molecule structure. The descriptors derived by this representation can be called **1D-descriptors** and are typically used in  $\rightarrow$  *substructural analysis* and  $\rightarrow$  *substructure searching*.

The two-dimensional representation of a molecule considers how the atoms are connected, that is, it defines the connectivity of atoms in the molecule in terms of the presence and nature of chemical bonds. Approaches based on the  $\rightarrow$  *molecular graph* allow a two-dimensional representation of a molecule, usually known as **topological representation**. Molecular descriptors derived from the algorithms applied to a topological representation are called **2D-descriptors**, that is, they are the so-called  $\rightarrow$  *graph invariants*. In the last few years, several efforts have been made to formalize the several formulas and algorithms dealing with molecular graph information: "a graph operator applies a mathematical equation to compute a whole class of related molecular graph descriptors, using different molecular matrices and various weighting schemes. ... In this way, molecular graph operators introduce a systematization of topological indices and graph invariants by assembling together all descriptors computed with the same mathematical formula or algorithm, but with different parameters or molecular matrices" [Ivanciuc, 2000i].

Two-dimensional representations alternative to the molecular graph are the **linear notation systems**, for example, **Wiswesser Line Notation system (WLN)** [Smith and Baker, 1975], **SMILES** [Weininger, 1988, 1990, 2003; Weininger, Weininger *et al.*, 1989; Convard, Dubost *et al.*, 1994; Hinze and Welz, 1996], and **SMARTS** (SMART – Daylight Chemical Information Systems, 2004). **CAST** (*CAnonical representation of STereochemistry*) is a method that gives a linear notation that canonically represents stereochemistry around a specific site in a molecule [Satoh, Koshino *et al.*, 2000, 2001, 2002].

The three-dimensional representation views a molecule as a rigid geometrical object in space and allows not only a representation of the nature and connectivity of the atoms, but also the overall spatial configuration of the molecule. This representation of a molecule is called **geometrical representation** and molecular descriptors derived from this representation are called **3D-descriptors**. Examples of 3D descriptors are the → *geometrical descriptors*, several → *steric descriptors*, and → *size descriptors*.

Several molecular descriptors derive from multiple molecular representations and can then be classified with difficulty. For example, graph invariants derived from a molecular graph weighted by properties obtained by → *computational chemistry* are both 2D and 3D descriptors.

The **bulk representation** of a molecule describes the molecule in terms of a physical object with 3D attributes such as bulk and steric properties, surface area, and volume.

The **stereoelectronic representation** (or **lattice representation**) of a molecule is a molecular description related to those molecular properties arising from electron distribution, interaction of the molecule with probes characterizing the space surrounding them (e.g., → *molecular interaction fields*). This representation is typical of the → *grid-based QSAR techniques*. Descriptors at this level can be considered **4D-descriptors**, being characterized by a scalar field, that is, a lattice of scalar numbers, associated with the 3D → *molecular geometry*.

Finally, the **stereodynamic representation** of a molecule is a time-dependent representation, which adds structural properties to the 3D representations, such as flexibility, conformational behavior, transport properties, and so on. → *Dynamic QSAR*, → *4D-Molecular Similarity Analysis*, and → *4D-QSAR Analysis* are examples of a multiconformational approach.

Within the two main classes of descriptors, experimental measurements and theoretical descriptors, several other subclasses of molecular descriptors can be recognized on the basis of a rational analysis of the molecular descriptor properties.

The main properties of the descriptors can be represented by a four-level taxonomy. Together with the first-level classification based on the molecular representation, as defined above, the other three levels are summarized below.

- **mathematical representation of molecular descriptors**

The descriptors can be represented by a scalar value, a vector, a two-way matrix, a tensor, or a scalar field, which can be discretized into a lattice of grid points.

- **invariance properties of molecular descriptors**

The invariance properties of molecular descriptors can be defined as the ability of the algorithm for their calculation to give a descriptor value that is independent of the particular characteristics of the molecular representation, such as atom numbering or labeling, spatial reference frame,



molecular conformations, and so on. Invariance to molecular numbering or labeling is assumed as a minimal basic requirement for any descriptor. **Chemical invariance** of a molecular descriptor means that its values are independent of the atom types and multiple bonds, that is, the descriptor is not able to account for heteroatoms and  $\rightarrow$  *bond multiplicity* in the molecules. Such invariance is considered explicitly in classifying topological indices as  $\rightarrow$  *topostructural indices* and  $\rightarrow$  *topochemical indices*.

Two other important invariance properties, **translational invariance** and **rotational invariance**, are the invariance of a descriptor value to any translation or rotation of the molecules in the chosen reference frame. Molecular descriptors being invariant to translation and rotation of a molecule are referred to as **TRI descriptors**. These invariance properties have to be considered when dealing with descriptors derived from  $\rightarrow$  *molecular geometry*. For all descriptors based on  $\rightarrow$  *internal coordinates*, rototranslational invariance is naturally guaranteed. For descriptors based on spatial atomic coordinates, translational invariance is usually easily attained by centering the atomic coordinates; rotational invariance may be satisfied by using, as the reference frame, an univocally defined frame such as the principal axes of each molecule. In some QSAR methods, such as grid-based QSAR techniques, the problem of invariance to rotation is, at least in principle, overcome by adopting  $\rightarrow$  *alignment rules*.

**Conformational invariance** means that molecular descriptor values are independent of the conformational changes in molecules. *Conformations* of molecules are the different atom dispositions in the 3D space, that is, configurations that flexible molecules can assume without any change to their connectivity. Usually interest in different conformations of a molecule is related to those conformations for which the total energy is relatively close to the minimum energy, that is, within a cutoff energy value of some kcal/mol.

Molecular descriptors can be distinguished according to their conformational invariance degree in four classes, as suggested by Charton [Charton, 1983]:

- (a) **No Conformational Dependence (NCD descriptors)**: This is typical of all descriptors, which do not depend on 3D molecular geometry, such as  $\rightarrow$  *molecular weight*,  $\rightarrow$  *count descriptors*, and  $\rightarrow$  *topological indices*.
- (b) **Low Conformational Dependence (LCD descriptors)**: This is the case of molecular descriptors whose values show small variations only in the presence of relevant conformational changes, such as *cis/trans* configurations. Examples are  $\rightarrow$  *cis/trans descriptors* and usually  $\rightarrow$  *charge descriptors*.
- (c) **Intermediate Conformational Dependence (ICD descriptors)**: These are molecular descriptors whose values show small variations in the presence of any conformational changes. Typical descriptors of this class are  $\rightarrow$  *EVA descriptors* and descriptors based on mass distribution, for example,  $\rightarrow$  *radius of gyration*.
- (d) **High Conformational Dependence (HCD descriptors)**: This is the case of descriptors with values very sensitive to any conformational change in the molecule. Typical descriptors of this class are  $\rightarrow$  *interaction energy values* obtained from  $\rightarrow$  *molecular interaction fields*,  $\rightarrow$  *3D-MoRSE descriptors*,  $\rightarrow$  *WHIM descriptors*,  $\rightarrow$  *G-WHIM descriptors*,  $\rightarrow$  *spectrum-like descriptors*,  $\rightarrow$  *shape descriptors* based on molecular geometries, and so on.

Among the  $\rightarrow$  *quantum-chemical descriptors*, descriptors of different kinds of conformational dependence can be found:  $\rightarrow$  *ionization potential*,  $\rightarrow$  *electron affinity*, and molecular orbital

energies are often LCD or ICD descriptors, whereas molecular energies are usually HCD descriptors.

To quantify the conformational sensitivity of molecular descriptors, the **conformational pairwise sensitivity** (*CPS*) was proposed as the difference between conformationally dependent physico-chemical properties *P*, such as, for example, dipole moments, polar surface area, or 3D calculated log*P*, over the difference between geometry-dependent molecular descriptors *D* [Vistoli, Pedretti *et al.*, 2005]. This quantity is defined as

$$CPS_{st}(P, D) = \frac{|P_s - P_t|}{|D_s - D_t|}$$

where *s* and *t* are two different molecular conformations. The **conformational global sensitivity** of a descriptor (*CGS*) is then defined as the average of the pairwise sensitivities for all the possible pair combinations of *N* conformers:

$$CGS(P, D) = \frac{\sum_{st} CPS_{st}(P, D)}{N \cdot (N-1)}$$

It should be noted that some invariance properties such as invariance to atom numbering and rototranslations are mandatory for molecular descriptors used in QSAR/QSPR modeling; in several cases, chemical invariance is required, particularly when dealing with a series of compounds with different substituents; moreover, conformational invariance is closely dependent on the considered problem.

#### • degeneracy of molecular descriptors

This property refers to the ability of a descriptor to avoid equal values for different molecules. In this sense, descriptors can show no degeneracy at all (N), low (L), intermediate (I), or high (H) degeneracy. The degree of degeneracy of a descriptor can naturally be measured by  $\rightarrow$  *Shannon's entropy*. Moreover, the degree of degeneracy depends on the molecules present in the considered data set. Suitable measures of molecular descriptor degeneracy can be provided by using a data set consisting of an extended hydrocarbon series as well as heteroatoms and cycles.

Information content and Shannon's entropy of molecular descriptors were extensively studied by Bajorath, Godden, and coworkers in several papers [Godden, Stahura *et al.*, 2000; Godden and Bajorath, 2000, 2002, 2003].

Degeneracy of 735 molecular descriptors as well as their pairwise correlations were estimated on the NCI database for 221,860 compounds and made available on a software module called Molecular Descriptor Correlations (MDC) [MDC – Milano Chemometrics, 2006].

Degeneracy is considered an undesirable characteristic for all molecular descriptors that are used for the characterization of molecules in store and retrieval database systems; however, in QSAR modeling, degenerate properties are better modeled by molecular descriptors showing analogous degeneracy [Todeschini, Consonni *et al.*, 1998].

Based on the previous criteria, examples of an indicative classification of molecular descriptors are shown in Table M10.

**Table M10** Mathematical properties of some molecular descriptors.

Descriptors	Molecular representation	Mathematical representation	Invariance properties	Degeneracy
Molecular weight	0D	Scalar	NCD	H
Atom-type counts	0D	Scalar	NCD	H
Fragment counts	1D	Scalar	NCD	H
Topological information indices	2D	Scalar	NCD	L/I
Molecular profiles	2D	Vector	NCD	N
2D autocorrelation descriptors	2D	Vector	NCD	N/L
3D autocorrelation descriptors	3D	Vector	MCD	N
Substituent constants	3D	Scalar	NCD/LCD	L/I
WHIM descriptors	3D	Vector	HCD	N
3D-MoRSE descriptors	3D	Vector	LCD/MCD	N
GETAWAY descriptors	3D	Vector	MCD	N
Surface/volume descriptors	3D	Scalar	HCD/MCD	L
Quantum-chemical descriptors	3D	Scalar	MCD/HCD	N/L
Compass descriptors	3D	Vector	HCD	N
Interaction energy values	4D	Lattice	HCD/RD	N
GRIND descriptors	4D	Vector	HCD	N

Suitable molecular descriptors, besides the trivial invariance properties, should satisfy some basic requirements. The list of desirable requirements of chemical descriptors suggested by Randić [Randić, 1996a] is shown in Table M11.

**Table M11** List of desirable requirements for molecular descriptors.

#	Descriptors
1	Should have structural interpretation
2	Should have good correlation with at least one property
3	Should preferably discriminate among isomers
4	Should be possible to apply to local structure
5	Should be possible to generalize to “higher” descriptors
6	Descriptors should be preferably independent
7	Should be simple
8	Should not be based on properties
9	Should not be trivially related to other descriptors
10	Should be possible to construct efficiently
11	Should use familiar structural concepts
12	Should have the correct size dependence
13	Should change gradually with gradual change in structures

#### • transformations of molecular descriptors

Some formal definitions of molecular descriptors are only suitable for small molecules, but for big molecules, they take so large values that they are not algorithmically manageable. In these cases, a proper transformation should be applied; the most common transformations are

logarithmic transformations, such as

$$\mathcal{D}' = \log(\mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \ln(\mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \log(1 + \mathcal{D}) \quad \text{or} \quad \mathcal{D}' = \ln(1 + \mathcal{D})$$

Another common transformation of molecular descriptors is their conversion into binary descriptors [Xue, Godden *et al.*, 1999a, 2003b], that is, each descriptor represented by continuous or integer numbers is reduced to a single binary value (0 or 1) or, in some cases, a series of binary values. A simple way to assign 0 or 1 values to the descriptor in place of the actual descriptor value is by comparing the descriptor value with a cutoff value; 0 or 1 values are assigned to those descriptor values that fall either below or above the cutoff value. The cutoff can be, for example, the median of the descriptor estimated from the data set or the chemical library.

Transformations of a set of molecular descriptors are often performed when there is the need of a  $\rightarrow$  *variable reduction* or the need to modify binary vectors, such as site and substituent-oriented variables, into real-valued variable vectors. The milestone of these techniques is the  $\rightarrow$  *Principal Component Analysis* (PCA), but also  $\rightarrow$  *Fourier analysis* and  $\rightarrow$  *Wavelet analysis* are often used, especially for  $\rightarrow$  *spectra descriptors* compression.

Fourier analysis was, for instance, applied to change site and substituent-oriented binary variables in the  $\rightarrow$  *Free–Wilson analysis*, into a few latently dependent real coefficients [Holik and Halamek, 2002].

Wavelet analysis was also proposed for variable reduction problems and, in particular, the wavelet coefficients obtained from *discrete wavelet transforms* (DWT) were proposed as a molecular representation in  $\rightarrow$  *PEST descriptor methodology* and their sums as molecular descriptors [Breneman, Sundling *et al.*, 2003; Lavine, Davidson *et al.*, 2003].

Another interesting tool to obtain linear combinations of descriptors is that based on the **Andrews' curves** [Andrews, 1972]. The Andrews' curves are a pictorial way to represent and compare multivariate objects. In this representation of a  $p$ -dimensional space, each  $i$ th object is represented by a mathematical function as

$$f_i(t) = x_{i1}/\sqrt{2} + x_{i2} \cdot \sin(t) + x_{i3} \cdot \cos(t) + x_{i4} \cdot \sin(2t) + x_{i5} \cdot \cos(2t) + \dots$$

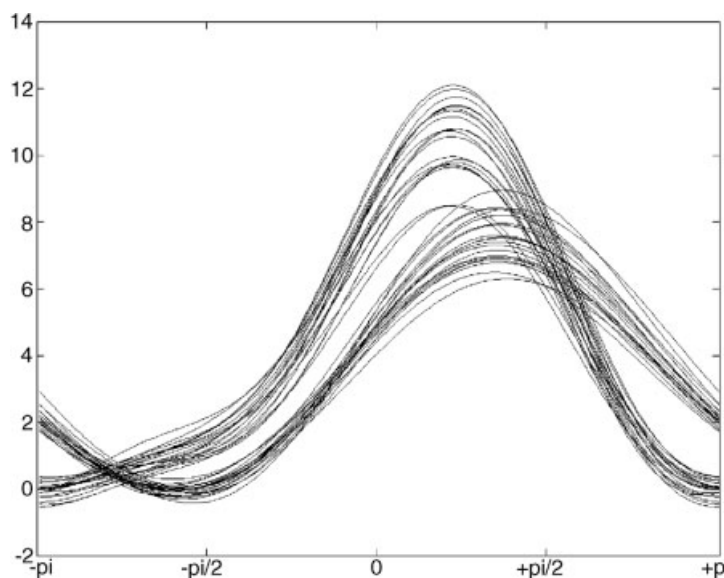
plotting this function over the range  $-\pi \leq t \leq \pi$ . The length of the  $f_i(t)$  vector depends on the resolution by which the range is spanned by the  $t$  parameter.

A set of molecules will thus appear as a set of curves drawn across the plot (Figure M2). This transformation preserves the means, the variances and the Euclidean distances calculated from the original variables  $x$ . However, the curves are not invariant with respect to the order of the variables: in general, the low frequencies ( $x_1, x_2, x_3$ ) are distinguished more readily on the plot than the high frequencies ( $x_p, x_{p-1}, x_{p-2}$ ). For this reason, it is better to associate the most important variables with the low frequencies [Todeschini, Consonni *et al.*, 1998].

#### • susceptibility of molecular descriptors

The susceptibility of a molecular descriptor  $\mathcal{D}$  is defined according to the following expression [Perdih and Perdih, 2002e, 2003d]:

$$S_{a,b} = \frac{\mathcal{D}_b}{\mathcal{D}_a} - 1$$



**Figure M2** Andrews' plot of 40 samples described by four variables.

where subscript *a* refers to a reference structure with respect to some molecule structural feature, whereas *b* refers to a reference structure with “opposite” characteristic with respect to the studied structural feature.

Susceptibilities were defined for alkanes for the increase in carbon number and branching; moreover, differential susceptibilities were proposed to highlight the contribution of the number of branches, the position of branches, the separation between branches, and the change of the substituent from methyl to ethyl [Perdih and Perdih, 2003d].

Susceptibilities can be evaluated also for the influence of heteroatoms assuming benzene (or *n*-alkane) as the reference molecule and heteroatom-substituted benzenes (or heteroatom-substituted alkanes).

#### • comparisons of molecular descriptors

Comparisons among molecular descriptors and among different classes of descriptors are important for at least two reasons: (a) to enhance the comprehension of the chemical meaning of complex descriptors by comparing them with other more interpretable descriptors; (b) to evaluate their different prediction ability relatively to the different kinds of response to be modeled.

Together with the comparison of the → *degeneracy of molecular descriptors*, several papers contain pairwise correlation tables of molecular descriptors as well as extended discussions about different classes of descriptors. These are: [Duperray, Chastrette *et al.*, 1976a; Hall and Kier, 1978a; Todeschini, Cazar *et al.*, 1992; Moriguchi, Hirono *et al.*, 1994; Basak, Gute *et al.*, 1996c; Mannhold and Dross, 1996; Das, Dömötör *et al.*, 1997; Dearden and Ghafourian, 1999; Viswanadhan, Ghose *et al.*, 2000; Clare, 2002; Consonni, Todeschini *et al.*, 2002b; Cruciani, Pastor *et al.*, 2002; Zissimos, Abraham *et al.*, 2002c; Doweiko, 2004; Asikainen, Ruuskanen

*et al.*, 2005; Fechner, Paetz *et al.*, 2005; Hollas, 2005b, 2005c; Perdih, 2000b, 2000c; Quigley and Nauhton, 2002; Geddeck, Rohde *et al.*, 2006].

Molecular descriptors are usually classified into several classes by a mixed taxonomy based on different points of view. For example, descriptors are often distinguished by their *physico-chemical meaning* such as  $\rightarrow$  *electronic descriptors*,  $\rightarrow$  *steric descriptors*,  $\rightarrow$  *lipophilicity descriptors*,  $\rightarrow$  *hydrogen-bonding descriptors*,  $\rightarrow$  *shape descriptors*,  $\rightarrow$  *charge descriptors*,  $\rightarrow$  *electric polarization descriptors*, and  $\rightarrow$  *reactivity descriptors*; moreover, on the basis of the specific mathematical tool used for the calculation of the molecular descriptors,  $\rightarrow$  *autocorrelation descriptors*,  $\rightarrow$  *spectral indices*,  $\rightarrow$  *determinant-based descriptors*,  $\rightarrow$  *Wiener-type indices*,  $\rightarrow$  *Schultz-type indices*,  $\rightarrow$  *characteristic polynomial-based descriptors*,  $\rightarrow$  *connectivity-like indices*, and  $\rightarrow$  *Balaban-like indices* can be distinguished.

📖 [Duperray, Chastrette *et al.*, 1976a; Jurs, Chou *et al.*, 1979; Jurs, Stouch *et al.*, 1985; Lavenhar and Maczka, 1985; Mekenyan and Bonchev, 1986; Jurs, Hasan *et al.*, 1988; Dearden, 1990; Govers, 1990; Randić, 1990b, 1991a, 1991c; Silipo and Vittoria, 1990; Weininger and Weininger, 1990; Ash, Warr *et al.*, 1991; Cronin, 1992; Horvath, 1992; Bonchev, Mountain *et al.*, 1993; Katritzky and Gordeeva, 1993; Randić and Trinajstić, 1993b; Rücker and Rücker, 1993; Dearden, Cronin *et al.*, 1995b; Basak, Gute *et al.*, 1996c, 1997, 1998b; Karelson, Lobanov *et al.*, 1996; Balaban, 1997a; Basak, Grunwald *et al.*, 1997; Klein and Babić, 1997; Matter, 1997; Gasteiger, 1998; Lee, Park *et al.*, 1998; Baumann, 1999; Jalali-Heravi and Parastar, 1999; Andersson, Sjöström *et al.*, 2000; Godden, Stahura *et al.*, 2000; Karelson, 2000; Randić and Basak, 2000b; Todeschini and Consonni, 2000; Vidal, Thormann *et al.*, 2005; Todeschini, 2006]

### ■ molecular distance-edge vector

The Molecular Distance-Edge vector (or **MDE vector**), denoted by  $\lambda$ , was proposed as a molecular 10-dimensional vectorial descriptor based on the geometric means of the topological distances between carbon atoms of predefined type [Liu, Cao *et al.*, 1998; Liu, Liu *et al.*, 1999]; the four types of carbon atoms were classified simply as primary carbon  $C_1$  (three bonded hydrogens), secondary  $C_2$  (two bonded hydrogens), tertiary  $C_3$  (one bonded hydrogens), and quaternary  $C_4$  (no bonded hydrogens). The single elements are defined as

$$\lambda_{uv} = \frac{n_{uv}}{\bar{d}_{uv}^2} \quad u = 1, 2, 3, 4; \quad v \geq u$$

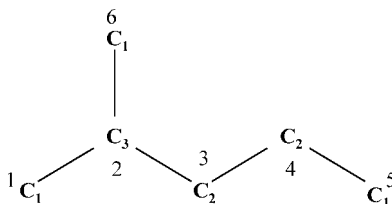
where

$$\bar{d}_{uv} = \prod_{u \leq v} (d_{i(u)j(v)})^{1/(2n_{uv})}$$

The geometric mean takes into account all the  $\rightarrow$  *topological distances* between carbon atoms  $i$  and  $j$  of types  $u$  and  $v$ ;  $n_{uv}$  is the number of possible atom pairs for a fixed combination of carbon types.  $\lambda_{uv}$  is set at zero by definition if no atom pairs with types  $u$  and  $v$  are present in the molecular graph.

**Example M2**

MDE vector for 2-methylpentane.



C <sub>1</sub> C <sub>1</sub>	C <sub>1</sub> C <sub>2</sub>	C <sub>1</sub> C <sub>3</sub>	C <sub>1</sub> C <sub>4</sub>	C <sub>2</sub> C <sub>2</sub>	C <sub>2</sub> C <sub>3</sub>	C <sub>2</sub> C <sub>4</sub>	C <sub>3</sub> C <sub>3</sub>	C <sub>3</sub> C <sub>4</sub>	C <sub>4</sub> C <sub>4</sub>
$d_{16} = 2$	$d_{13} = 2$	$d_{12} = 1$	—	$d_{34} = 1$	$d_{32} = 1$	—	—	—	—
$d_{15} = 4$	$d_{14} = 3$	$d_{62} = 1$			$d_{42} = 2$				
$d_{65} = 4$	$d_{63} = 2$	$d_{52} = 3$							
	$d_{64} = 3$								
	$d_{53} = 2$								
	$d_{54} = 1$								

$$\begin{aligned} \bar{d}_{C_1C_1} &= (2 \cdot 4 \cdot 4)^{1/(2 \cdot 3)} = 32^{1/6} = 1.7818 & \lambda_{C_1C_1} &= 3/3.1748 = 0.9449 & \lambda_{C_1C_2} &= 6/2.0398 = 2.9415 \\ \bar{d}_{C_1C_2} &= (2 \cdot 3 \cdot 2 \cdot 3 \cdot 2 \cdot 1)^{1/(2 \cdot 6)} = 72^{1/12} = 1.4282 & \lambda_{C_1C_3} &= 3/1.4422 = 2.0802 & \lambda_{C_1C_4} &= 0 \\ \bar{d}_{C_1C_3} &= (1 \cdot 1 \cdot 3)^{1/(2 \cdot 3)} = 3^{1/6} = 1.2009 & \lambda_{C_2C_2} &= 1/1 = 1 & \lambda_{C_2C_3} &= 2/1.4142 = 1.4142 \\ \bar{d}_{C_2C_2} &= 1^{1/2} = 1 & \lambda_{C_2C_4} &= 0 & \lambda_{C_3C_3} &= 0 \\ \bar{d}_{C_2C_3} &= (1 \cdot 2)^{1/(2 \cdot 2)} = 2^{1/4} = 1.1892 & \lambda_{C_3C_4} &= 0 & \lambda_{C_4C_4} &= 0 \end{aligned}$$

$$\lambda = (0.9449, 2.9451, 2.0802, 0, 1, 1.4142, 0, 0, 0, 0)$$

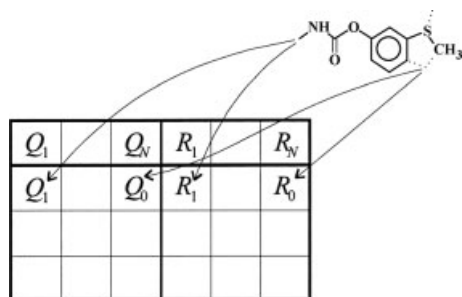
- **Molecular Diversity Analysis descriptor**  $\equiv$  *MolDiA descriptor*  $\rightarrow$  substructure descriptors ( $\odot$  structural keys)
- **molecular eccentricity**  $\rightarrow$  shape descriptors
- **molecular electronegativity**  $\rightarrow$  quantum-chemical descriptors ( $\odot$  electronic chemical potential)
- **Molecular Electronegativity Distance Vector**  $\rightarrow$  MEDV-13 descriptor
- **Molecular Electronegativity Edge Vector**  $\rightarrow$  autocorrelation descriptors
- **Molecular Electrostatic Potential**  $\rightarrow$  quantum-chemical descriptors
- **molecular electrostatic potential contour surface**  $\rightarrow$  molecular surface
- **molecular energies**  $\rightarrow$  quantum-chemical descriptors

### ■ Molecular Field Topology Analysis (MFTA)

The method of Molecular Field Topology Analysis is among the  $\rightarrow$  *hyperstructure-based QSAR techniques*. It was proposed as a “topological analogue” of the  $\rightarrow$  *CoMFA* method because it is based on topological rather than spatial alignment of structures [Zefirov, Palyulin *et al.*, 1997; Palyulin, Radchenko *et al.*, 2000; Melnikov, Palyulin *et al.*, 2007].

The quantitative description of structural features is provided by local physico-chemical parameters. First, for a set of structures of known activity (a training set), the so-called  $\rightarrow$  *molecular supergraph* (MSG) is automatically constructed. The MSG is a certain graph, such that each training set structure can be represented as its subgraph. It enables the construction of  $\rightarrow$

uniform-length descriptors for all structures in the set. To build each descriptor vector, the MSG vertices and edges corresponding, respectively, to the atoms and bonds of a given structure are assigned the values of local descriptors (e.g., atomic charge  $q$  and  $\rightarrow$  van der Waals radius  $R^{vdw}$ ) and the remaining vertices and edges are labeled with neutral descriptor values that provide a reasonable simulation of properties in an unoccupied region of space. The descriptor vector formation is illustrated in Figure M3.



**Figure M3** Descriptor vector formation in MFTA.

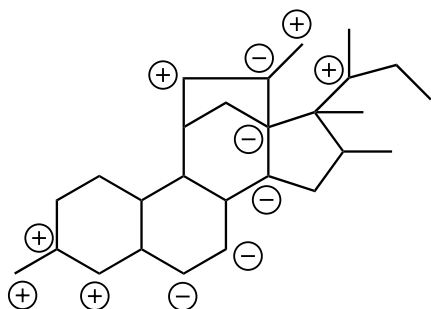
The following local descriptors are currently calculated: Gasteiger's  $\rightarrow$  atomic charge  $q$  estimated by the  $\rightarrow$  partial equalization of orbital electronegativity approach, Sanderson's  $\rightarrow$  electronegativity  $\chi^S$ , Bondi's van der Waals radius  $R^{vdw}$ , atomic contribution to the  $\rightarrow$  van der Waals molecular surface  $SA^{vdw}$ , relative steric accessibility defined as  $Ac = SA^{vdw}/SA_{free}$  (where  $SA_{free}$  is the van der Waals surface of the "free" (isolated) atom of the same type),  $\rightarrow$  electrotopological state  $S$ , atomic lipophilicity contribution  $l$  taking into account the environment of an atom, and group lipophilicity  $L_g$  defined as a sum of contributions for both a non-hydrogen atom and attached hydrogens, the ability of an atom in a given environment to be a donor and acceptor of a hydrogen bond characterized by the binding constants, local stereochemical indicator variables, and the site occupancy factors for atoms  $I_a$  and bonds  $I_b$  (which have the value 1, if a given feature is present in the structure and 0 otherwise). This set of local descriptors provides sufficient coverage of major interaction types that are important for the interaction of a ligand with a biological target. However, the set is open and can be easily extended to account for the specific features of the problem.

Since the number of descriptors is rather large, partial least squares (PLS) regression is used to analyze the descriptor-activity relationships. As a result, the quantitative characteristic of the influence on activity of each descriptor in each position, including common structural fragments, can be determined (Figure M4). Such characteristics provide a basis for designing new, potentially more active structures as well as being anchor points for spatial structure alignment.

MFTA often gives models that are comparable in quality of description and prediction to models based on the widely used classical QSAR methods and 3D approaches.

- **molecular fingerprints**  $\rightarrow$  substructure descriptors
- **molecular flexibility**  $\rightarrow$  flexibility indices
- **molecular flexibility number**  $\rightarrow$  flexibility indices





**Figure M4** Molecular supergraph of steroid data set with major atomic charge contributions into the CBG affinity.

- **molecular formula**  $\equiv$  *chemical formula*  $\rightarrow$  molecular descriptors
- **molecular fragments**  $\rightarrow$  count descriptors

### ■ molecular geometry

A molecule is the smallest fundamental group of atoms of a chemical compound that can take part in a chemical reaction. The atoms of the molecule are organized in a 3D structure; the **molecular matrix**, denoted by **M**, is a rectangular matrix  $A \times 3$  whose rows represent the molecule atoms and the columns the atom **Cartesian coordinates** ( $x, y, z$ ) with respect to any rectangular coordinate system with axes  $X, Y, Z$ . The Cartesian coordinates of a molecule usually correspond to some optimized molecular geometry obtained by the methods of  $\rightarrow$  *computational chemistry*. The molecular geometry can also be obtained from crystallographic coordinates or 2D–3D automatic converters.

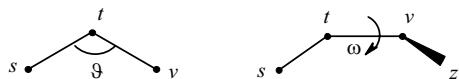
The **connectivity table** of a molecule is a rectangular table whose rows represent atoms and row entries the labels of all the bonded atoms.

Since the molecular matrix does not contain information on atom adjacencies, it usually is given as an augmented matrix **M'**, obtained by union of the molecular matrix and the connectivity table, where the first column denotes the atom type (e.g., carbon, hydrogen, chlorine atoms) and the last four columns contain the labels of the atoms connected to the  $i$ th atom:

$$\mathbf{M} = \begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \dots & \dots & \dots \\ x_A & y_A & z_A \end{vmatrix} \quad \mathbf{M}' = \begin{vmatrix} \text{at. 1} & x_1 & y_1 & z_1 & c_{11} & c_{12} & c_{13} & c_{14} \\ \text{at. 2} & x_2 & y_2 & z_2 & c_{21} & c_{22} & c_{23} & c_{24} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{at. A} & x_A & y_A & z_A & c_{A1} & c_{A2} & c_{A3} & c_{A4} \end{vmatrix}$$

Note that the last four columns of the **M'** matrix constitute the connectivity table of the molecules.

An alternative to the molecular matrix representation of a molecule is that of **internal coordinates**, where the relative position of each atom to the other atoms in the molecule is given: these coordinates are bond distances, bond angles, and torsion angles. **Bond distances**  $r_{st}$  are the interatomic distances between bonded atoms (usually expressed in Ångström); **bond angles**  $\vartheta_{stv}$  are plane angles among triples of connected atoms ( $s, t, v$ ) within the molecule; **torsion angles**  $\omega_{stvz}$  are dihedral angles among quadruples of connected atoms ( $s, t, v, z$ ) (Figure M5). Note that



**Figure M5** Bond and torsion angles.

bond distances and bond angles are less sensitive to conformational change than interatomic distances and torsion angles.

Internal coordinates are collected in the so-called **Z-matrix**, which is a rectangular matrix, whose rows are the atoms, defined as

$$Z = \begin{array}{cccccc} \text{at. 1} & & & 0 & 0 & 0 \\ \text{at. 2} & r_{12} & & 1 & 0 & 0 \\ \text{at. 3} & r_{23} & \vartheta_{321} & 2 & 1 & 0 \\ \text{at. 4} & r_{34} & \vartheta_{432} & \omega_{4321} & 3 & 2 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \text{at. A} & r_{As} & \vartheta_{Ast} & \omega_{Astv} & s & t & v \end{array}$$

where  $r$ ,  $\vartheta$ , and  $\omega$  are the molecular internal coordinates considered among the  $\rightarrow$  *geometrical descriptors*. The last three columns contain the labels of atoms involved in bonds, bond angles, and torsion angles.

Other simple geometrical descriptors are **interatomic distances**  $r_{st}$  between pairs of atoms  $s$  and  $t$ . Interatomic distances are distinguished into **intramolecular interatomic distances**, that is, distances between any pair of atoms ( $s$ ,  $t$ ) within the molecule and **intermolecular interatomic distances**, that is, distances between atoms of a molecule and atoms of a receptor structure, a reference compound or another molecule. While classical computational chemistry describes molecular geometry in terms of three-dimensional Cartesian coordinates or internal coordinates, the  $\rightarrow$  *distance geometry* (DG) method takes the interatomic distances as the fundamental coordinates of molecules, exploiting their close relationship to experimental quantities and molecular energies. Moreover, for series of congeneric compounds, bond distances, optimized by quantum-chemistry approaches and selected by genetic algorithms, were directly used as molecular descriptors in QSAR studies [Smith and Popelier, 2004].

The molecular matrix **M** and the matrix **Z** are the natural starting point for the calculation of several 3D atomic and molecular descriptors, such as  $\rightarrow$  *quantum-chemical descriptors*,  $\rightarrow$  *molecular interaction fields*,  $\rightarrow$  *EVA descriptors*,  $\rightarrow$  *WHIM descriptors*,  $\rightarrow$  *GETAWAY descriptors*,  $\rightarrow$  *CoMMA descriptors*,  $\rightarrow$  *Compass descriptors*, and  $\rightarrow$  *molecular surface descriptors*.

Another common source of geometrical descriptors is the geometry matrix.

The **geometry matrix** (or **geometric distance matrix**) of a molecule, denoted by **G**, obtained from the molecular matrix **M**, is a square symmetric matrix  $A \times A$ , where each entry  $r_{st}$  is the **geometric distance** calculated as the  $\rightarrow$  *Euclidean distance* between the atoms  $s$  and  $t$ :

$$G \equiv \begin{array}{cccc} 0 & r_{12} & \dots & r_{1A} \\ r_{21} & 0 & \dots & r_{2A} \\ \dots & \dots & \dots & \dots \\ r_{A1} & r_{A2} & \dots & 0 \end{array}$$

Diagonal entries are always zero. Geometric distances are intramolecular interatomic distances.

Like the molecular matrix, the geometry matrix contains information about molecular configurations and conformations; however, the geometry matrix does not contain information about atom connectivity. Thus, for several applications, it is accompanied by a connectivity table where, for each atom, there is listed the identification number of the atoms bonded to it. The geometry matrix can also be calculated on geometry-based standardized bond lengths and bond angles and derived by embedding a graph on a regular two-dimensional or three-dimensional grid; in these cases, the geometry matrix is often referred to as the **topographic matrix T** and the interatomic distance to the **topographic distance** [Balaban, 1997a]. Depending on the kind of grid used for graph embedding, different topographic matrices can be obtained.

The **bond length-weighted adjacency matrix** is obtained from the geometry matrix **G** as [Mihalić, Nikolić *et al.*, 1992]

$${}^b\mathbf{A} = \mathbf{G} \otimes \mathbf{A}$$

where  $\otimes$  indicates the  $\rightarrow$  *Hadamard matrix product* and **A** is the  $\rightarrow$  *adjacency matrix*.

From the geometry matrix used to represent a  $\rightarrow$  *molecular graph*, a number of  $\rightarrow$  *local vertex invariants* and related  $\rightarrow$  *graph invariants*, called **topographic indices**, can be derived [Randić and Wilkins, 1979b; Randić, 1988a; Randić, Jerman-Blazic *et al.*, 1990; Diudea, Horvath *et al.*, 1995b; Randić and Razingar, 1995a; Balaban, 1997b].

Analogously to the  $\rightarrow$  *vertex distance degree*, the *i*th row sum of the geometry matrix is called **geometric distance degree**  ${}^G\sigma_i$  (or **Euclidean degree**) [Balasubramanian, 1995b]:

$${}^G\sigma_i = \sum_{j=1}^A r_{ij}$$

This is a local vertex invariant used, for example, in the definition of the  $\rightarrow$  *3D-connectivity indices*  $\chi\chi$  and the  $\rightarrow$  *Euclidean connectivity index*. In general, the row sum of this matrix represents a measure of the centrality of an atom; atoms that are close to the  $\rightarrow$  *center of the molecule* have smaller atomic sums, whereas those far from the center have large atomic sums. The smallest and the largest row sums give the extreme values of the first eigenvalue of the geometry matrix; therefore, when all the atoms are equivalent, that is, the distance degrees are all the same, the geometric distance degree yields exactly the first eigenvalue. The average sum of all geometric distance degrees is a molecular invariant called **average geometric distance degree**, that is,

$${}^G\bar{\sigma} = \frac{1}{A} \cdot \sum_{i=1}^A {}^G\sigma_i = \frac{1}{A} \cdot \sum_{i=1}^A \sum_{j=1}^A r_{ij}$$

whereas the half sum of all geometric distance degrees is another molecular descriptor called **3D-Wiener index** by analogy with the  $\rightarrow$  *Wiener index* calculated from the topological distance matrix. The 3D Wiener index is calculated as

$${}^{3D}W_H \equiv Wi(\mathbf{G}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A r_{ij}$$

where  $r_{ij}$  is the interatomic distance between the *i*th and *j*th atom [Mekenyan, Peitchev *et al.*, 1986a, 1986b; Bogdanov, Nikolić *et al.*, 1989, 1990; Randić, Jerman-Blazic *et al.*, 1990]. This index is obviously more discriminant than the 2D Wiener index as it accounts for spatial molecular geometry; it shows different values for different molecular conformations, the largest values corresponding to the most extended conformations, the smallest to the most compact

conformations. Therefore, it is considered among  $\rightarrow$  *shape descriptors* since it decreases with increasing sphericity of a structure [Nikolić, Trinajstić *et al.*, 1991]. The 3D Wiener index can be calculated both considering  ${}^3\text{D}W_{\text{H}}$  and not considering  ${}^3\text{D}W$  hydrogen atoms [Basak, Gute *et al.*, 1999a]. Moreover, a strictly related molecular descriptor is the  $\rightarrow$  *bond length-weighted Wiener index* calculated by using as the distance between two atoms the sum of the bond lengths along the shortest path.

A 3D local vertex invariant based on the geometric distance was proposed as [Toropov, Toropova *et al.*, 1998; Krenkel, Castro *et al.*, 2002]

$$3\text{D}W_i = \sum_{j=1}^A (1 - a_{ij}) \cdot \exp(r_{ij}^{-2}) \quad j \neq i$$

where the summation accounts only for contributions from the pairs of nonadjacent atoms,  $a_{ij}$  being the elements of the adjacency matrix equal to one only for pairs of adjacent atoms, and zero otherwise. The exponential form of the distance was chosen from a series of terms approximating the attracting interatomic potentials.

From these local invariants,  $\rightarrow$  *Zagreb indices*,  $\rightarrow$  *connectivity-like indices*, and  $\rightarrow$  *Wiener-type indices* were derived as

$$\begin{aligned} 3\text{D}M_1 &= \sum_{i=1}^A 3\text{D}W_i & 3\text{D}M_2 &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (3\text{D}W_i \cdot 3\text{D}W_j) \\ 3\text{D}^0\chi &= \sum_{i=1}^A (3\text{D}W_i)^{-1/2} & 3\text{D}^1\chi &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (3\text{D}W_i \cdot 3\text{D}W_j)^{-1/2} \\ 3\text{D}Wi &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A \exp(r_{ij}) \end{aligned}$$

where  $A$  is the number of molecule atoms.

Molecular descriptors based on this kind of local vertex invariant are called **MIS indices**, being defined in the framework of the **Method of Ideal Symmetry** (MIS), based on a partial optimization procedure of the molecular geometry, where bond lengths and bond angles are kept fixed and only free rotations around C–C bonds are varied [Toropov, Toropova *et al.*, 1994].

The maximum value entry in the  $i$ th row of the geometry matrix is a local descriptor called **geometric eccentricity**  ${}^G\eta_i$ , representing the longest geometric distance from the  $i$ th atom to any other atom in the molecule:

$${}^G\eta_i = \max_j(r_{ij})$$

From the eccentricity definition, **geometric radius**  ${}^GR$  and **geometric diameter**  ${}^GD$  can immediately characterize a molecule. The radius of a molecule is defined as the minimum geometric eccentricity and the diameter is defined as the maximum geometric eccentricity in the molecule, according to the following:

$${}^GR = \min_i({}^G\eta_i) \quad \text{and} \quad {}^GD = \max_i({}^G\eta_i)$$

These parameters are  $\rightarrow$  *size descriptors* also depending on the molecular shape ( $\rightarrow$  *Petitjean shape indices*), such as their topological counterpart, that is,  $\rightarrow$  *topological radius* and  $\rightarrow$  *topological diameter*.

Derived from the geometry matrix, the **neighborhood geometry matrix** (or **neighborhood Euclidean matrix**), denoted as  ${}^N\mathbf{G}$ , was also proposed as [Bajzer, Randić *et al.*, 2003]

$$[{}^N\mathbf{G}]_{ij} = \begin{cases} r_{ij} & \text{if } r_{ij} \leq R_t \\ 0 & \text{if } r_{ij} > R_t \end{cases}$$

where  $R_t$  is a user-defined distance threshold. This matrix was used to calculate descriptors of  $\rightarrow$  *proteomics maps* by the additional constraint that the matrix element  $i$ - $j$  is set at zero also for nonconnected protein spots.

The **reciprocal geometry matrix**, denoted as  $\mathbf{G}^{-1}$ , is obtained from the geometry matrix as the following:

$$[\mathbf{G}^{-1}]_{ij} = \begin{cases} r_{ij}^{-1} & i \neq j \\ 0 & i = j \end{cases}$$

In the same way, the **reciprocal topographic matrix**, denoted as  $\mathbf{T}^{-1}$ , is defined in terms of the reciprocal of topographic distances instead of the reciprocal of geometric distances.

### Example M3

Geometry matrix  $\mathbf{G}$ , geometric distance degrees  ${}^G\sigma$ , eccentricities  ${}^G\eta$ , geometric radius  ${}^G R$  and diameter  ${}^G D$  for 2-methylpentane.  ${}^G\bar{\sigma}$  and  ${}^{3D}W$  are the average geometric distance degree and the 3D-Wiener index, respectively.

Atom	1	2	3	4	5	6	${}^G\sigma$	${}^G\eta$
1	0	1.519	2.504	3.856	5.014	2.498	15.391	5.014
2	1.519	0	1.530	2.521	3.864	1.521	10.955	3.864
3	2.504	1.530	0	1.521	2.509	2.507	10.571	2.509
4	3.856	2.521	1.521	0	1.511	3.038	12.447	3.856
5	5.014	3.864	1.511	1.511	0	4.348	17.246	5.014
6	2.498	1.521	3.038	3.038	4.348	0	13.912	4.348

$${}^G R = \min_i ({}^G\eta_i) = 2.509 \quad {}^G\bar{\sigma} = \frac{1}{6} \cdot (15.391 + 10.955 + 10.571 + 12.447 + 17.246 + 13.912) = 13.420$$

$${}^G D = \max_i ({}^G\eta_i) = 5.014 \quad {}^{3D}W = \frac{1}{6} \cdot (15.391 + 10.955 + 10.571 + 12.447 + 17.246 + 13.912) = 40.261$$

From the geometry matrix, the usual  $\rightarrow$  *graph invariants* can be calculated such as  $\rightarrow$  *characteristic polynomial*,  $\rightarrow$  *spectral indices*,  $\rightarrow$  *ID numbers*,  $\rightarrow$  *3D-Balaban index*,  $\rightarrow$  *3D-Schultz index*, and so forth [Randić, 1988b; Nikolić, Trinajstić *et al.*, 1991]. It is noteworthy that all these indices despite their topological counterparts are sensitive to molecular geometry. Moreover, geometry matrix is used for the calculation of  $\rightarrow$  *size descriptors* and  $\rightarrow$  *3D-MorSE descriptors*.

Important derived matrices are the powers of the geometry matrix, used to define  $\rightarrow$  *molecular profiles* descriptors. Moreover, **distance/distance matrices**, denoted as  $\mathbf{D}/\mathbf{D}$ , were defined as  $\rightarrow$  *quotient matrices* in terms of geometric  $r_{ij}$  or topographic distances  $t_{ij}$  and  $\rightarrow$

topological distances  $d_{ij}$  to unify 2D and 3D information about the structure of molecules [Randić, 1994, 1999]:

$$\begin{aligned} [\mathbf{G}/\mathbf{D}]_{ij} &= \begin{cases} \frac{r_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} & [\mathbf{T}/\mathbf{D}]_{ij} &= \begin{cases} \frac{t_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \\ [\mathbf{D}/\mathbf{G}]_{ij} &= \begin{cases} \frac{d_{ij}}{r_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} & [\mathbf{D}/\mathbf{T}]_{ij} &= \begin{cases} \frac{d_{ij}}{t_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \end{aligned}$$

The **geometric distance/topological distance quotient matrix**, denoted as  $\mathbf{G}/\mathbf{D}$ , is a square symmetric matrix  $A \times A$ ,  $A$  being the number of molecule atoms, whose entries are the quotient of the corresponding elements of the molecular geometry matrix  $\mathbf{G}$  and the graph  $\rightarrow$  distance matrix  $\mathbf{D}$ . An alternative to the geometric distance/topological distance quotient matrix is the **topographic distance/topological distance quotient matrix** ( $\mathbf{T}/\mathbf{D}$ ), derived by using the  $\rightarrow$  topographic matrix  $\mathbf{T}$  instead of the geometry matrix. Note that in the original papers, both these matrices were indifferently referred to as distance/distance matrix and denoted by  $\mathbf{DD}$ .

The **topological distance/geometric distance quotient matrix**, denoted by  $\mathbf{D}/\mathbf{G}$ , is the reciprocal matrix of the  $\mathbf{G}/\mathbf{D}$  matrix, and the **topological distance/topographic distance quotient matrix**, denoted by  $\mathbf{D}/\mathbf{T}$ , the reciprocal matrix of the  $\mathbf{T}/\mathbf{D}$  matrix.

The row sums of these matrices contain information on the molecular folding; in effect, in highly folded structures, they tend to be relatively small as the interatomic distances are small while the topological distances increase as the size of the structure increases. Therefore, the average row sum is a molecular invariant called **average distance/distance degree**, that is,

$$ADDD = \frac{1}{A} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{r_{ij}}{d_{ij}} \quad j \neq i$$

while the half sum of all distance/distance matrix entries is another molecular descriptor called **D/D index**, that is,

$$D/D \equiv Wi(\mathbf{G}/\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \frac{r_{ij}}{d_{ij}} \quad j \neq i$$

where  $Wi$  is the  $\rightarrow$  Wiener operator.

From the largest eigenvalue of the distance/distance matrix a  $\rightarrow$  folding degree index was also defined.

Other matrices that combine topological and geometrical information are **distance–distance combined matrices**, which are defined in terms of geometric ( $r_{ij}$ ) or topographic ( $t_{ij}$ ) distances as [Janežič, 2007]:

$$\begin{aligned} [\mathbf{G} \wedge \mathbf{D}]_{ij} &= \begin{cases} r_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases} & [\mathbf{T} \wedge \mathbf{D}]_{ij} &= \begin{cases} t_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases} \\ [\mathbf{D} \wedge \mathbf{G}]_{ij} &= \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ r_{ij} & \text{if } i > j \end{cases} & [\mathbf{D} \wedge \mathbf{T}]_{ij} &= \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ t_{ij} & \text{if } i > j \end{cases} \end{aligned}$$

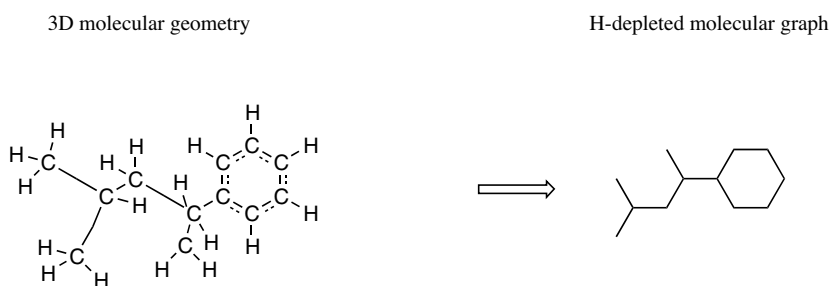
where  $\mathbf{D}$  is the topological distance matrix and  $d_{ij}$  the topological distance between vertices  $v_i$  and  $v_j$ ;  $\mathbf{G} \wedge \mathbf{D}$  is the **geometric distance–topological distance combined matrix**,  $\mathbf{T} \wedge \mathbf{D}$  is the **topographic distance–topological distance combined matrix**,  $\mathbf{D} \wedge \mathbf{G}$  is the **topological distance–geometric distance combined matrix**, and  $\mathbf{D} \wedge \mathbf{T}$  is the **topological distance–topographic distance combined matrix**. Note that  $\mathbf{D} \wedge \mathbf{G}$  and  $\mathbf{D} \wedge \mathbf{T}$  are the transpose matrices of  $\mathbf{G} \wedge \mathbf{D}$  and  $\mathbf{T} \wedge \mathbf{D}$ .

📖 [Turro, 1986; Mihalić and Trinajstić, 1991; Mihalić, Nikolić *et al.*, 1992; Kunz, 1993, 1994; Warthen, Schmidt *et al.*, 1993; Balasubramanian, 1995b; Estrada and Ramirez, 1996; Zhu and Klein, 1996; Laidboeur, Cabrol-Bass *et al.*, 1997; Randić and Razinger, 1997; Ivanciuc, Ivanciuc *et al.*, 1998b; Ivanciuc and Ivanciuc, 1999; Tao and Lu, 1999; Blatova, Blatov *et al.*, 2001, 2002; Imre, Veress *et al.*, 2003; Todeschini and Consonni, 2003; Wisniewski, 2003; Wang, Wang *et al.*, 2006]

■ **molecular graph** ( $\equiv$  *structural graph, constitutional graph*)

It is a nondirected connected  $\rightarrow$  *graph*  $G$ , which represents a chemical compound, that is, a graph where vertices and edges are chemically interpreted as atoms and covalent bonds [Harary, 1969a, 1969b; Balaban and Harary, 1976; Rouvray and Balaban, 1979; Rouvray, 1990a; Bonchev and Rouvray, 1991; Trinajstić, 1992]. A molecular graph obtained excluding all the hydrogen atoms is called **H-depleted molecular graph** (or **hydrogen-depleted molecular graph** or **Labeled Hydrogen-Suppressed molecular Graph**, LHSG), whereas a molecular graph where also hydrogens are graph vertices is called **H-filled molecular graph** (or **hydrogen-included molecular graph** or **hydrogen-filled molecular graph** or **Labeled Hydrogen-Filled molecular Graph**, LHFG).

Such a graph depicts the connectivity of atoms in a molecule irrespective of the metric parameters such as equilibrium  $\rightarrow$  *interatomic distances* between nuclei,  $\rightarrow$  *bond angles*, and  $\rightarrow$  *torsion angles*, representing the 3D  $\rightarrow$  *molecular geometry*. Thus, a molecular graph is a  $\rightarrow$  *topological representation* of the molecule, and it is from this that several  $\rightarrow$  *molecular descriptors* are derived (Figure M6).

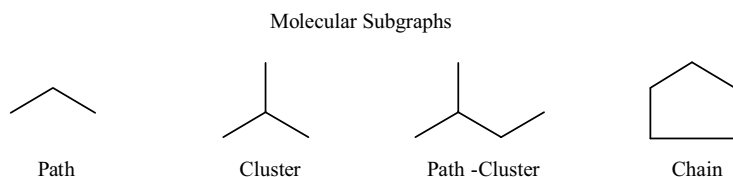


**Figure M6** The transition from 3D geometry to 2D topology.

Some vertices of the H-depleted molecular graph can be more precisely defined as the **hydride group**, which is a heavy atom plus its bonded hydrogens. For example, hydride groups are  $-\text{CH}_3$ ,  $-\text{CH}_2-$ ,  $=\text{NH}$ ,  $-\text{NH}_2$ , and  $-\text{OH}$ .

A **molecular subgraph** is a subset of atoms and related bonds, which is in itself a valid graph usually representing molecular fragments and functional groups.

There are four commonly used subgraph types: **path subgraph**, **cluster subgraph**, **path-cluster subgraph**, and **chain subgraph** (or *Ring*), emphasizing different aspects of atom connectivity within the molecule. They are defined according to the following rules: (1) if the subgraph contains a cycle, it is of type Chain (CH); otherwise, (2) if all  $\rightarrow$  *vertex degrees* in the subgraph (not in the whole graph) are either greater than 2 or equal to 1, the subgraph is of type Cluster (C); otherwise, (3) if all vertex degrees in the subgraph are either equal to 2 or 1, the subgraph is of type Path (P); otherwise, (4) the subgraph is of type Path-Cluster (PC) (Figure M7).



**Figure M7** Elementary molecular subgraphs.

The **order of a subgraph** is the number of edges within it. Note that subgraphs of order 0 are considered of type Path, subgraphs of order 1 and 2 are only of type Path, and subgraphs of order 3 can be of type Path, Cluster, or Chain only.

Referring to the subgraph order,  $r$ th order indices can be defined as  $\rightarrow$  *count descriptors*, that is, the number of  $r$ th order subgraphs in the graph  $G$ . The zero order index is simply the  $\rightarrow$  *atom number*  $A$ , that is, the number of graph vertices; first order index is the  $\rightarrow$  *bond number*  $B$ , that is, the number of graph edges; second order index is the  $\rightarrow$  *connection number*, that is, the second order  $\rightarrow$  *molecular path count*; third order indices are the number of paths of length 3, the number of three-edge clusters, and the number of three-edges cycles.

The total number  $K$  of connected subgraphs of a molecular graph  $G$  is a very simple measure of  $\rightarrow$  *molecular complexity*, obviously referring only to structural complexity of the molecule; it is called **total subgraph count**.

The simplest form to represent the chemical information contained in a molecular graph is by  $\rightarrow$  *graph-theoretical matrices*. Examples are  $\rightarrow$  *adjacency matrix*  $A$ ,  $\rightarrow$  *edge adjacency matrix*  $E$ , vertex  $\rightarrow$  *distance matrix*  $D$ ,  $\rightarrow$  *edge distance matrix*  $E^D$ ,  $\rightarrow$  *incidence matrix*  $I$ ,  $\rightarrow$  *Wiener matrix*  $W$ ,  $\rightarrow$  *Hosoya Z-matrix*  $Z$ ,  $\rightarrow$  *Cluj matrices*  $CJ$ ,  $\rightarrow$  *detour matrix*  $\Delta$ ,  $\rightarrow$  *Szeged matrix*  $SZ$ ,  $\rightarrow$  *geometric distance/topological distance quotient matrix*  $G/D$ , and  $\rightarrow$  *detour-distance combined matrix*  $\Delta/D$ .

A **reduced graph** is a molecule representation aimed at the storage and retrieval of generic chemical structures. The internal representation of the molecule is hierarchically tree structured as a topological graph, the chemical nature of the various parts of the generic structure being represented in the vertices of the graph and the information about their connections and relationships in its edges. Since information on the chemical nature of each part is predominantly based on conventional  $\rightarrow$  *connectivity tables*, the whole is a sort of superconnection table or connection table of connection tables, and is referred to as an **Extended Connection Table Representation (ECTR)**. Within the ECTR, each vertex is called a partial structure and each edge a gate [Barnard, Lynch *et al.*, 1982; Gillet, Downs *et al.*, 1991; Gillet, Willett *et al.*, 2003].

Because in a reduced graph, groups of atoms within the structure are collapsed together to form single vertices and smaller graphs are obtained, more quick searching can be performed on large molecule data sets, using any of the conventional methods [Barnard, 1993].



The reduced graph may contain vertices representing the cyclic and acyclic portions of the molecule or contiguous groups of carbon or heteroatoms.

Common structural features between molecules are searched for by means of graph-theoretical approaches, such as the determination of the maximal  $\rightarrow$  *cliques* of the docking graph. A clique in the docking graph corresponds to a grouping of functional groups in the original reduced graphs, where all the intragrouping distances are the same in both original graphs.

An example of a reduced graph is a molecular graph composed of weighted edges and with vertex number equal to the number of functional groups perceived by using predefined rules [Takahashi, Sukekawa *et al.*, 1992].

Once all the functional atomic groups are perceived, the interrelations between them are checked. The relationships evaluated are described in terms of matrix expression, and they are divided into the following two cases:

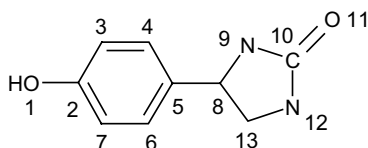
- (1) The case in which two functional groups are partially overlapping (overlap matrix).
- (2) The case in which one of the functional atomic groups is completely included by another one (inclusion matrix).

For the former case, the relationship is described in the overlap matrix, and for the latter, the relationship is described in the inclusion matrix to avoid the duplication of vertices in the reduced graph representation. They are used to determine the  $\rightarrow$  *topological distance* between functional groups.

Therefore, this reduced graph representation is a graph constituted by a number of vertices equal to the number of perceived functional groups and weighted edges, whose weights correspond to the shortest topological distance between the different functional groups. If the structure has ring(s), there are several possible paths that can be drawn simultaneously between functional groups. In such a case, some of the edges are weighted with multiple values.

#### Example M4

Derivation of a reduced graph.



- A : Hydroxy {1}  
 B : Benzene {2, 3, 4, 5, 6, 7}  
 C : Amido {9, 10, 11}  
 D : Amido {12, 10, 11}  
 E : Urea {9, 10, 11, 12}

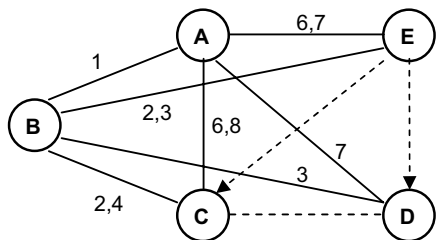
Overlap matrix

	A	B	C	D	E
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	1	1
D	0	0	1	1	1
E	0	0	1	1	1

Inclusion matrix

	A	B	C	D	E
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	1	1	1

The reduced graph obtained by defining the functional groups outlined above can be drawn as the following:



Other approaches to generate and/or manage reduced graphs were proposed, such as the SwiFT index method, created to identify redundant subtrees [Fischer and Rarey, 2007], and homeomorphically reduced graph formed by deleting all atoms of connectivity 2 [Cringean and Lynch, 1989].

The **feature tree** is a representation of a molecule similar to a reduced graph. A feature tree represents hydrophobic fragments and functional groups of the molecule and the way these groups are linked together [Rarey and Dixon, 1998; Rarey and Stahl, 2001]. The vertices of the feature tree are molecular fragments and edges connect vertices that represent fragments that are connected in the simple molecular graph. Moreover, each vertex in the tree is associated with a set of features representing chemical properties of the molecular fragment corresponding to the vertex.

Feature trees are used in  $\rightarrow$  *similarity/diversity* analysis of compounds: the comparison of the feature trees of two compounds is based on matching subtrees of the two feature trees onto each other.

📖 [Mason, 1943; Gutman and Trinajstić, 1972; Gutman *et al.*, 1975; Gutman, Rusčić *et al.*, 1975; Randić, 1975b; Balaban, 1976d, 1978a, 1985a, 1985b, 1987, 1991, 1993d, 1995a; Polanski and Rouvray, 1976a, 1976b; Balaban and Rouvray, 1980; Mekenyan, Bonchev *et al.*, 1980, 1981; Trinajstić, Jericevic *et al.*, 1983; Hosoya, 1986; Trinajstić, Klein *et al.*, 1986; Hansen and Jurs, 1988a; Dias, 1993; Klein, 1997; Bytautas and Klein, 1998, 1999; John, Mallion *et al.*, 1998; Bonchev and Rouvray, 2000; Kruja, Marks *et al.*, 2002; Ivanciuc, 2003f; Kerber, Laue *et al.*, 2004; García-Domenech, Gálvez *et al.*, 2008]

➤ **molecular graphics and modeling descriptors**  $\rightarrow$  molecular graphics descriptors

#### ■ molecular graphics descriptors

These are descriptors derived from high-quality 2D projections of molecules or molecular aggregates obtained by current molecular graphic techniques, which can be an extensive source of quantitative information on molecular properties [Kiralj and Ferreira, 2003a].

In general, quantities directly “measured” from pictures using some digital or analogue technique can be 1D (such as molecular dimensions), 2D (such as surface areas), or 3D (such as molecular volumes).

Combination of these descriptors with some structural information from other sources (such as data from experimental structure determination or other descriptors useful in molecular modeling) yields composite functions, which are called **molecular graphics–structural descriptors** and **molecular graphics and modeling descriptors**, respectively. Both classes of descriptors can be global (describing the entire molecule) or local (being related to some molecular fragment).

- **molecular graphics–structural descriptors** → molecular graphics descriptors
- **molecular holographic distance vector** → MEDV-13 descriptor
- **molecular holograms** → substructure descriptors (⊙ fingerprints)
- **molecular ID numbers** → ID numbers
- **molecular influence matrix** → GETAWAY descriptors

### ■ **molecular interaction fields** ( $\equiv$ *interaction fields*)

A molecular interaction field is a scalar field of → *interaction energy values* between a molecule, whose → *molecular geometry* is known, and a → *probe* [Wade, 1993; Andrews, 1993; Leach, 1996]. For QSAR studies, molecular interaction fields are calculated using one or more probes for a number of compounds previously aligned by specific → *alignment rules* and embedded in the same fixed → *grid*, that is, a regular 3D array of  $N_G$  points, each point **p** being characterized by grid coordinates ( $x, y, z$ ). The interaction energy values are calculated by moving the probe in each grid point.

Depending on the selected probe and the defined potential energy function, several molecular interaction fields can be calculated. The most common are *steric fields* and *electrostatic fields*, sometimes referred to as **CoMFA fields** because originally implemented in → *CoMFA*. Several interaction fields are actually calculated in → *GRID method*.

Derived from a topological approach → *E-state fields* and → *HE-fields* were defined.

The **enthalpic fields** are all of the molecular interaction fields accounting for enthalpic contributions to the free energy of ligand–receptor binding, such as *steric fields* and *electrostatic fields*. On the other hand, the **entropic fields** are all of the molecular interaction fields accounting for entropic contributions to the free energy of ligand–receptor binding. The entropy of binding is related to hydrophobic interactions between nonpolar ligand and receptor lipophilic chemical groups after the release of water molecules formerly structured around the receptor groups, and to the loss of conformational freedom due to ligand immobilization at the binding site. The entropy of binding is mainly modeled by *hydrophobic fields* or *hydrogen bonding fields*; however, sometimes also the degrees of torsional freedom in the molecule were considered to account for the entropy change resulting from the reduced conformational freedom of the ligand in the receptor complex [Greco, Novellino *et al.*, 1997].

Some molecular interaction fields are listed below.

#### • **steric interaction fields** ( $\equiv$ *van der Waals interaction fields*)

A steric interaction field is obtained calculating the van der Waals interaction energy  $E_{vdw}$  between probe and target in each grid point [Kim, 1992b]. Different potential energy functions were proposed to model van der Waals interactions between atoms. The most common are *Lennard–Jones potential*, *Buckingham potential*, and *Hill potential* [Leach, 1996].

The general formula of **Lennard–Jones 6–12 potential function** [Lennard–Jones, 1924, 1929] is

$$E_{vdw} = \sum_s \sum_t \left( \frac{A_{st}}{r_{st}^{12}} - \frac{C_{st}}{r_{st}^6} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule;  $r_{st}$  is the interatomic distance between the  $s$ th atom of the probe and the  $t$ th atom of the target;  $A$  and  $C$  are two functions defined as

$$A_{st} = \sqrt{\epsilon_s \cdot \epsilon_t} \cdot (R_s + R_t)^{12} \quad \text{and} \quad C_{st} = 2 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot (R_s + R_t)^6$$

where  $\epsilon$  is the well depth and  $R$  one half the separation at which the energy passes through a minimum (i.e., the  $\rightarrow$  *van der Waals radius*). The Lennard–Jones potential is characterized by an attractive component that varies as  $r^{-6}$  and a repulsive component that varies as  $r^{-12}$ . The energy function modeling the steric repulsion between pairs of atoms becomes large and positive at interatomic distances  $r$  less than the sum of the van der Waals radii of the probe atom and the target atom.

The **Buckingham potential function** [Buckingham, 1938] is defined in an exponential form as

$$E_{vdw} = \sum_s \sum_t \left( A \cdot \exp^{-B \cdot r_{st}} - \frac{C}{r_{st}^6} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule;  $r_{st}$  is the interatomic distance between the  $s$ th atom of the probe and the  $t$ th atom of the target;  $A$ ,  $B$ , and  $C$  are functions of the well depth  $\epsilon$ , the van der Waals radius  $R$ , and an adjustable parameter  $\alpha$ . The exponential energy function is commonly used for small molecules, the Lennard–Jones 12–6 function for macromolecules.

The **Hill potential function** is an exponential function defined as

$$E_{vdw} = \sum_s \sum_t \left[ -2.25 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot \left( \frac{R_s + R_t}{r_{st}} \right)^6 + 8.28 \cdot 10^5 \cdot \sqrt{\epsilon_s \cdot \epsilon_t} \cdot \exp \left( -\frac{r_{st}}{0.073 \cdot (R_s + R_t)} \right) \right]$$

where  $\epsilon$  is the well depth and  $R$  the van der Waals radius. The coefficients were determined by fitting them to data for the rare gases [Hill, 1948].

#### • electrostatic interaction fields

These are molecular interaction fields obtained by calculating electrostatic interaction energy  $E_{el}$  between probe and target in each grid point. Besides the  $\rightarrow$  *molecular electrostatic potential* (MEP), the most common energy function for electrostatic interactions is the **Coulomb potential energy function** defined as

$$E_{el} = \sum_s \sum_t \frac{q_s \cdot q_t}{4\pi \cdot \epsilon_0 \cdot \epsilon_m \cdot r_{st}}$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule;  $r_{st}$  is the interatomic distance between the  $s$ th atom of the probe and the  $t$ th atom of the target;  $q$  is the  $\rightarrow$  *partial atomic charge*; and  $\epsilon_0$  is the permittivity of the free space and  $\epsilon_m$  is the relative dielectric constant of the surrounding medium.

The **GRID electrostatic energy function** was proposed to account for the dielectric discontinuity between a solute and the solvent as [Goodford, 1985]

$$E_{el} = \sum_s \sum_t \frac{q_s \cdot q_t}{K \cdot \zeta} \cdot \left( \frac{1}{r_{st}} + \frac{(\zeta - \epsilon)/(\zeta + \epsilon)}{\sqrt{(r_{st}^2 + 4 \cdot s_s \cdot s_t)}} \right)$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule;  $r_{st}$  is the interatomic distance between the  $s$ th atom of the probe and the  $t$ th atom of the target;  $q$  is the partial atomic charge;  $K$  is a constant;  $\zeta$  and  $\epsilon$  are the relative dielectric constants of the protein and the target solution phases, respectively;  $s_s$  and  $s_t$  are the nominal depths at which the probe atom and the target atom are respectively buried in the target phase. These depths are calculated by counting the number of neighboring target atoms within a distance of 4 Å and translating this into an equivalent depth using a calibrated scale.

#### • molecular orbital fields

These are fields restricted to the regions occupied by selected molecular orbitals; of particular interest are fields related to the  $\rightarrow$  *highest occupied molecular orbital* (HOMO) and to the  $\rightarrow$  *lowest unoccupied molecular orbital* (LUMO) [Navajas, Poso *et al.*, 1996; Oprea and Waller, 1997; Durst, 1998].

Molecular orbital fields are descriptors particularly useful when an ionic or charge transfer reaction is part of the ligand–receptor interaction; in this case, electrostatic fields are not able to fully represent the electronic characteristics of molecules.

To calculate a molecular orbital field, semiempirical single-point calculations are performed on the molecule-optimized geometry and the electron density at each grid point in the region of the selected orbital is determined.

#### • hydrophobic fields

These are molecular descriptors based on hydrophobic interaction energy between nonpolar surfaces of ligand and receptor. The energy of hydrophobic interactions derives from the disruption of the water structure around nonpolar surfaces resulting in a gain of entropy [Abraham and Kellogg, 1993].

**Kellogg and Abraham interaction field**, also called **Hydropathic Interactions** (HINT), is a hydrophobic field calculated by  $\rightarrow$  *Leo–Hansch hydrophobic fragmental constants* scaled by surface area and a distance-dependent function [Kellogg, Semus *et al.*, 1991; Kellogg and Abraham, 1992; Abraham and Kellogg, 1993]. **Hydropathy** is a term used in structural molecular biology to represent the hydrophobicity of amino acid side chains.

The hydropathic field in each grid point is calculated as

$$E_{hy} = \sum_s \sum_t (SA_s \cdot h_s \cdot SA_t \cdot h_t \cdot R_{st} + R'_{st})$$

where the first sum runs over all probe atoms and the second over all atoms of the target molecule;  $SA$  is the atomic  $\rightarrow$  *solvent-accessible surface area*,  $h$  the hydropathic atom constant, and  $R_{st}$  and  $R'_{st}$  are functions of the interatomic distance  $r_{st}$  between the  $s$ th atom of the probe and the  $t$ th atom. The function  $R_{st}$  scales the product between solvent-accessible surface area and hydropathic constant with a distance usually defined as

$$R_{st} = I_{st} \cdot \exp^{-r_{st}}$$

where  $I_{st}$  is a sign-flip function recognizing acid–base interactions. The function  $R'_{st}$  is a Lennard–Jones-type potential accounting for close contacts of atoms by van der Waals radius term:

$$R'_{st} = A \cdot \epsilon_{st} \cdot \left[ \left( \frac{R_s + R_t}{r_{st}} \right)^{12} - 2 \cdot \left( \frac{R_s + R_t}{r_{st}} \right)^6 \right]$$

where  $A$  is a scaling factor,  $\epsilon_{st}$  is the depth of the Lennard–Jones potential well, and  $R$  is the van der Waals radius of the considered atoms. The probe is usually taken as a single atom and its parameters are set to unity.

**Hydrophathic atom constants**  $h$  are derived from Leo–Hansch hydrophobic fragmental constants in such a way that

- the sum of hydrophathic atom constants in a group is consistent with the group fragmental constant;
- frontier atoms in a group are more important than shielded atoms;
- bond, chain, branch, and proximity factors are applied in an additive scheme, the former three to all eligible atoms, the last to the central atoms of polar groups.

Positive hydrophathic constants indicate hydrophobic atoms, whereas negative constants indicate hydrophilic atoms.

The **Molecular Lipophilicity Potential** (MLP) describes the combined lipophilic effect of all fragments in a molecule on its environment and can be calculated at any point in space around the molecule [Audry, Dubost *et al.*, 1986, 1992; Fauchère, Quarendon *et al.*, 1988; Furet, Sele *et al.*, 1988; Audry, Dallet *et al.*, 1989]. It is defined by considering a molecule surrounded by nonpolar or low polarity organic solvent molecules, and assuming that the solvent molecule distribution around the considered molecule depends on the fragmental or atomic contributions to  $\log P$  and the distances at which the solvent molecules are from the target molecule. Therefore, the molecular lipophilicity potential at each  $k$ th grid point is calculated as

$$\text{MLP}_k = \sum_{i=1}^A \frac{a_i}{1 + r_{ki}}$$

where the sum runs over all atoms (or fragments) of the target molecule;  $a_i$  are the  $\rightarrow$  *Ghose–Crippen hydrophobic atomic constants* for the  $i$ th atom (or fragments) in the target molecule, and  $r_{ki}$  is the distance between the considered atom (or fragments) and the  $k$ th grid point. Only non-hydrogen atoms  $A$  of the molecule are usually considered.

In contrast to other potentials, the lipophilicity potential is not obtained by calculating the interactions between a probe and the molecule.

Different MLP functions can be obtained according to the selection of the fragmental constant values and the distance function [Croizet, Langlois *et al.*, 1990; Gaillard, Carrupt *et al.*, 1994a, 1994b; Testa, Carrupt *et al.*, 1996; Carrupt, Testa *et al.*, 1997]. The MLP has been later adapted to a new atomic hydrophobic parameter called **Topological Lipophilicity Potential** (TLP) defined for each  $j$ th atom of the molecule as [Langlois, Audry *et al.*, 1993; Dubost, 1993]

$$\text{TLP}_j = \sum_{i=1}^A \frac{a_i}{1 + d_{ij}}$$

where  $d_{ij}$  is the  $\rightarrow$  *topological distance* between atoms  $i$  and  $j$  of the molecule.

📖 [Gussio, Pattabiraman *et al.*, 1996; Masuda, Nakamura *et al.*, 1996; Testa, Raynaud *et al.*, 1999]

#### • hydrogen bonding fields

These are descriptors accounting for hydrogen-bonding interactions between ligand and receptor. Hydrogen bonding fields are obtained by calculating the energy  $E_{hb}$  due to the

formation of hydrogen-bonds between probe and target in each grid point [Leach, 1996; Oprea and Waller, 1997].

The hydrogen bonding potential energy is calculated as [Wade, 1993]

$$E_{hb} = \sum_s \sum_t E_r(r_{st}) \cdot E_s \cdot E_t$$

where the first sum runs over probe atoms and the second over atoms of the target molecule;  $E_r$  is an energy component dependent on the interatomic distance  $r_{st}$  between probe and target atoms involved in the hydrogen-bond;  $E_s$  and  $E_t$  are energy components dependent on the angle made by the hydrogen bond at the probe and target atoms, respectively.  $E_s$  and  $E_t$  values are between 0 and 1. The component  $E_r$  is usually defined by a Lennard-Jones function as

$$E_r = \frac{A}{r_{st}^m} - \frac{C}{r_{st}^n}$$

where  $A$  and  $C$  are constants dependent on the chemical type of the hydrogen-bonding atoms;  $m$  and  $n$  are parameters taking different values; for example,  $m = 12$  and  $n = 10$  are commonly used values,  $m = 8$  and  $n = 6$  were used in the GRID hydrogen bonding energy function [Boobbyer, Goodford *et al.*, 1989].

A more sophisticated hydrogen bonding potential energy based on the geometry of the hydrogen-bonding systems was proposed by Kim [Kim, 1993a, 1993f; Kim, Greco *et al.*, 1993] and implemented in the GRID program:

$$E_{hb} = \left( \frac{C}{r_{st}^6} - \frac{D}{r_{st}^4} \right) \cdot \cos(m \cdot \theta)$$

where the energy is evaluated in each grid point;  $r_{st}$  is the interatomic distance between probe and target atoms involved in the hydrogen-bond;  $C$  and  $D$  are parameters taken from tables;  $m$  is usually equal to one; and  $\theta$  is the angle made by donor, hydrogen, and acceptor atoms. The probe used is a neutral  $H_2O$  molecule with an effective radius of 1.7 Å, free to rotate around the grid point.

#### • total interaction energy fields

These are potential energy descriptors accounting for the total noncovalent interaction potential energy, which determines the binding affinity of a molecule to the considered receptor. They are generally calculated as the pairwise sum of the interaction energies between each probe atom and each target atom as [Wade, 1993]

$$E = \sum_s \sum_t (E_{vdw} + E_{el} + E_{hb})_{st}$$

where the first sum runs over probe atoms and the second over atoms of the target molecule;  $E_{vdw}$  is the van der Waals interaction energy,  $E_{el}$  the electrostatic energy, and  $E_{hb}$  the hydrogen-bonding energy. Other noncovalent energy contributions can be included.

#### • desolvation energy fields

These are potential energy descriptors proposed as an indicator of hydrophobicity [Oprea and Waller, 1997]. Originally, they were calculated using the finite difference approximation method; the linearized Poisson-Boltzmann equation was solved numerically to compute the electrostatic

contribution to solvation at each grid point. Desolvation energy field values were calculated as the difference between solvated (grid dielectric = 80) and *in vacuum* (grid dielectric = 1).

📖 [Richard, 1991; Balogh and Naray-Szabo, 1993; Kim, 1993b; Naray-Szabo and Balogh, 1993; Nusser, Balogh *et al.*, 1993; van de Waterbeemd, Camenisch *et al.*, 1996; Liljefors, 1998] [Cruciani, Pastor *et al.*, 2001a; Cruciani, Benedetti *et al.*, 2004; Cianchetta, Li *et al.*, 2006; Goodford, 2006; Wade, 2006]

- **molecular lipophilicity potential** → molecular interaction fields (⊙ hydrophobic fields)
- **molecular lipophilicity potential model** → lipophilicity descriptors
- **molecular matrix** → molecular geometry
- **Molecular Modeling** ≡ *Computer-Aided Molecular Modeling* → drug design
- **molecular moment of energy** → self-returning walk counts
- **molecular negentropy** → information content
- **molecular orbital contour surface** → molecular surface
- **molecular orbital energies** → quantum-chemical descriptors
- **molecular orbital fields** → molecular interaction fields
- **molecular path code** → path counts
- **molecular path count** → path counts
- **molecular path number** ≡ *molecular path count* → path counts
- **molecular path/walk indices** → shape descriptors (⊙ path/walk shape indices)
- **molecular polarizability** → electric polarization descriptors
- **Molecular Polarizability Effect Index** → electric polarization descriptors (⊙ Polarizability Effect Index)

### ■ molecular profiles

These are molecular descriptors denoted by  ${}^kD$  and derived from the → *distance distribution moments* of the → *geometry matrix* **G**, defined as the average row sum of its entries raised at the *k*th power, normalized by the factor *k*!:

$${}^kD = \frac{1}{k!} \cdot \frac{\sum_{i=1}^A \sum_{j=1}^A r_{ij}^k}{A}$$

where  $r_{ij}^k$  is the *k*th power of the *i*-*j* entry of the geometry matrix and *A* the number of atoms (Figure M8) [Randić, 1995a, 1995b; Randić and Razinger, 1995b].

Using several increasing *k* values, a sequence of molecular invariants called *molecular profile* is obtained as

$$\{{}^1D, {}^2D, {}^3D, {}^4D, {}^5D, {}^6D, \dots\}$$

As the exponent *k* increases, the contributions of the most distant pairs of atoms become the most important.

The maximum nonzero value of  ${}^kD$  is for the power corresponding to the number of atoms of the molecule (*k* = *A*); to obtain → *uniform-length descriptors*, values for *k* > *A* are set at zero.

For large *k* values,  ${}^kD$  tends to zero, due to the effect of the factorial normalization factor.



Another set of theoretical invariants can be obtained by averaging the row sums as

$${}^k d = \frac{1}{k!} \cdot \frac{\sum_{i=1}^A \sum_{j=1}^A r_{ij}^k}{A^2}$$

obtaining the vector

$$\{ {}^1 d, {}^2 d, {}^3 d, {}^4 d, {}^5 d, {}^6 d, \dots \}$$

For characterization of 2D structures, molecular profiles are computed in the same way by the  $\rightarrow$  *distance distribution moments* of the topological  $\rightarrow$  *distance matrix D*.

If one is interested in the characterization of molecular local features, that is, **local profiles**, the calculation of the  ${}^k D$  values can be restricted to the local environment of interest, that is, only the row sums corresponding to atoms of interest are considered, obtaining a vector of local theoretical invariants:

$$\{ {}^1 L, {}^2 L, {}^3 L, {}^4 L, {}^5 L, {}^6 L, \dots \}$$

By this way, different types of profiles can be derived, such as **shape profiles**, which are local profiles taking into account only atoms on molecular periphery:

$$\{ {}^1 S, {}^2 S, {}^3 S, {}^4 S, {}^5 S, {}^6 S, \dots \}$$

In this case, the row sums of the geometry matrix are obtained by summing only the geometric distance powers of the atoms belonging to the periphery and the average is made by the number of the contributing atoms only. Each atomic distance sum is considered as a local indicator of molecular shape and each molecular invariant  ${}^k S$  is considered a global shape descriptor.

In the case of 3D space-filled molecular models, one can represent the molecule by **contour profiles**, which are shape profiles calculated for all individual contours used to map the molecule. Each contour profile is then defined by a sequence:

$$\{ {}^1 C, {}^2 C, {}^3 C, {}^4 C, {}^5 C, {}^6 C, \dots \}$$

where each element of the profile is the normalized average row sum of an augmented geometry matrix, where additional points defining the contour are also considered.

Particular contour profiles are obtained by randomly distributed points over the surface of the molecule.

Moreover, an arbitrary number of points can be considered along the molecule bonds, thus deriving **bond profiles**:

$$\{ {}^1 B, {}^2 B, {}^3 B, {}^4 B, {}^5 B, {}^6 B, \dots \}$$

Bond profiles constitute a generalization of atomic molecular profiles since they provide a characterization of molecular connectivity, which is not explicitly contained in the geometry matrix [Randić, 1996a; Randić and Krilov, 1996].

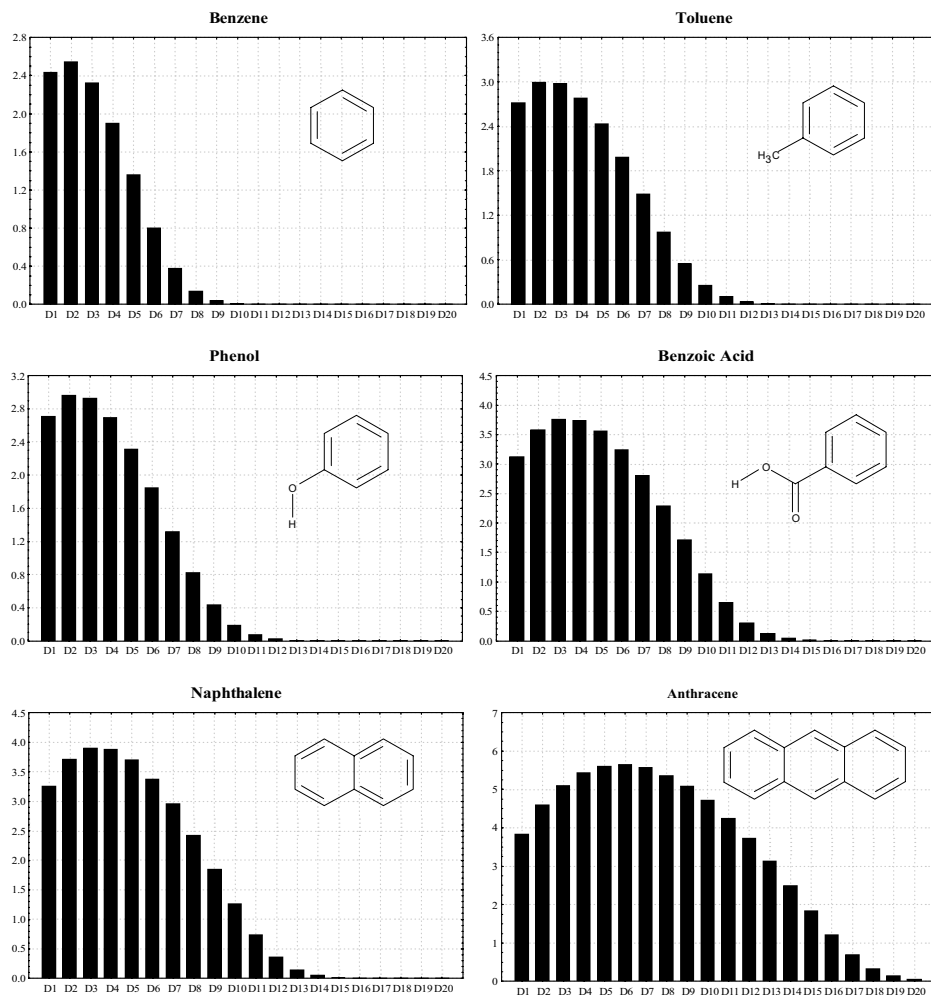



Figure M8 Molecular profiles for some compounds.

**Volume profiles**  $^kV$  of a molecule can be calculated by distributing random points throughout the molecular interior defined by  $\rightarrow$  *van der Waals molecular surface* and then constructing the corresponding augmented geometry matrices whose elements are raised at the  $k$ th power [Randić and Krilov, 1997b]. Moreover, to characterize the molecular surface, random points are restricted to the surface, thus obtaining **surface profiles**  $^kSA$ . Based on the same principles of the  $\rightarrow$  *surface-volume ratio*  $G' = SA/V$ , the **volume-to-surface profiles**  $^kV/^kSA$  have been proposed as  $\rightarrow$  *shape descriptors* defined as the ordered sequence of the ratios of the volume over surface profile elements of corresponding order.

 [Randić, 1996c; Randić and Krilov, 1997a; Randić and Razinger, 1997; Zefirov and Tratch, 1997]

- **molecular pseudograph's adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Molecular Quantum Self-Similarity Measures** → quantum similarity
- **molecular quantum similarity** ≡ *quantum similarity*
- **Molecular Quantum Similarity Indices** → quantum similarity
- **Molecular Quantum Similarity Measures** → quantum similarity
- **molecular representation** → molecular descriptors
- **molecular rigidity** → flexibility indices
- **molecular self-returning walk count** → self-returning walk counts
- **molecular sequence code** → sequence matrices
- **molecular sequence count** → sequence matrices

### ■ Molecular Shape Analysis (MSA)

A QSAR approach based on a set of methods, which combines molecular shape similarity and commonality measures with other → *molecular descriptors* to search both for similarities among molecules and build QSAR models [Hopfinger, 1980; Burke and Hopfinger, 1993]. The term *molecular shape similarity* refers to molecular similarity on the basis of a comparison of three-dimensional molecular shapes represented by some property of the atoms composing the molecule, such as the van der Waals spheres. The *molecular shape commonality* is the measure of molecular similarity when conformational energy and molecular shape are simultaneously considered [Hopfinger and Burke, 1990].

The main assumption of this approach is that the shape of the molecules is closely related to the shape of the → *binding site cavity* and, as a consequence, to the biological activity. Therefore, a shape reference compound is chosen, which represents the binding site cavity, and the similarity (or commonality) measured between the reference shape and the shape of other compounds is used to determine the biological activity of these compounds. Besides the shape similarity measures, other molecular descriptors such as those in → *Hansch analysis* can be used to evaluate the biological response. The MSA model is thus defined as

$$\hat{y}_i = b_0 + \sum_k b_k \cdot \Phi_{ik} + [f_0(M(i,j)) + \rho(n_j - n_i) - \beta \cdot \Delta E_i]$$

where  $i$  refers to any compound of the data set and  $j$  to the reference compound;  $\hat{y}_i$  is the estimated biological response of the  $i$ th compound, usually expressed as a logarithm of the ligand inverse concentration;  $b_0$  is a constant characteristic of the reference compound ( $j$ );  $\Phi_k$  is any molecular descriptor representing → *physico-chemical properties* such as → *Hansch descriptors*, topological, geometrical, electronic, or thermodynamic features of the molecules. The last term (in squared brackets) is a 3D molecular structure term involving molecular shape and conformational thermodynamics;  $f_0(M(i,j))$  is a molecular shape similarity function, that is, a function of the measure of the relative shape similarity between  $i$  and  $j$ ,  $\rho \cdot (n_j - n_i)$  is the difference in intramolecular conformational entropy (flexibility) between  $i$  and  $j$ , and  $\beta \cdot \Delta E_i$  is a measure of the relative stability of the bioactive conformation of compound  $i$  with respect to its global intramolecular energy minimum. The quantity  $I_c(i,j) = f_0(M(i,j)) - \beta \cdot \Delta E_i$  is the shape commonality index, which takes into account the balance between a gain in molecular shape similarity at the expense of loss in conformational stability.

There are seven operations involved in the MSA approach:

- (a) conformational analysis;
- (b) active conformation hypothesis;
- (c) shape reference compound selection;
- (d) pairwise molecular superimposition;
- (e) molecular shape similarity (or commonality) measure calculation;
- (f) other molecular descriptor calculation; and
- (g) trial QSAR model development.

For each MSA operation, there exists a set of choices that are experimented in the trial QSAR model; the final selection of the requirements for each MSA operation is based on optimizing the fitting ability of the QSAR model.

The most active compound is usually assumed as the reference structure, but also a set of overlapped structures can be assumed to define a reference shape.

Some **molecular shape similarity descriptors** (or **MSA descriptors**) are mentioned below. They represent a measure of the matching between the shapes of two molecules  $i$  and  $j$ , one of them being by definition the reference structure; the representation of molecular shape is given in different ways.

• **Common Overlap Steric Volume (COSV)**

Defined as the volume shared by two superimposed molecules, that is,

$$M_0(i,j) \equiv V_0(i,j) = V_i \cap V_j$$

where  $V_i$  and  $V_j$  are the  $\rightarrow$  *van der Waals volume* of the  $i$ th and  $j$ th molecules, respectively.

Two arbitrary functions of the common overlap steric volume were also introduced as alternative molecular shape descriptors:

$$S_0 = V_0^{2/3} \quad \text{and} \quad L_0 = V_0^{1/3}$$

where  $S_0$  is the **common overlap surface** (or **overlap surface**) and  $L_0$  the **common overlap length**. Despite the terms,  $S_0$  has the dimensions of area but is not a physical measure of the common surface area between two molecules, and the same holds for  $L_0$ . Therefore, if the shape of the reference molecule is a good approximation for the acceptor site cavity,  $V_0$  should measure the part of the cavity volume occupied by the considered ligand, whereas  $S_0$  should be an approximation for the contact surface area of the ligand with receptor.

The **nonoverlap steric volume**  $V_{\text{non}}$  is another MSA descriptor defined as [Tokarski and Hopfinger, 1994]

$$V_{\text{non}}(i,j) = V_{ij} - V_j$$

where  $V_{ij}$  is the composite steric volume of the two aligned molecules  $i$  and  $j$ . In practice, the nonoverlap volume measures the regions of the  $i$ th molecule volume not shared by the reference compound, that is, it represents the  $\rightarrow$  *steric misfit*.

- **atom-pair matching function**

Defined as

$$M_r(i, j) = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} K_{aa'} \cdot r_{aa'}$$

where the sums run over all pairs of atoms of the two considered molecules  $i$  and  $j$ ,  $r_{aa'}$  is the interatomic distance between each pair of atoms from the  $i$ th and  $j$ th molecules, and  $K_{aa'}$  is a user-defined constant providing the relative importance of the considered distance. For  $M_r \rightarrow 0$ , the superposition between  $i$  and  $j$  becomes better.

- **charge-matching function**

Defined as

$$M_c(i, j) = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} \frac{q_a \cdot q_{a'}}{Q_T} \cdot r_{aa'}$$

where the sums run over all pairs of atoms of the two considered molecules  $i$  and  $j$ ,  $q$  are atomic partial charges,  $r$  the  $\rightarrow$  *interatomic distances* between atoms from molecules  $i$  and  $j$ , and  $Q_T$  is a normalization term calculated as

$$Q_T = \sum_{a=1}^{A_i} \sum_{a'=1}^{A_j} q_a \cdot q_{a'}$$

The partial charges  $q_a$  and  $q_{a'}$  are assumed to always have the same sign, otherwise they would not be matched.

- **Integrated Spatial Difference in Field Potential (ISDFP)**

A field-based shape descriptor derived from the representation of molecular body by  $\rightarrow$  *molecular interaction fields* and defined as

$$M_p(i, j) = \frac{1}{\Phi} \cdot \left[ \int_{\Phi} [E_i(R, \Theta, \phi) - E_j(R, \Theta, \phi)]^2 \cdot d\Phi \right]^{1/2}$$

where  $E$  are the  $\rightarrow$  *interaction energy values*, as measured by a probe, at the spherical coordinate position  $(R, \Theta, \phi)$ , and  $\Phi$  is the considered integration volume. To calculate this descriptor, it is assumed that molecules  $i$  and  $j$  are superimposed.

- **weighted combination of COSV and ISDFP**

A combination of two complementary measures of shape similarity defined as

$$M_w(i, j) = w \cdot [(V_j \cap V_i) - M_0(i, j)] + (1 - w) \cdot M_p(i, j)$$

where  $M_0(i, j)$  and  $M_p(i, j)$  are the common overlap steric volume and the integrated spatial difference in field potential, and  $w$  is a weighting factor between zero and one. The two descriptors are considered complementary in the sense that the overlap volume measures the shape within the van der Waals surface formed by superimposition of  $i$  and  $j$ , whereas *ISDFP* measures the shape outside the van der Waals surface.

📖 [Battershell, Malhotra *et al.*, 1981; Hopfinger, 1981, 1983, 1984; Hopfinger and Potenzzone Jr, 1982; Mabilia, Pearlstein *et al.*, 1985; Walters and Hopfinger, 1986; Hopfinger, Compadre *et al.*, 1987; Rohrbaugh, Jurs *et al.*, 1988; Nagy, Tokarski *et al.*, 1994; Rowberg, Even *et al.*, 1994; Rhyu, Patel *et al.*, 1995; Holzgrabe and Hopfinger, 1996]

### ■ Molecular Shape Field (MSF)

The molecular shape field (MSF) is constituted by values of the  $\rightarrow$  *molecular interaction potential* (MEP) of selected grid points that compose the molecular surface [Urbano-Cuadrado, Carbó *et al.*, 2007].

To obtain data suitable for later analysis, the local curvature values at each of the grid points of the molecular surface are computed using a cosine expression similar to that used by Pastor *et al.* [Pastor, Cruciani *et al.*, 2000]. The MSF values range between 0 and  $-1$  for convex areas and 0 and  $+1$  for concave ones.

From MSF and MEP values,  $\rightarrow$  *autocorrelation descriptors* MSF–MSF, MEP–MEP, and MSF–MEP were proposed as molecular descriptors.

- **molecular shape similarity descriptors**  $\rightarrow$  molecular shape analysis
- **molecular similarity matrices**  $\rightarrow$  similarity/diversity

### ■ molecular structure

“... the term *molecular structure* represents a set of non-equivalent conceptual entities. There is no reason to believe that when we discuss different topics (e.g., organic synthesis, reaction rates theories, spectroscopic transitions, reaction mechanisms, *ab initio* calculations) using the concept of molecular structure, the different meaning we attach to the term *molecular structure* ultimately flows from the same concept” [Basak and Gute, 1997].

Together with the concepts of synthesis and chemical composition, the concept of molecular structure is one of the most fruitful of twentieth century scientific researches. This concept is conveniently studied by considering several levels of description, that is, the molecular structure is a part of a hierarchical system organized in different levels; to each level correspond characteristic language, properties, and relationships within its constitutional elements at that level as well as relationships between higher and lower levels. Thus, particles, atoms, molecules, compounds, cells, bodies, and so on are hierarchically organized levels of a complex system. At each level emergent properties arise from the organization of elements characterizing that level, that is, the presence of organizing relationships gives birth to new properties and constraints that influence the complexity of the system.

The above considerations also hold for different hierarchical descriptions of a system at a given level, that is, the same level is traversed by an inner hierarchical organization due to different descriptions of the same elements. The molecular representations are hierarchical descriptions of the molecular system; therefore, derived from the different representations of the molecular structure, several  $\rightarrow$  *molecular descriptors* are calculated with different chemical information content.

Each molecular representation reflects hypotheses, ideas, a theory on unknown but supposed relevant relationships between molecules and their behavior. Much chemical research makes efforts to accurately predict properties, or to accurately classify chemical structures according to their properties, on the basis of chemical structure alone.

Each  $\rightarrow$  *molecular representation* is a model that highlights only a part of the chemical reality, and, then, explaining only a part of the experimental evidence. Also a simple chemical formula

such as  $C_6H_5Cl$  already gives chemical information, at least about chemical composition and stoichiometric atom-type relationships.

Although chemical theories are the framework within which molecular structure has been developed, experimental properties define the reference framework in which the concept, or, better still, the concepts of molecular structure have been continuously verified, evaluated, and modified.

📖 [Woolley, 1978a, 1978b; Primas, 1981; Weininger, 1984; Turro, 1986; Wirth, 1986; Rouvray, 1989b; Weininger and Weininger, 1990; Ash, Warr *et al.*, 1991; Randić, 1992a; Wentang, Ying *et al.*, 1993; Dietz, 1995; Bauerschmidt and Gasteiger, 1997; Testa, Kier *et al.*, 1997; Ivanciuc, 2001a; Xu, 2003; Kuz'min, Artemenko *et al.*, 2005; Clark, Labute *et al.*, 2006]

- **molecular subgraph** → molecular graph
- **molecular supergraph** → hyperstructure-based QSAR techniques

### ■ molecular surface

The term molecular surface is usually referred to any surface surrounding some or all of the nuclei of the molecule. In the strict quantum mechanical sense, molecules do not have precisely defined surfaces; however, in analogy to macroscopic objects, the electron distribution may be regarded to as a 3D *molecular body* whose boundary is the molecular surface. In other words, the molecular surface can be viewed as the formal boundary that separates the 3D space into two parts: within the surface one is expected to find the whole molecule and beyond the rest of the universe [Meyer, 1986b, 1991c].

Different physical properties and molecular models have been used to define the molecular surface, the most common are reported below together with the descriptors proposed as measures of **surface areas** and molecular volume (→ *volume descriptors*). Molecular surface area and volume are parameters of molecules that are very important in understanding their structure and chemical behavior such as their ability to bind ligands and other molecules. An analysis of molecular surface shape is also an important tool in QSAR and → *drug design*, in particular both → *molecular shape analysis* and → *Mezey 3D shape analysis* were developed to search for similarities among molecules, based on their molecular shape.

#### • van der Waals molecular surface

The surface that envelops fused hard spheres centered at the atom coordinates (atomic nuclei) and having radii equal to some of the recommended values of the van der Waals radii. The spheres interpenetrate one another in such a way that the distance between the centers of two spheres equals the formal bond length.

In the hard-sphere model [Ciubotariu, Medeleanu *et al.*, 2004], the **van der Waals molecular surface**  $SA^{vdw}$  (also known as **Total molecular Surface Area, TSA**) is then defined as the exterior surface of the union of all such spheres in the molecule, that is, the area of the van der Waals molecular surface. It can be calculated by generating a uniform grid around each sphere of the molecule atoms, followed by the counting of the number of points generated on the surface  $n_s$ , consisting in the points that satisfy at least one of the following equalities:

$$(X_i - x)^2 + (Y_i - y)^2 + (Z_i - z)^2 \leq (R_i^{vdw})^2 \quad i = 1, \dots, A$$

where  $A$  is the number of atoms and  $R^{vdw}$  the van der Waals radius;  $X_i$ ,  $Y_i$ , and  $Z_i$  are the coordinates of the  $i$ th atom and  $x$ ,  $y$ , and  $z$  the coordinates of the generated points.

Then, the number of points  $n_e$  that are external to the surface have to be counted, that is, the number of points that do not satisfy the inequalities.

Therefore, the van der Waals surface  $SA_i^{vdw}$  of each  $i$ th atom is calculated as

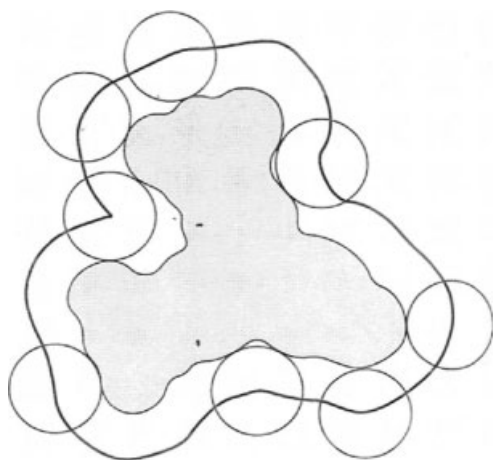
$$SA_i^{vdw} = \frac{(n_e)_i}{n_s} \cdot 4 \cdot \pi \cdot (R_i^{vdw})^2$$

and the total van der Waals surface is calculated as the sum of the atomic van der Waals surfaces:

$$SA^{vdw} = \sum_{i=1}^A SA_i^{vdw}$$

#### • solvent-accessible molecular surface

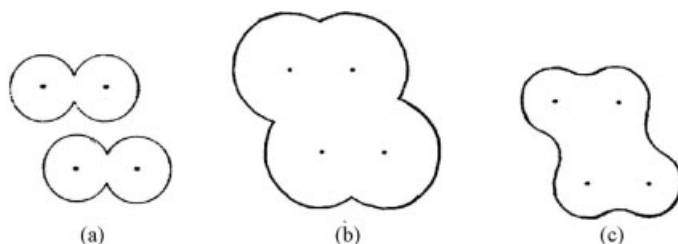
In the case of large complex and folded molecular structures, a part of the van der Waals surface is buried in the interior and is thus inaccessible to solvent interactions, which mainly govern the chemical behavior of molecules in solution. Therefore, to obtain the best representation of the outer surface and overall shape of the molecule the solvent-accessible molecular surface was proposed. It was originally defined [Lee and Richards, 1971] as the surface across which the center of a spherical approximation of the solvent is passed when the solvent sphere is rolled over the van der Waals surface of the molecule (Figure M9). The radius (1.5 Å) of the solvent sphere is usually chosen to approximate the contact surface formed when a water molecule interacts with the considered molecule. If there are several grooves or minor cavities on the van der Waals surface where the rolling sphere cannot enter, then the solvent-accessible surface will be significantly different from the van der Waals surface (Figure M10).



**Figure M9** Solvent-accessible molecular surface defined by the centers of spheres rolled along the molecular contour surface. The radius of the sphere is chosen according to the size of the solvent molecule.



A few years later, Richards [Richards, 1977] gave a new definition of solvent-accessible surface, dividing it into two parts: the *contact surface* and the *reentrant surface*. The **contact surface** is that part of the van der Waals surface that is accessible to the probe sphere representing the solvent molecule. The **reentrant surface** comes from the inward-facing surface of the probe sphere when it is simultaneously in contact with more than one atom.



**Figure M10** Comparison among (a) van der Waals surface, (b) solvent-accessible surface, and (c) contact surface.

The area of the solvent-accessible surface is called **Solvent-Accessible Surface Area**, **SASA** (or **Total Solvent-Accessible Surface Area**, **TSASA**). Several algorithms were proposed that implement both the first original definition of **SASA** and that of Richards. One of the most popular algorithms that implements Richards' solvent-accessible surface was proposed by Connolly [Connolly, 1983]. It is an analytical method for computing molecular surface, and is based on surface decomposition into a set of curved regions of spheres and tori that join at circular arcs; spheres, tori, and arcs are defined by analytical expressions in terms of atomic coordinates, van der Waals radii, and the probe radius. The molecular surface calculated in such a way is sometimes referred to as **Connolly surface area**. This algorithm also allows the calculation of solvent-accessible atomic areas.

An alternative to the hard sphere model is a recently proposed method for **SASA** calculations, based on atomic Gaussian functions describing the exposure of atoms and molecular fragments to solvent. A simple integral function of these atomic Gaussians is used to define a Gaussian neighborhood, which behaves in a complementary fashion to the conventional definition of solvent accessibility, that is, the smaller the Gaussian neighborhood, the more exposed the atom and hence the larger its accessibility [Grant and Pickup, 1995; Grant, Gallardo *et al.*, 1996].

Several  $\rightarrow$  **charged partial surface area descriptors (CPSA)** and  $\rightarrow$  **hydrogen-bond charged partial surface area descriptors (HB-CPSA)** are based on portions of the solvent-accessible surface area relative to polar or hydrophobic regions of the molecule, in some cases weighted by the corresponding local charges. Moreover, the **Hydrated Surface Area (HSA)** is the portion of the solvent-accessible surface area associated with hydration of polar functional groups.

The **Isotropic Surface Area (ISA)** is the surface of the molecule accessible to nonspecific interactions with the solvent, that is, the surface of the molecule involved in specific hydrogen-bonding with water is not considered [Collantes and Dunn III, 1995; Koehler, Grigoros *et al.*, 1988]. A hydration complex model needs to estimate the isotropic surface area. The **Polar Surface Area (PSA)** is defined as the part of the surface area of the molecule associated with

oxygen, nitrogen, sulfur, and the hydrogen bonded to any of these atoms. This surface descriptor is related to the hydrogen-bonding ability of compounds [Palm, Luthman *et al.*, 1998; Winiwarter, Bonham *et al.*, 1998] and is used to define some  $\rightarrow$  *drug-like indices*.

The volume of space bounded by the solvent-accessible molecular surface is called **solvent-excluded volume** because it is the volume of space from which solvent is excluded by the presence of the molecule when the solvent molecule is also modeled as a hard sphere. Moreover, the **interstitial volume** is the volume consisting of packing defects between the atoms that are too small to admit a probe sphere of a given radius; in practice, it is calculated as the difference between the solvent-excluded volume and the van der Waals volume. An analytical method developed by Connolly was able to calculate the solvent-excluded volume [Connolly, 1983a]; several other numerical and analytical approaches have been proposed.

📖 [Silla, Tunon *et al.*, 1991; Hirono, Qian *et al.*, 1991]

- **electron isodensity contour surface**

The collection of all those points  $\mathbf{r}$  of the space where the value of the  $\rightarrow$  *electron density*  $\rho(\mathbf{r})$  is equal to a threshold value  $m$  [Mezey, 1991b], that is,

$$G(m) = \{\mathbf{r} : \rho(\mathbf{r}) = m\}$$

Any positive value as threshold  $m$  can be chosen, even if a relatively small value is usually used to define a suitable molecular surface because the electron density converges rapidly to zero at short distances from the nuclei. The positive values of the threshold are due to the usual convention that a large negative charge means a large positive value of the electron density. For large values of  $m$ , the molecular surface is composed of several disconnected surfaces each surrounding one nucleus, whereas for too small values of  $m$ , the surface is an essentially spherical surface surrounding all of the nuclei and containing no information on the shape of the molecule.

- **molecular electrostatic potential contour surface**

The collection of all those points  $\mathbf{r}$  of the space for which the value of the  $\rightarrow$  *molecular electrostatic potential* (MEP)  $V(\mathbf{r})$  is equal to a threshold value  $m$  [Mezey, 1991b], that is,

$$G(m) = \{\mathbf{r} : V(\mathbf{r}) = m\}$$

The contour parameter  $m$  as well as the electrostatic potential can take both positive and negative values. An analysis of the shape of MEP surfaces is of particular interest in  $\rightarrow$  *drug design* as the electrostatic potential has a marked influence on the binding interactions between ligand and receptor. Moreover, the sum of all the surface minima values of the electrostatic potential, denoted as  $\sum V_S^-$ , was proposed as a molecular descriptor able to account for lipophilicity [Zou, Zhao *et al.*, 2002].

- **molecular orbital contour surface**

A molecular surface defined as the contour surface of individual molecular orbitals such as HOMO and LUMO, other frontier orbitals, or localized and delocalized orbitals [Mezey, 1991b]. In practice, it is the collection of all those points  $\mathbf{r}$  of the space for which the value of the electronic wavefunction  $\Psi(\mathbf{r})$  of the considered molecular orbital is equal to a threshold

value  $m$ , that is,

$$G(m) = \{\mathbf{r} : \psi(\mathbf{r}) = m\}$$

The contour parameter  $m$  can take both positive and negative values.

📖 [Hermann, 1972; Amidon, Yalkowsky *et al.*, 1975; Arteca, Jammal *et al.*, 1988b; Leicester, Finney *et al.*, 1988; Marsili, 1988; Lipkowitz, Baker *et al.*, 1989; Pascual-Ahuir and Silla, 1990; Valkó and Slegel, 1992; Brusseau, 1993; Leicester, Finney *et al.*, 1994a, 1994b; Schüürmann, 1995; Lee, Kwon *et al.*, 1996; Palm, Luthman *et al.*, 1996; Brickmann, 1997; Hermann, 1997; Randić and Krilov, 1997b; Zweerszeilmaker, Horbach *et al.*, 1997; Whitley, 1998; Jørgensen, Jensen *et al.*, 2001; Deanda and Pearlman, 2002; King, 2002]

### ■ molecular surface interaction terms (MSI)

These constitute a set of molecular descriptors including the molecular surface area and empirically derived descriptors accounting for dispersion, polar, and hydrogen-bonding interactions [Grigoras, 1990]. They were proposed to empirically express the molecular surface energy using atomic contributions to the total molecular surface. The molecular surface interaction terms are

$$A = \sum_i SA_i \quad A_- = \sum_{a-} SA_a \cdot b_a \cdot q_a^- \quad A_+ = \sum_{a+} SA_a \cdot b_a \cdot q_a^+ \quad A_{HB} = \sum_i SA_i \cdot b_i \cdot q_i^H$$

where  $A$  is the  $\rightarrow$  *total molecular surface area* calculated as the sum of all the atomic surface areas  $SA_i$ ; this is a dispersion molecular surface interaction term.  $A_-$  is the electrostatic negative molecular surface interaction term calculated as the sum of surface areas of negatively charged atoms multiplied by their corresponding scaled  $\rightarrow$  *net atomic charge*  $q_a^-$ .  $A_+$  is the electrostatic positive molecular surface interaction term calculated as the sum of surface areas of positively charged atoms multiplied by their corresponding scaled net atomic charge  $q_a^+$ .  $A_{HB}$  is the hydrogen-bonding molecular surface interaction term calculated as the sum of the surface areas of hydrogen-bonding hydrogen atoms multiplied by their corresponding scaled net atomic charge  $q_i^H$ . The coefficient  $b_i$  is an empirical charge scaling factor, which is the same for the atoms of the same chemical type (Table M12).

**Table M12** Charge scaling factors  $b$  for different atom chemical types.

Atom/hybrid	Coefficient $b$	Atom/hybrid	Coefficient $b$
H (at $C_{sp^3}$ )	3.29	$N_{sp^3}$	0.155
H (at $C_{sp^2}$ )	7.77	$N_{sp}$	1.59
$H_{HB}$	1.00	$N_{AROM}$	2.79
$C_{sp^3}$	1.00	$O_{sp^3}$	1.32
$C_{sp^2}$	0.00	$O_{sp^2}$	1.51
$C_{sp}$	2.33	F	0.00
$C_{AROM}$	9.25	Cl	1.78

The hydrogen atom considered for the calculation of the hydrogen bonding term are not taken into account in the  $A_+$  term.

- **molecular topological index**  $\equiv$  *Schultz molecular topological index*
- **molecular topological indices**  $\equiv$  *topological indices*  $\rightarrow$  graph invariants

### ■ molecular transforms

These are descriptors based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves [Soltzberg and Wilkins, 1976, 1977]. A generalized scattering function can be used as the functional basis for deriving, from a known molecular structure, the specific analytic relationship of both X-ray and electron diffraction. The general molecular transform is

$$G(\mathbf{s}) = \sum_{i=1}^A f_i \cdot \exp(2\pi i \cdot \mathbf{r}_i \cdot \mathbf{s})$$

where  $\mathbf{s}$  represents the scattering in various directions by a collection of  $A$  atoms located at points  $\mathbf{r}_i$ ;  $f_i$  is a form factor taking into account the direction dependence of scattering from a spherical body of finite size. The scattering parameter  $s$  has the dimension of a reciprocal distance and depends on the scattering angle as

$$s = \frac{4\pi}{\lambda} \cdot \sin(\vartheta/2)$$

where  $\vartheta$  is the scattering angle and  $\lambda$  the wavelength of the electron beam.

Usually, the above equation is used in a modified form as suggested in 1931 by Wierl [Wierl, 1931]. On substituting the form factors by an  $\rightarrow$  *atomic property*  $w_i$ , considering the molecule to be rigid and setting the instrumental constant equal to one, the following function, usually called **radial distribution function**, is used to calculate molecular transforms:

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}}$$

where  $I(s)$  is the scattered electron intensity,  $w$  an atomic property, chosen as the atomic number  $Z$  by Soltzberg and Wilkins [Soltzberg and Wilkins, 1976],  $r_{ij}$  the  $\rightarrow$  *geometric distance* between the  $i$ th and  $j$ th atom, and  $A$  the number of atoms in the molecule. The sum is performed over all the pairs of atoms in the molecule.

Soltzberg and Wilkins introduced a number of simplifications to obtain a binary code. Only the zero crossing of the  $I(s)$  curve, that is, the  $s$  values at which  $I(s) = 0$ , in the range  $1\text{--}31 \text{ \AA}^{-1}$  were considered. The  $s$  range was then divided into 100 equal-sized bins, each described by a binary variable equal to 1 if the bin contains a zero crossing, 0 otherwise. Thus, a vectorial descriptor consisting of 100 bins was finally calculated for each molecule.

Gabányi *et al.* proposed a modified molecular transform by replacing the geometric distance  $r_{ij}$  with the  $\rightarrow$  *topological distance*  $d_{ij}$  [Gabanyi, Surjan *et al.*, 1982]. Moreover, two different functions of the scattering parameter  $s$  were evaluated to be used in place of the trigonometric term:

$$(a) \quad I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-1/2 \cdot (s \cdot d_{ij})^2} \quad (b) \quad I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot \left(1 - \frac{s \cdot d_{ij}}{\Delta_{ij}}\right)$$

where  $d_{ij}$  is the topological distance between atoms  $v_i$  and  $v_j$ ,  $\Delta_{ij}$  the corresponding  $\rightarrow$  *detour distance* (i.e., the length of the largest path between two atoms), and  $w$  is a chosen atomic property [Csorvassy and Tötsér, 1991].

Raevsky and coworkers applied the molecular transform to study ligand–receptor interactions by using hydrogen-bond abilities, hydrophobicity, and charge of the atoms, instead of the atomic numbers  $Z$  [Novikov and Raevsky, 1982]. For each atomic property, a spectrum of interatomic distances, **interatomic interaction spectrum**, was derived to represent the 3D structure of molecules and the scattered intensities in selected regions of the spectrum were used as the molecular descriptors [Raevsky, Dolmatova *et al.*, 1995]. These  $\rightarrow$  *vectorial descriptors* are based on local characteristics of different pairs of centers in the molecule. For a selected distance  $R$ , the following function is evaluated [Raevsky, Dolmatova *et al.*, 1995; Raevsky, 1997a, 1977b; Raevsky, Trepalin *et al.*, 2000]:

$$I(R) = \sum_{r_{\min}}^{r_{\max}} \sum_{i=1}^A \sum_{j=1}^A \frac{w_i \cdot w_j}{1 + \sqrt{\frac{(R-r_{ij})^2}{0.1}}} \quad i \neq j$$

where  $A$  is the number of atoms in the molecule,  $w_i$  and  $w_j$  are  $\rightarrow$  *atomic properties* of the  $i$ th and  $j$ th atom, respectively,  $r_{ij}$  is the geometric interatomic distance;  $r_{\min}$  and  $r_{\max}$  define a distance range around  $R$ , which accounts for vibrations of atoms and allows to obtain a band instead of a line in the final spectrum for each pair of centers defined by  $R$ . Distances  $R$  are varied from 1.1 to 20 Å with step 0.1 Å, resulting in a total of 190 signals per spectrum.

Superimposition of all the bands for all the possible pairs of centers forms the final interatomic interaction spectrum. Seven types of spectrum are calculated for each molecule by using different atomic properties  $w$ . These include *atomic van der Waals radius* (steric interaction spectrum), *atomic charges* (spectrum of interactions between positively charged atoms, spectrum of interactions between negatively charged atoms, and spectrum of interactions of positively charged atoms with negatively charged atoms), *hydrogen-bond abilities* (spectrum of interactions between hydrogen-bond donors, spectrum of interactions between hydrogen-bond acceptors, and spectrum of interactions of hydrogen-bond donors with hydrogen-bond acceptors).

The **integrated molecular transform** (FT<sub>m</sub>) is a molecular descriptor calculated from the square of the molecular transform, by integrating the squared molecular transform in a selected interval of the scattering parameter  $s$  to obtain the area under the curve and finally taking the square root of the area [King, Kassel *et al.*, 1990, 1991]. The square root of the integrated molecular transform, called **SQRT index**, was also proposed as molecular descriptor [Famini, Kassel *et al.*, 1991].

Applications of integrated molecular transforms found in literature are: [King and Kassel, 1992; King, 1993, 1994; Molnar and King, 1995, 1998; King and Molnar, 1996, 1997, 2000].

To calculate **3D-MoRSE descriptors** (3D-MOLEcule Representation of Structures based on Electron diffraction, or simply **MoRSE descriptors**), Gasteiger *et al.* [Schuur and Gasteiger, 1996, 1997] returned to the initial  $I(s)$  curve and maintained the explicit form of the curve. As the atomic  $\rightarrow$  *weighting scheme*  $w$ , various  $\rightarrow$  *physico-chemical properties* such as atomic mass, partial atomic charges, and atomic polarizability were considered. To obtain  $\rightarrow$  *uniform-length descriptors*, the intensity distribution  $I(s)$  was made discrete, calculating its value at a sequence of evenly distributed values of, for example, 32 or 64 in the range of 1–31 Å<sup>−1</sup>. Clearly, the more the values are chosen, the finer the resolution in the representation of the molecule.

Applications of 3D-MoRSE descriptors found in literature are: [Gasteiger, Sadowski *et al.*, 1996; Schuur, Selzer *et al.*, 1996a, 1996b; Gasteiger, Schuur *et al.*, 1997; Baumann, 1999; Jelcic, 2004;

Pérez González and Moldes Teran, 2004; Pérez González, Helguera Morales *et al.*, 2004; Caballero and Fernández, 2006; Helguera Morales, Perez *et al.*, 2006; Saiz-Urra, Pérez González *et al.*, 2006, 2007; Yap, Li *et al.*, 2006].

**RDF descriptors** (or **Radial Distribution Function descriptors**) were proposed based on a radial distribution function different from that commonly used to calculate molecular transforms  $I(s)$  [Hemmer, Steinhauer *et al.*, 1999; Selzer, Gasteiger *et al.*, 2000]. The radial distribution function selected here is that quite often used for interpretation of the diffraction patterns obtained in powder X-ray diffraction experiments.

Formally, the radial distribution function of an ensemble of  $A$  atoms can be interpreted as the probability distribution to find an atom in a spherical volume of radius  $R$ . The general form of the radial distribution function is represented by

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-\beta \cdot (R-r_{ij})^2}$$

where  $f$  is a scaling factor,  $w$  characteristic atomic properties of the atoms  $i$  and  $j$ ,  $r_{ij}$  the interatomic distance between the  $i$ th and  $j$ th atom, and  $A$  the number of atoms. The exponential term contains the distance  $r_{ij}$  between the atoms  $i$  and  $j$  and the smoothing parameter  $\beta$ , which defines the probability distribution of the individual interatomic distances;  $\beta$  can be interpreted as a temperature factor that defines the movement of atoms.  $g(R)$  is generally calculated at a number of discrete points with defined intervals. An RDF vector of 128 values was proposed, using a step size for  $R$  about 0.1–0.2 Å, whereas the  $\beta$  parameter is fixed in the range between 100 and 200 Å<sup>-2</sup>. By including characteristic atomic properties  $w$  of the atoms  $i$  and  $j$ , RDF descriptors can be used in different tasks to fit the requirements of the information to be represented. These atomic properties enable the discrimination of the atoms of a molecule for almost any property that can be attributed to an atom.

The radial distribution function in this form meets all the requirements for a 3D structure descriptor: It is independent of the number of atoms, that is, the size of a molecule, it is unique regarding the three-dimensional arrangement of the atoms, and invariant against translation and rotation of the entire molecule. In addition, the RDF descriptors can be restricted to specific atom types or distance ranges to represent specific information in a certain three-dimensional structure space, for example, to describe sterical hindrance or structure/activity properties of a molecule.

Moreover, the RDF vectorial descriptor is interpretable by using simple rules and, thus, it provides a possibility of  $\rightarrow$  *reversible decoding*. Besides information about distribution of interatomic distances in the entire molecule, the RDF vector provides further valuable information; for example, about bond distances, ring types, planar and nonplanar systems, and atom types. This fact is a most valuable consideration for a computer-assisted code elucidation.

To account for stereochemistry of molecules, the  $\rightarrow$  *Chirality Code* was proposed as a modification of the RDF code [Aires-de-Sousa and Gasteiger, 2001].

Applications of RDF descriptors reported in literature are: [Razdol'skii, Trepalin *et al.*, 2000; Yan and Gasteiger, 2003; Caballero and Fernández, 2006; Helguera Morales, Perez *et al.*, 2006; Podlipnik, Solmajer *et al.*, 2006; Saiz-Urra, Pérez González *et al.*, 2006, 2007; Schuffenhauer, Brown *et al.*, 2006; Yap, Li *et al.*, 2006; Hristozov, Da Costa *et al.*, 2007].

- **molecular volume**  $\rightarrow$  volume descriptors (⊙ molar volume)
- **molecular volume index**  $\rightarrow$  volume descriptors

- **molecular walk count** → walk counts
- **molecular weight** → physico-chemical properties
- **molecule center**  $\equiv$  center of a molecule
- **MOLORD algorithm** → iterated line graph sequence

### ■ MOLMAP descriptors

MOLMAP (*MO*Lecular *MA*p of *At*om-level *Pr*operties) descriptors are uniform-length → *vectorial descriptors* derived by mapping physico-chemical properties of all the bonds in a molecule into a 2D Kohonen → *self-organizing map* (SOM) [Zhang and Aires-de-Sousa, 2005; Gupta, Matthew *et al.*, 2006]. These descriptors encode local features of a chemical structure, being calculated on the basis of properties of single elements in a molecule, such as bonds.

A Kohonen map consists of a set of neurons (i.e., vectors of weights) organized into a square grid, each having as many weights as the number of input variables. In the MOLMAP approach, objects used for training the neural network are chemical bonds and the input variables are seven selected properties of bonds: resonance of stabilization, difference between the  $\sigma$  electronegativity of two bonded atoms, difference between the total charge of two bonded atoms, difference between the  $\pi$  charge of two bonded atoms, mean bond polarizability, bond dissociation energy, and bond polarity. These physico-chemical properties were chosen to encode information on the chemical reactivity of compounds, which is related to propensity for bond breaking and bond making.

All the bonds in all the molecules in the data set in analysis are used for network training. As some properties depend on the bond orientation, each bond is taken twice with different orientation. To focus on functional groups, only bonds involving a heteroatom, or an atom of a  $\pi$  system, can be considered. Once the training has been completed, the SOM provides similarities among chemical bonds. Indeed, similar bonds are mapped into the same or closely adjacent neurons.

By using a trained SOM, bonds in a molecule are mapped into the SOM and the pattern of activated neurons is interpreted as a fingerprint of the bonds of the molecule. For numerical processing, each neuron is assigned a value equal to the number of times it was activated by bonds of the molecule. The map, that is, a matrix, is then transformed into a vector by concatenation of columns. This vector is called MOLMAP descriptor. To account for proximity relationship, a value of 0.3 is added to each neuron multiplied by the number of times a neighbor was activated by a bond.

Unlike the common molecular descriptors, MOLMAP descriptors are data set dependent, which means that their values for a molecule change if another training set is used for SOM training or a different map size is chosen. However, their use in QSAR applications can lead to the identification of structural features responsible for the molecular property in analysis.

MOLMAP descriptors were originally proposed for automatic classification of chemical reactions with the name **reaction MOLMAPs** [Zhang and Aires-de-Sousa, 2005; Latino and Aires-de-Sousa, 2006]. These descriptors are calculated as the difference map between the MOLMAPs of the products of a reaction and the MOLMAPs of the reactants of the same reaction. This difference map can be interpreted as the reaction fingerprint. Zero values in the difference map are related to bonds far apart from the reaction center, remaining unchanged during the reaction; negative values concern bonds of the reactants that brake or change properties in the reaction; positive values concern new bonds appearing in the products. If more reactants (products) are involved in the reaction, the MOLMAPs of all reactants (products) are numerically summed.

Self-organizing map for describing chemical information of a molecule is also used in → *Comparative Molecular Surface Analysis* and → *topological feature maps*.

MOLMAP descriptors were used to predict mutagenicity (positive or negative Ames test) by CART classification tree and random forest [Zhang and Aires-de-Sousa, 2007]. Combined with other global molecular descriptors, error percentage of 15 and 16% were achieved for an external data set with 472 compounds and for the training set with 4038 compounds, respectively. They were also applied in modeling the radical scavenging activity of 47 naturally occurring phenolic antioxidants by counterpropagation neural networks obtaining a cross-validated  $Q^2$  of 0.71 [Gupta, Matthew *et al.*, 2006].

- **MOLPRINT-2D fingerprints** → substructure descriptors (⊙ fingerprints)
- **MOLPRINT-3D fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)

### ■ MolSurf descriptors

MolSurf descriptors comprise a set of physico-chemical properties estimated by quantum-chemical calculations [Norinder, Österberg *et al.*, 1997; Sjöberg, 1997; Norinder, Sjöberg *et al.*, 1998].

MolSurf descriptors include  $\log P$ ,  $\log D$ ,  $pK_a$ , polarizability, polarity, number and strength of H-bond acceptor nitrogen and oxygen atoms, number of H-bond donor atoms, and charge-transfer characteristics for all carbon atoms. The definition of specific substituents for which descriptors are calculated is also allowed. Moreover, MolSurf software includes a module allowing the construction of QSAR models by the Partial Least Squares (PLS) regression.

📖 [Norinder, Österberg *et al.*, 1999; Alifrangis, Christensen *et al.*, 2000; Egan, Merz Jr *et al.*, 2000; Stenberg, Norinder *et al.*, 2001; Norinder and Haeblerlein, 2002; Nordqvist, Nilsson *et al.*, 2004]

- **moments about the mean** → statistical indices (⊙ moment statistical functions)

### ■ moments indices

These are molecular descriptors defined in terms of the weighted absolute central moment of first order, which is a statistical quantity used to measure variability of a distribution around a center. They are defined as

$$M(w) = \frac{\sum_{i=1}^A w_i \cdot r_i}{\sum_{i=1}^A w_i}$$

where  $w$  is an → *atomic property* and  $r$  the distance of an atom from the geometric center of the molecule;  $A$  is the number of atoms in the molecule.

The moment index based on the atomic weights  $m_i$  was called **normalized molecular moment**,  $M_n$ , and defined as [King and Molnar, 1997]

$$M_n = \frac{\sum_{i=1}^A m_i \cdot r_i}{MW}$$



where MW is the  $\rightarrow$  *molecular weight* and  $r$  the distance of the atom from the geometric center of the molecule. This descriptor is a measure of absolute deviation of the distribution of the atomic masses and is similar to the  $\rightarrow$  *radius of gyration*, which is defined in terms of the second-order central moments. Moreover, other moment indices encoding information on electronic features of the molecule were derived from quantum-chemical calculations, by replacing the atomic masses with atomic electron densities and charges [Molnar and King, 1998; King and Molnar, 2000].

- **moment of inertia**  $\rightarrow$  principal moments of inertia
- **moment statistical functions**  $\rightarrow$  statistical indices
- **Monge-Arrault-Marot-Morin-Allory scoring functions**  $\rightarrow$  scoring functions
- **Monte Carlo version of MTD**  $\rightarrow$  minimal topological difference
- **Moran coefficient**  $\rightarrow$  autocorrelation descriptors
- **Moreau–Broto autocorrelation**  $\rightarrow$  autocorrelation descriptors
- **Moreau chirality index**  $\rightarrow$  chirality descriptors
- **morphological similarity**  $\rightarrow$  Compass method
- **morphologic index**  $\rightarrow$  functional coordination index
- **Morgan's extended connectivity algorithm**  $\rightarrow$  canonical numbering
- **Moriguchi model based on structural parameters**  $\equiv$  *MLOGP*  $\rightarrow$  lipophilicity descriptors
- **Moriguchi model based on surface area**  $\rightarrow$  lipophilicity descriptors
- **Moriguchi polar parameter**  $\rightarrow$  lipophilicity descriptors

#### ■ Morovitz information index ( $I_{MOR}$ )

An information index accounting for the structural features of a molecule [Morovitz, 1955]. It is defined as

$$I_{MOR} = I_{AC} + I_{PB}$$

where the first term is the  $\rightarrow$  *total information index on atomic composition* and the second term is the **information on the possible valence bonds**  $I_{PB}$  defined as

$$I_{PB} = \sum_{g=1}^G A_g \cdot \log_2 V_g$$

where  $g$  runs over all the different atom types,  $A_g$  is the number of atoms of  $g$ th type, and  $V_g$  the number of possible bonds that can be formed by an atom of  $g$ th type, calculated as

$$V_{g,\delta} = \binom{6 + \delta - 1}{\delta} = \frac{(6 + \delta - 1)!}{\delta! \cdot 5!}$$

where 6 is assumed as the maximum possible valence and  $\delta$  is the actual valence of the  $g$ th-type atom. For example,  $V_H = 6$ ,  $V_O = 21$ ,  $V_{N,3} = 56$ , and  $V_{C,4} = 126$  [Bonchev, 1983].

- **MoRSE descriptors**  $\equiv$  *3D-MoRSE descriptors*  $\rightarrow$  molecular transforms
- **motor octane number**  $\rightarrow$  technological properties
- **Mozley similarity coefficient**  $\equiv$  *Forbes–Mozley similarity coefficient*  $\rightarrow$  similarity/diversity (Table S9)

- **MPEI**  $\equiv$  *Molecular Polarizability Effect Index*  $\rightarrow$  electric polarization descriptors ( $\odot$  polarizability effect index)
- **M-PEOE**  $\equiv$  *modified partial equalization of orbital electronegativities*  $\rightarrow$  electronegativity
- **MP-MFP descriptors**  $\rightarrow$  substructure descriptors ( $\odot$  structural keys)

### ■ MPR approach

This is an approach designed to calculate  $\rightarrow$  *local vertex invariants* (LOVIs) as the solutions of a linear equation system [Filip, Balaban *et al.*, 1987; Ivanciuc, Balaban *et al.*, 1992]:

$${}^a\mathbf{M} \cdot \mathbf{s} = \mathbf{r} \quad {}^a\mathbf{M} = \mathbf{M} + \mathbf{p} \cdot \mathbf{I}$$

where  ${}^a\mathbf{M}$  is a square  $A \times A$  matrix representing the  $\rightarrow$  *molecular graph* and defined according to the scheme for the  $\rightarrow$  *augmented matrices*,  $\mathbf{p}$  an  $A$ -dimensional column vector containing weights for graph vertices that are used as diagonal elements of the matrix  $\mathbf{M}$ ,  $\mathbf{r}$  the  $A$ -dimensional column vector of atomic properties,  $\mathbf{I}$  the identity matrix, and  $\mathbf{s}$  the  $A$ -dimensional column vector that is the solution of the system.

MPR (*Matrix–Property–Response*) descriptors are thus the elements of the vector **MPR** calculated as

$$\mathbf{MPR} \equiv \mathbf{s} = ({}^a\mathbf{M}^T \cdot {}^a\mathbf{M})^{-1} \cdot {}^a\mathbf{M}^T \cdot \mathbf{r}$$

The vertex properties encoded in the column vectors  $\mathbf{p}$  and  $\mathbf{r}$  can be either topological, for example,  $\rightarrow$  *vertex degree*,  $\rightarrow$  *vertex distance degree*, or chemical, for example, atomic number,  $\rightarrow$  *electronegativity*, and  $\rightarrow$  *ionization potential*.

Among the LOVIs obtained by this general approach, the most known are **AZV descriptors** derived from the  $\rightarrow$  *adjacency matrix*  $\mathbf{A}$  whose diagonal elements are substituted by the atomic numbers  $Z_i$  and the  $A$ -dimensional vector  $\mathbf{r}$  containing the vertex degrees  $\delta_i$ .

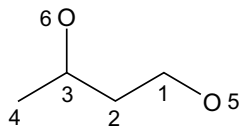
Different sets of LOVIs can be obtained by different choices of matrices and vectors defining the linear equation system; several combinations were studied on linear alkanes (Table M13).

**Table M13** A, adjacency matrix; D, distance matrix; V, vector of vertex degrees  $\delta_i$ ; S, vector of distance sums  $\sigma_i$ ; Z, vector of atomic numbers; N, vector of numbers of graph vertices; 1, unit vector. LOVI range values for linear alkanes.

ID	MPR	LOVIs range	ID	MPR	LOVIs range
1	AZV	0.1–1	11	DSN	0.05–0.7
2	ASV	0.01–0.2	12	DN <sup>2</sup> N	0.06–0.2
3	DSV	–0.02–0.12	13	ANS	1–4
4	AZS	2–9	14	ANV	0.08–0.5
5	ASZ	0.1–1	15	AZN	0.3–1.5
6	DN <sup>2</sup> S	0.1–0.3	16	ANZ	0.5–1.7
7	DN <sup>2</sup> 1	0–0.09	17	AN1	0.1–0.3
8	AS1	0.02–0.1	18	DSZ	0.06–0.6
9	DS1	0–0.3	19	ANN	0.7–0.9
10	ASN	0.2–0.7	20	DN <sup>2</sup> Z	0.03–0.5

**Example M5**

AZV descriptors for the H-depleted molecular graph of 1,3-butandiol.

	Atom	1	2	3	4	5	6	×	Atom	$s_i$	=	Atom	$\delta_i$
		1	6	1	0	0	1		1	$s_1$		1	2
		2	1	6	1	0	0		2	$s_2$		2	2
		3	0	1	6	1	0		3	$s_3$		3	3
		4	0	0	1	6	0		4	$s_4$		4	1
		5	1	0	0	0	8		5	$s_5$		5	1
		6	0	0	1	0	8		6	$s_6$		6	1

$\begin{cases} 6 \cdot s_1 + s_2 + s_5 = 2 \\ s_1 + 6 \cdot s_2 + s_3 = 2 \\ s_2 + 6 \cdot s_3 + s_4 + s_6 = 3 \\ s_3 + 6 \cdot s_4 = 1 \\ s_1 + 8 \cdot s_5 = 1 \\ s_3 + 8 \cdot s_6 = 1 \end{cases} \Rightarrow$	$\begin{aligned} s_1 &\equiv \text{AZV}_1 = 0.28284 \\ s_2 &\equiv \text{AZV}_2 = 0.21334 \\ s_3 &\equiv \text{AZV}_3 = 0.43708 \\ s_4 &\equiv \text{AZV}_4 = 0.09382 \\ s_5 &\equiv \text{AZV}_5 = 0.08996 \\ s_6 &\equiv \text{AZV}_6 = 0.07036 \end{aligned}$
--	--

**Triplet topological indices** (*TTI*) or, simply, **triplet indices**, are derived from local vertex invariants calculated by the MPR approach by using the common functions defined for the calculation of  $\rightarrow$  *graph invariants*. The most frequent functions used to generate triplet TIs are [Basak, Balaban *et al.*, 2000; Basak, Gute *et al.*, 2003]:

$$\begin{aligned} 1. \quad TTI_1(\text{MPR}) &= \sum_{i=1}^A \text{MPR}_i^\lambda \\ 2. \quad TTI_2(\text{MPR}) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\text{MPR}_i \cdot \text{MPR}_j)^\lambda \\ 3. \quad TTI_3(\text{MPR}) &= A \cdot \left( \prod_{i=1}^A \text{MPR}_i \right)^{1/A} \end{aligned}$$

where  $\lambda$  is a real exponent, usually taking values 1, 1/2, or 2 in function 1 and value  $-1/2$  in function 2;  $A$  is the number of graph vertices and  $a_{ij}$  are the elements of the adjacency matrix.

Closely related to MPR descriptors are local vertex invariants called **graph potentials** and denoted by  $U_i$  [Golender, Drboglav *et al.*, 1981; Ivanciuc, Balaban *et al.*, 1992]. They are calculated as the solutions of a linear equation system defined as

$$\mathbf{W} \cdot \mathbf{s} = \mathbf{r}$$

where  $\mathbf{W}$  is a weighted graph-theoretical matrix,  $\mathbf{r}$  the  $A$ -dimensional column vector of atomic properties, and  $\mathbf{s}$  the  $A$ -dimensional column vector of solutions of the system, which are local invariants  $U_i$ .

The weighted matrix  $\mathbf{W}$  is defined as

$$[\mathbf{W}]_{ij} = \begin{cases} w_i + \sum_k w_{ik} & \text{if } i = j \\ -w_{ij} & \text{if } (i, j) \in E(G) \\ 0 & \text{if } (i, j) \notin E(G) \end{cases}$$

where  $w_i$  is any topological or chemical semipositive definite atomic property and the sum runs over the first neighbors of the  $i$ th atom;  $w_{ij}$  is any topological or chemical semipositive definite bond weight, and  $E(G)$  the set of edges of the molecular graph  $G$ . If the weights  $w$  are all set equal to one, then the  $W$  matrix is

$$W = V - A + I = L + I$$

where  $V$ ,  $A$ ,  $I$ , and  $L$  are the diagonal  $\rightarrow$  *vertex degree matrix*, the adjacency matrix, the identity matrix, and the  $\rightarrow$  *Laplacian matrix*, respectively.

Similar to graph potentials, another set of LOVIs was proposed based on the  $\rightarrow$  *geometry matrix*  $G$ , using as the diagonal terms the  $\rightarrow$  *Balaban distance connectivity index* and as the response vector the adjacency matrix  $A$  multiplied by the column vector  $z$  collecting the atomic numbers of all the non-hydrogen atoms [Beteringhe, Filip *et al.*, 2005]:

$$MPR = {}^aG^{-1} \cdot A \cdot z \quad \text{and} \quad {}^aG = G + J \cdot I$$

where  ${}^aG$  is the augmented geometry matrix; it must be noted that the diagonal terms of the geometry matrix are filled in by a constant term (the Balaban index  $J$  of the molecule and not by local vertex invariants). Moreover, the atomic properties  $r_i$  are obtained from the adjacency matrix and the atomic numbers  $Z_i$  as

$$r_i \equiv [A \cdot z]_i = \sum_{j=1}^A a_{ij} \cdot Z_j$$

where the summation accounts for the atomic numbers of vertices adjacent to the  $i$ th vertex.

From these local vertex invariants  $r_i$ , a molecular descriptor, called **Beteringhe–Filip–Tarko descriptor** and denoted as  $GJ(AZ)$ , was proposed as

$$GJ(AZ) = \frac{B}{A+B} \cdot \sum_{i=1}^A \log(r_i)^2$$

where  $A$  is the number of graph vertices and  $B$  the number of edges.

**Note.** The authors called this index as topological, although it depends on the molecular geometry.

📖 [Balaban, 1993a, 1994b]

- **MPS topological index**  $\equiv$  *detour index*  $\rightarrow$  detour matrix
- **MSA descriptors**  $\equiv$  *molecular shape analysis descriptors*  $\rightarrow$  molecular shape analysis
- **MS-WHIM descriptors**  $\rightarrow$  grid-based QSAR techniques ( $\odot$  G-WHIM descriptors)
- **MTD-ADJ method**  $\rightarrow$  minimal topological difference
- **MTD descriptors**  $\rightarrow$  minimal topological difference
- **MTD-MC method**  $\rightarrow$  minimal topological difference
- **MTD model**  $\rightarrow$  minimal topological difference
- **MTI' index**  $\equiv$  *S index*  $\rightarrow$  Schultz molecular topological index
- **MTD-PLS method**  $\rightarrow$  minimal topological difference
- **$m$ th order sparse matrix**  $\rightarrow$  algebraic operators ( $\odot$  sparse matrices)
- **Mulliken electronegativity**  $\rightarrow$  atomic electronegativity

- **Mulliken population analysis** → quantum-chemical descriptors
- **MULTICASE** → lipophilicity descriptors (⊖ Klopman hydrophobic models)
- **multicriteria decision making** → chemometrics (⊖ ranking methods)
- **multigraph** → graph
- **multigraph distance degree** → weighted matrices (⊖ weighted distance matrices)
- **multigraph distance matrix** → weighted matrices (⊖ weighted distance matrices)
- **multigraph factor**  $\equiv$  *atomic multigraph factor* → bond order indices (⊖ conventional bond order)
- **multigraph information content indices**  $\equiv$  *indices of neighborhood symmetry*
- **MultiLevel Chemical Compatibility** → scoring functions
- **Multilevel Neighborhoods of Atoms descriptors** → substructure descriptors (⊖ fingerprints)
- **multiple arc** → graph
- **multiple bond count** → multiple bond descriptors

### ■ multiple bond descriptors

The presence of multiple bonds in molecules is a fundamental chemical aspect, which characterizes molecular properties and reactivity.

The **bond multiplicity**  $m_{ij}$  represents the degree of bonding between two adjacent vertices  $v_i$  and  $v_j$  and the most common way to quantify it is by using → *bond orders* derived from quantum-chemical calculations or → *conventional bond orders*.

The most simple descriptors of the degree of unsaturation of a molecule are → *count descriptors* based on the presence of double bonds, triple bonds, and aromatic bonds; they are the **double-bond count (DB)**, the **triple-bond count (TB)**, and the **aromatic-bond count (AB)**.

The **MCB index** was proposed as the number of multiple CC bonds in the molecule accounting for double, triple, and aromatic bonds [Bakken and Jurs, 1999a] as

$$MCB = DB + TB + AB$$

→ *Partial Wiener indices* are other multiple bond descriptors derived by a splitting of the → *Wiener index* into different multiple bond contributions.

Among the first proposed simple multiple bond descriptors [Pellegrin, 1983], there are the **number of unsaturation sites (US)** defined as the number of double bonds plus the number of triple bonds in the molecule, that is,  $US = DB + TB$ , and the **unsaturation number (UN)** defined as the number of double bonds plus twice the number of triple bonds, that is,  $UN = DB + 2 \times TB$ . Moreover, the same author proposed the **degree of unsaturation (DU)** by also considering the number of rings  $C$  (the → *cyclomatic number*), that is,  $DU = DB + 2 \times TB + C = UN + C$ .

The unsaturation index  $UN$  can be expressed by a more general form, called **multiple bond count**, as

$$b^* = \sum_b \left( \pi_{ij}^* \right)_b - B$$

where  $\pi^*$  is the → *conventional bond order* and the summation runs over all  $B$  bonds. For saturated compounds,  $b^* = UN = 0$ .

The **formal oxidation number** of a carbon atom equals the sum of the → *conventional bond orders* with electronegative atoms; the C–N bond order in pyridine may be considered as 2 since

there is one such bond and 1.5 when there are two such bonds; the C–X bond order in pyrrole or furan may be considered as 1.

The **unsaturation index** *UI* was also defined as

$$UI = \log_2(1 + b)$$

where *b* is calculated as

$$b = 2N_C + 2 - N_H - N_X + N_N + N_P + 2(N_{O-S} - N_{SO_3})/2 - C$$

$N_C$ ,  $N_H$ ,  $N_X$ ,  $N_N$ ,  $N_P$ , and  $C$  are the number of carbon atoms, hydrogen, halogen, nitrogen, phosphorous, and independent cycles, respectively.  $N_{O-S}$  and  $N_{SO_3}$  are the number of oxygen atoms bonded to sulfur and the number of  $SO_3$  groups, respectively. When no sulfur atoms are present, this index can be easily calculated from the chemical formula; otherwise, it is coincident with the index calculated replacing *b* with  $b^*$ , the multiple bond count.

A general expression [Pellegrin, 1983], valid for any organic compound, was also given by defining the atom valencies as the following:

Symbol	Atom	Symbol	Atom
$\alpha$	Monovalent	$\delta$	Tetravalent
$\beta$	Divalent	$\epsilon$	Pentavalent
$\gamma$	Trivalent	$\xi$	Hexavalent

The total number of atoms in a compound is

$$A = \alpha + \beta + \gamma + \delta + \epsilon + \xi$$

and the total number of bonds is

$$B = \frac{1}{2} \cdot \alpha + \frac{2}{2} \cdot \beta + \frac{3}{2} \cdot \gamma + \frac{4}{2} \cdot \delta + \frac{5}{2} \cdot \epsilon + \frac{6}{2} \cdot \xi$$

Then, starting from the  $\rightarrow$  *Euler's formula* for a graph, corresponding to the usual expression for the calculation of the number of rings, that is, the  $\rightarrow$  *cyclomatic number* *C*:

$$C = B - A + 1$$

the degree of unsaturation was defined as

$$DU = -\frac{\alpha}{2} + \frac{\gamma}{2} + \delta + \frac{3}{2} \cdot \epsilon + 2 \cdot \xi + 1$$

The sum of all the multiple bonds, that is, the *MCB* index, plus the number of rings is called **index of hydrogen deficiency** (*IHD*) or **double bond equivalents** (*DBE*). This last index can be derived from the following general equation [Pellegrin, 1983; Badertscher, Bsichofberger *et al.*, 2001]:

$$IHD = 1 + \frac{1}{2} \cdot \left[ \sum_{i=1}^A (v_i - 2) \right]$$

where *A* is the number of atoms and  $v_i$  is the formal valence of the *i*th atom ( $\alpha$ ,  $\beta$ , ...,  $\xi$ ).

One drawback of this index is that the formal valence of each element must be known. This is not a problem with most of the organic molecules containing only C, H, N, O, and halogens, but could become a problem for molecules containing sulfur and phosphorous. Moreover, it cannot be applied to radicals, ions, and disjoint parts. Finally, it is not invariant to different molecular representations.

Unlike the index of hydrogen deficiency, the **degree of unsaturation** has the same value for any structural representation corresponding to a molecular formula and can be calculated for much variety of structure representations. In this approach, the valence electrons of an element are partitioned into bond electrons and electrons localized on an atom, as shown in Table M14.

**Table M14** Number of valence electrons, bond electrons, and localized valence electrons for some chemical elements.

Atom	No. valence electrons	Std. no. bond electrons	Std. no. localized valence electrons
H	1	1	0
Li	1	1	0
C	4	4	0
N	5	3	2
O	6	2	4
Halogens	7	1	6
Si	4	4	0
P	5	3	2
S	6	2	4

Then, the degree of unsaturation is defined for the molecular formula as [Badertscher, Bsichofberger *et al.*, 2001]

$$DU = 1 + \frac{1}{2} \cdot \left[ -Q + \sum_{i=1}^A (b_i - 2) \right]$$

where  $Q$  is total charge of the molecule (signed) and  $b_i$  the standard number of bond electrons of the  $i$ th element (Table M14). For any structural molecule representation,  $DU$  is defined as

$$DU = DB + 2 \cdot TB + C + (1 - D) + \frac{1}{2} \cdot ELE$$

where  $DB$  and  $TB$  are the number of double and triple bonds, respectively,  $C$  the number of molecule rings, the so-called  $\rightarrow$  *cyclomatic number*,  $D$  the number of disconnected parts.  $ELE$  is the number of excess localized electrons of a molecule, calculated as the difference between the actual number of electrons localized on all atoms and the sum of the standard numbers, as given in column 4 of Table M14.

Another unsaturation measure is the **Unsat index** proposed as [Zheng, Luo *et al.*, 2005]:

$$\text{Unsat} = NRG_{567} + DB + 2 \cdot TB + \frac{AB + 1}{2}$$

where  $NRG_{567}$  is the number of 5-, 6-, and 7-member rings,  $DB$  the number of double bonds,  $TB$  the number of triple bonds, and  $AB$  the number of aromatic bonds. A relative unsaturation

measure called **Unsat-p index** was also proposed as the ratio of the Unsat index to the number of atoms that do not have bonded hydrogens and halogens.

A multiple bond descriptor was proposed in terms of  $\rightarrow$  *valence vertex degree*  $\delta^v$ , called **DV index**, and defined as

$$DV = \sum_b \left[ (\delta_i^v)^{-1/2} + (\delta_j^v)^{-1/2} \right]_b$$

where the sum runs over all the multiple bonds and  $i$  and  $j$  denote the atoms forming the considered bond [Millership and Woolfson, 1980].

The **induction parameter** was proposed to estimate the interaction ability of polar and nonpolar groups in the molecule [Thomas and Eckert, 1984]; it is based on the degree of unsaturation and defined as

$$q_{ind} = 1 - \frac{DB}{A}$$

where  $DB$  is the number of double bonds in the molecule. For saturated molecules,  $q_{ind} = 1$ .  $\rightarrow$  *Multigraph information content indices* are  $\rightarrow$  *information indices* encoding the bond multiplicity in the molecules.

To take into account the absolute contribution that a single double-bond makes to the whole size and shape of alkene molecules, **second-grade structural parameters** were derived from a  $\rightarrow$  *molecular graph* [Zhang, Liu *et al.*, 1997]. The topological descriptors representing the size  $w$  and the shape  $P_{\bar{w}}$  related to the presence of a double-bond are, respectively, as

$$w = \frac{\sum_{i=1}^A (d_{ik} + d_{il})}{2W} \quad P_{\bar{w}} = {}^3f_k + {}^3f_l$$

where  $W$  is the  $\rightarrow$  *Wiener index*, that is, the total sum of distances in the molecular graph, and  $d$  represents the  $\rightarrow$  *topological distance* between two vertices;  $k$  and  $l$  denote the vertices incident with the considered double-bond and the sum runs over all  $A$  vertices of the molecular graph. In the shape descriptor  $P_{\bar{w}}$ ,  ${}^3f_k$  and  ${}^3f_l$  are the number of vertices at a distance 3 from vertices  $v_k$  and  $v_l$ , respectively, that is, their  $\rightarrow$  *vertex distance count*. The size descriptor  $w$  is derived from the Wiener index, whereas  $P_{\bar{w}}$  from the  $\rightarrow$  *polarity number*. An extension giving information about the presence of several double bonds can be the sum of the  $w$  and  $P_{\bar{w}}$  values defined above over all double bonds.

- **multiple correlation coefficient**  $\rightarrow$  regression parameters
- **multiple edge**  $\rightarrow$  graph
- **multiple graph**  $\equiv$  *multigraph*  $\rightarrow$  graph
- **multiple pharmacophore descriptors**  $\rightarrow$  substructure descriptors ( $\odot$  pharmacophore-based descriptors)
- **multiplicative Wiener index**  $\rightarrow$  Wiener index
- **multivariate K correlation index**  $\rightarrow$  statistical indices ( $\odot$  correlation measures)
- **multivariate entropy**  $\rightarrow$  model complexity ( $\odot$  information content ratio)
- **mutation and selection uncover models**  $\rightarrow$  variable selection
- **mutation graph**  $\equiv$  *Sachs graph*  $\rightarrow$  graph
- **MVI**  $\equiv$  *molecular volume index*  $\rightarrow$  volume descriptors