

C

- **CACTVS screen vectors** → substructure descriptors (⊙ structural keys)
- **Calculated LOGP** \equiv *CLOGP* → lipophilicity descriptors (⊙ Leo–Hansch hydrophobic fragmental constants)
- **Camilleri model based on surface area** → lipophilicity descriptors
- **Cammarata–Yau analysis** → Free–Wilson analysis
- **Cammarata–Yau model** → Free–Wilson analysis
- **Canberra distance** → similarity/diversity (⊙ Table S7)

■ **canonical numbering** (\equiv *unique atomic ordering; unique atomic code*)

This is a procedure that assigns unique labels to the graph vertices so that the resulting matrix representations are in canonical form. The principal aim is to find a suitable numerical code for each given graph, which characterizes the graph up to isomorphism [Kvasnička and Pospichal, 1990; Faulon, 1998; Ivanciuc, 2003b].

The main canonical ordering procedures are listed below. Several different → *local vertex invariants* showing regular variation from central to terminal vertices were studied for canonical numbering of graph vertices [Filip, Balaban *et al.*, 1987; Bonchev and Kier, 1992]; examples are → *vertex distance degree*, → *local connectivity indices*, → *electrotopological state indices*, → *weighted atomic self-returning walk counts*, → *Randić atomic path code*, → *MPR descriptors*, → *centric operator*, → *centrocomplexity operator*, → *vertex complexity*, and → *vertex distance complexity*. The → *iterative vertex and edge centricity algorithm* (IVEC) also provides canonical ordering of vertices and edges in the graph and an algorithm based on the eigenvalues and eigenvectors of the adjacency matrix of the graph was also proposed [Liu and Klein, 1991].

• **Morgan's extended connectivity algorithm** (\equiv *extended connectivity algorithm, ECA*)

Graph vertices are ordered on the basis of their extended connectivity values obtained after a number of iterations of the Morgan method until constant atom ordering is obtained in two consecutive steps [Morgan, 1965]. The **extended connectivity** (or **extended vertex degree**), denoted as EC_i , of a vertex is calculated as the iterative summation of connectivities of all first neighbors as the following:

$$EC_i^{k+1} = \sum_{j=1}^A a_{ij} \cdot EC_j^k$$

where a_{ij} are the elements of the \rightarrow *adjacency matrix*, being equal to one only for pairs of adjacent vertices and zero otherwise; at the beginning ($k = 0$) the connectivity of each atom is simply the \rightarrow *vertex degree* δ .

It must be pointed out that the extended connectivity EC^k of Morgan coincides with the \rightarrow *atomic walk count* ($awc^{(k)}$) calculated as row sum of the k th power of the adjacency matrix **A** [Razinger, 1982; Rücker and Rücker, 1993; Figueras, 1993]. Then, the **extended connectivity indices**, denoted by EC^k and defined as [Rücker and Rücker, 1993]

$$EC^k = \sum_{i=1}^A EC_i^k$$

where A is the number of graph vertices, coincide with the \rightarrow *molecular walk counts*.

The **normalized extended connectivity** (NEC_i) is derived as

$$NEC_i = \lim_{k \gg 1} \left(\frac{EC_i^k}{EC^k} \right) \cdot \sum_{i=1}^A EC_i^1$$

where the last summation coincides with twice the number of bonds in the molecular graph [Bonchev, Kier *et al.*, 1993].

The Morgan algorithm was later improved by a better formalization and considering stereochemical aspects [Wipke and Dyott, 1974a, 1974b]. The Stereochemically Extended Morgan Algorithm (SEMA) resulted in a higher discriminating ability of graph vertices than ECA [Wipke, Krishnan *et al.*, 1978]; it is based on the iterative summation of the properties of neighboring atoms.

📖 [Ouyang, Yuan *et al.*, 1999]

• first eigenvector algorithm (FEVA)

Vertices in a graph are ordered according to the relative magnitudes of the coefficients of the first eigenvector (corresponding to the largest eigenvalue) of the \rightarrow *adjacency matrix* **A** [Randić, 1975d]. Generally nonequivalent vertices have different coefficient magnitudes, while equivalent vertices, that is, vertices constituting same orbits, have to be distinguished according to some alternative rules.

To obtain a vertex numbering similar to the Morgan algorithm, the convention to associate label 1 to the vertex with the largest coefficient and label A (the total number of vertices in the graph) to that with the smallest coefficient was established.

The largest coefficients correspond to \rightarrow *central vertices*, the smallest to \rightarrow *terminal vertices* and their neighbors.

• smallest binary label (SBL)

This is a binary label assigned to each graph vertex that consists of the corresponding row of the \rightarrow *adjacency matrix*; this binary label can also be expressed as decimal number, as in the \rightarrow *decimal adjacency vector*. The unique numbering given by the smallest binary label of each vertex can be achieved by iteratively renumbering the vertices of the graph, that is, iterative reordering of the adjacency matrix rows [Randić, 1974, 1975c; Mackay, 1975].

• Jochum–Gasteiger canonical numbering

A canonical numbering algorithm where nonterminal atoms are treated first, monovalent atoms (hydrogen and nonhydrogen atoms) then being numbered correspondingly to the nonterminal atoms [Jochum and Gasteiger, 1977]. The algorithm is based on the following steps:

1. the nonterminal atoms are put into the same equivalence class on the basis of their \rightarrow *vertex eccentricity*; the classes are ordered according to increasing eccentricity values and the atoms within each equivalent class are then ordered separately by the following sequential rules, beginning with the first equivalence class;
2. for each equivalent class, the atom with the highest atomic number has priority;
3. the atom with the most free electrons has priority;
4. the atom with the highest number of first neighbors (i.e., highest vertex degree) has priority;
5. the atom that has a first-neighbor atom with an atomic number higher than the others has priority;
6. the atom that has a first-neighbor atom with more free electrons than the others has priority;
7. the atom that has more bonds to first-neighbor atoms than the others has priority;
8. the atom with the highest bond order to the heavier first-neighbor atom has priority.
9. the atom that lies closer to an atom already numbered has priority;
10. the atom that has a higher bond order to an atom already numbered has priority.

Finally, terminal atoms are then numbered according to rules 2 and 9.

• hierarchically ordered extended connectivities algorithms (\equiv HOC algorithms)

HOC algorithm is an iterative procedure that finds topological equivalence classes (i.e., graph orbits) and provides canonical numbering of vertices in molecular graphs. It is based on the \rightarrow *extended connectivity* like Morgan's algorithm but also on the hierarchical ordering at each stage provided by the rank of the previous iteration [Balaban, Mekenyan *et al.*, 1985a].

The whole procedure consists of some algorithms that allow handling using graphs of different levels of complexity.

The main algorithm for ordering the vertices in graph orbits is called HOC-1 and the steps are

Step 1. Vertices of the \rightarrow *H-depleted molecular graph* are partitioned into equivalence classes according to their \rightarrow *vertex degree* δ_i , that is, a first rank ${}^1K_i = \delta_i$ is assigned to each vertex.

Step 2. For each vertex, the first ranks of its adjacent vertices are listed in increasing order as

$${}^1K_i^1, {}^1K_i^2, \dots, {}^1K_i^{\delta_i}$$

Step 3. An additional discrimination within each class is performed by means of the extended connectivities EC, which are the sums of the vertex degrees (ranks) of the nearest

neighbors, as

$${}^1\text{EC}_i = \sum_{r=1}^{\delta_i} {}^1K_i^r$$

where the sum runs over the neighbor ranks of the i th vertex considered in increasing order. A second rank ${}^{\text{II}}K_i$ is assigned to each vertex according to the increasing order of the extended connectivities. When two or more vertices are of the same rank, that is, ${}^{\text{II}}K_i = {}^{\text{II}}K_j$, but the individual values of the terms contributing to the extended connectivity are different, that is, ${}^1K_i^r \neq {}^1K_j^r$, then the ordering is made according to the rank of the first different addendum.

Step 4. Steps 2 and 3 are iteratively repeated, replacing first ranks by second ranks until all the ${}^{k+1}K_i$ ranks become equal to kK_i ranks of the preceding stage for all vertices.

The HOC-3 algorithm is used for the canonical numbering of graph vertices and is equal to the HOC-1 algorithm from step 1 to step 4 with an additional step based on an artificial discrimination into the largest orbits including two or more vertices. To make such a discrimination, one of the vertices inside the largest orbit, with the highest cardinality kK_i , is arbitrarily assigned a higher rank ${}^kK_i + 1$, and steps 2–4 are iteratively repeated.

The final vertex numbering is the inverse of the final HOC ranks so as to assign the lowest numbers to the most central vertices.

The HOC-2 and HOC-2A algorithms were proposed for the vertex ordering of special molecular graphs with pericondensed rings.

Based on HOC ranks, a *Unique Topological Representation* (UTR) was proposed in which topological equivalent vertices of the same rank in the graph are placed at the same level.

Moreover, two **HOC rank descriptors** based on the extended connectivity of graph vertices and the vertex rank obtained by HOC algorithms were proposed [Mekenyan, Bonchev *et al.*, 1984b]:

$$M = \sum_{i=1}^A M_i \quad \text{and} \quad N = \sum_{i=1}^A M_i^2$$

where the summation runs over all atoms and the term M_i is a local invariant defined as

$$M_i = \sum_{j=1}^i S_j \quad \text{and} \quad S_j = \sum_m K_m$$

where S_j is the sum of the ranks K of the adjacent vertices to the j th vertex restricted to those adjacent vertices of rank greater than j . The S_j values calculated for each vertex are ordered according to the increasing j index and are then summed up to i index to give the M_i local invariant. By this procedure, M_i gives a nondecreasing sequence.

Both descriptors increase with the size and cyclicity of the graph.

📖 [Mekenyan, Bonchev *et al.*, 1984a, 1985; Balaban, Mekenyan *et al.*, 1985b; Bonchev, Mekenyan *et al.*, 1985; Mekenyan, Balaban *et al.*, 1985; Ralev, Karabunarliev *et al.*, 1985]

• matrix method for canonical ordering

Graph vertices are partitioned and ordered into topological equivalence classes, that is, orbits, according to some special matrices developed for each atom [Bersohn, 1987]. These matrices give a representation of the whole molecule as seen from the considered atom.

The procedure consists first in assigning to each atom an atomic property P_i defined as

$$P_i = 1024 \cdot Z_i + 64 \cdot N_i^{\text{uns}} + 16 \cdot (4 - h_i)$$

where Z_i , N_i^{uns} , and h_i are the atomic number, the number of unsaturation, and the number of attached hydrogen atoms of the i th atom, respectively. The unsaturation number is 8 for an atom involved in a triple bond, 4 for either of the atoms in a double bond involving a heteroatom, 6 for an allene or ketene central atom with two double bonds, 2 for vinyl carbon atoms, 1 for aromatic atoms in a six-membered ring, and finally 0 for saturated atoms. The coefficients 1024, 64, and 16 have been chosen so that two atoms with the same atomic number cannot have the same property values.

The matrix representing the environment of the i th atom contains in the m th row the property values P_{mj} of the atoms located at a distance equal to m from the i th atom. The first row collects the property values of the first neighbors of the considered atom. The P_{mj} values are listed in descending order in the first entries of the matrix; the other entries are set to zero but are of no significance. The matrix dimension can be chosen for convenience.

The first partition of the vertices depends on their property values P_i ; two atoms X and Y are considered topologically equivalent if: (i) they have identical matrices and (ii) if X has a neighbor I belonging to a different equivalence class, then there must exist a neighbor J of the atom Y in the same equivalence class as the atom I .

Finally, the canonical ordering is performed on the basis of P_i values and in the case of equivalence according to the values of the matrix. Atoms with the greatest values are assigned the smallest numerical labels in the canonical ordering, such atoms being closest to the graph center.

Property values can also be modified to take into account geometric isomerism and chirality.

• self-returning walk ordering

Graph vertices are ordered on the basis of their \rightarrow *self-returning walk counts*, that is, the number of walks of a given length starting and ending at the same vertex. In particular, local invariants representing the relative occurrence of the self-returning walks of the considered atom to all self-returning walks (SRWs) in the molecule, that is, \rightarrow *topological atomic charge* TAC_i , provide a canonical numbering of graph vertices [Bonchev, Kier *et al.*, 1993]. The most important factors influencing the ordering, that is, the number of SRWs, are vertex branching, centrality, and cyclicity.

📖 [Balaban, 1976c; Randić, 1977c, 1978a, 1980b, 1995b; Hall and Kier, 1977a; Randić, Brissey *et al.*, 1979, 1981; Wilkins and Randić, 1980; Bonchev and Balaban, 1981; Bonchev, Mekenyan *et al.*, 1986; Polanski and Bonchev, 1986a; Klopman and Raychaudhury, 1988; Herndon, 1988; Diudea, Horvath *et al.*, 1992; Balaban, Filip *et al.*, 1992; Balasubramanian, 1995a; Babic, Balaban *et al.*, 1995; Laidboeur, Cabrol-Bass *et al.*, 1997; Agarwal, 1998; Lukovits, 1999]

■ Cao-Yuan indices

A set of three topological indices defined for a \rightarrow *H-depleted molecular graph* in terms of the \rightarrow *distance matrix* \mathbf{D} , the squared \rightarrow *Harary matrix* \mathbf{D}^{-2} , the vector $\boldsymbol{\delta}$ of \rightarrow *vertex degrees*, and the vector \mathbf{p} of *ring degrees* [Cao and Yuan, 2001; Yuan and Cao, 2003].

The **ring degree** of a vertex v_i belonging to a cycle, denoted as ρ_i , is a local vertex invariant defined as the minimum number of nonhydrogen vertices bonded to vertex v_i , which must be removed to transform the i th vertex into an acyclic one [Cao and Yuan, 2001].

The definitions of the three Cao–Yuan descriptors are the following:

- **odd–even index (OEI)**

The odd–even index is derived from the **odd–even matrix OEM**, which is a binary matrix, defined as

$$[\mathbf{OEM}]_{ij} = \begin{cases} (-1)^{d_{ij}-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where the off-diagonal elements of the **OEM** matrix are ± 1 , depending on whether the topological distance d_{ij} is odd or even. The **OEI** matrix is then calculated by the Hadamard product of the **OEM** and \mathbf{D}^{-2} matrices as

$$\mathbf{OEI} = \mathbf{OEM} \otimes \mathbf{D}^{-2}$$

where \mathbf{D}^{-2} is the matrix whose elements are the reciprocal of the square topological distances. Finally, the odd–even index is calculated as

$$\text{OEI} = \sum_{i=1}^A \sum_{j \neq i}^A [\mathbf{OEI}]_{ij} = \sum_{i=1}^A \sum_{j \neq i}^A [[\mathbf{OEM}]_{ij} \cdot [\mathbf{D}^{-2}]_{ij}]$$

This index encodes information on interactions between vertices v_i and v_j , which are proportional to the inverse of their square distance.

- **vertex degree–distance index (VDI)**

It is defined as

$$\text{VDI} = \left(\prod_{i=1}^A f_i \right)^{1/A}$$

where f_i are the elements of the A -dimensional vector \mathbf{f} defined as the product between the \mathbf{D}^{-2} matrix and the vertex degree vector \mathbf{v} :

$$\mathbf{f} = \mathbf{D}^{-2} \cdot \mathbf{v}$$

In this way, the interactions between vertices v_i and v_j are determined not only by their distance, but also by their vertex degrees.

- **ring degree–distance index (RDI)**

To distinguish the different freedom of atoms belonging and not belonging to cycles, the RDI index is defined as

$$\text{RDI} = \left(\prod_{i=1}^A g_i \right)^{1/A}$$

where g_i are the elements of the A -dimensional vector \mathbf{g} defined as the product between the matrix \mathbf{D}^{-2} and the ring degree vector \mathbf{p} :

$$\mathbf{g} = \mathbf{D}^{-2} \cdot \mathbf{p}$$

As the ring degree is zero for vertices not in cycles, the RDI index is obviously zero for acyclic molecules.

• **edge degree-distance index (EDI)**

It is defined as [Yuan and Cao, 2003]

$$\text{EDI} = \left(\prod_{i=1}^A \text{ES}_i \right)^{1/A}$$

where ES_i are the elements of the A -dimensional vector \mathbf{ES} defined as the product between the matrix \mathbf{D}^{-2} and the vector of the \rightarrow bond vertex degree δ^b that, unlike the vertex degree, accounts for bond multiplicity:

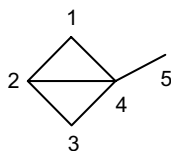
$$\mathbf{ES} = \mathbf{D}^{-2} \cdot \delta^b$$

This descriptor differs from VDI only for molecules containing multiple and/or aromatic bonds.

Note. The Authors use the name “edge degree” to refer to the bond vertex degree, but this is not correct because the edge degree was defined some years before [Bonchev, 1983] as the number of edges incident to an edge and not to a vertex.

Example C1

Cao-Yuan indices for 1-methylbicyclo[1.1.0]butane. \mathbf{v} is the vector of vertex degrees, \mathbf{p} the vector of ring degrees, \mathbf{f} the vector obtained by the product of matrix \mathbf{D}^{-2} and vector \mathbf{v} , \mathbf{g} the vector obtained by the product of matrix \mathbf{D}^{-2} and vector \mathbf{p} .


 $\mathbf{D} =$

Atom	1	2	3	4	5
1	0	1	2	1	2
2	1	0	1	1	2
3	2	1	0	1	2
4	1	1	1	0	1
5	2	2	2	1	0

 $\mathbf{D}^{-2} =$

Atom	1	2	3	4	5
1	0	1	0.25	1	0.25
2	1	0	1	1	0.25
3	0.25	1	0	1	0.25
4	1	1	1	0	1
5	0.25	0.25	0.25	1	0

 $\mathbf{OEI} =$

Atom	1	2	3	4	5
1	0	1	-0.25	1	-0.25
2	1	0	1	1	-0.25
3	-0.25	1	0	1	-0.25
4	1	1	1	0	1
5	-0.25	-0.25	-0.25	1	0

$\mathbf{v} = \{2, 3, 2, 4, 1\}$, $\mathbf{p} = \{1, 2, 1, 2, 0\}$, $\mathbf{f} = \{7.75, 8.25, 7.75, 8.00, 5.75\}$, $\mathbf{g} = \{4.25, 4.00, 4.25, 4.00, 3.00\}$

$\text{OEI} = 1 \times 12 - 0.25 \times 8 = 10.00$, $\text{VDI} = \text{EDI} = (7.75 \times 8.25 \times 7.75 \times 8.00 \times 5.75)^{1/5} = 7.44$
 $\text{RDI} = (4.25 \times 4.00 \times 4.25 \times 4.00 \times 3.00)^{1/5} = 3.87$

Boiling points of hydrocarbons were modeled by using Cao–Yuan indices, together with the fundamental contribution of the number of carbon atoms, as $N_C^{2/3}$.

- **capacity factor** → chromatographic descriptors
- **capacity factors** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **Carbó similarity index** → quantum similarity
- **cardinality layer matrix** → layer matrices
- **cardinality of a set** → algebraic operators
- **Carter resonance energy** → delocalization degree indices
- **Cartesian coordinates** → molecular geometry
- **CASE approach** → lipophilicity descriptors (⊙ Klopman hydrophobic models)
- **CAST** → molecular descriptors
- **CATS descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS2D descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS3D descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **CATS-charge descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **cavity term** → Linear Solvation Energy Relationships
- **cell-based density** → cell-based methods
- **cell-based entropy** → cell-based methods

■ cell-based methods (\equiv partition-based methods)

Cell-based methods, as well as clustering or distance-based methods, aim at extracting representative structurally diverse subsets of compounds from large chemical databases [Cummins, Andrews *et al.*, 1996; Mason and Pickett, 1997; Pearlman and Smith, 1999; Farnum, DesJarlais *et al.*, 2003]. They are mainly used in design and optimization of combinatorial libraries; the most important aspect being here to ensure maximum diversity within and between libraries before they are produced. Moreover, cell-based methods are used for lead discovery purposes allowing the selection of the compounds most similar to the active reference target.

Cell-based methods represent compounds in a p -dimensional space where each dimension represents either a molecular descriptor or a linear combination of molecular descriptors. Moreover, these methods partition the chemical space into hyper-rectangular regions, that is, the cells, in which the compounds are placed according to their property values, and measure the occupancy of the resulting cells by means of several cell-based diversity measures. The chemical space is defined by a number of selected molecular descriptors representing properties of compounds and ranges of the values of these descriptors are used to define the cells. Molecular descriptors selected to span the chemical space are commonly molecular properties that would be expected to affect ligand–receptor binding. The range of values of each molecular descriptor is divided into a set of subranges, that is, the bins. This can be accomplished by the use of different binning schemes. The binning scheme is the algorithm that is used to partition the descriptor value range into appropriate bins.

If n_j is the number of bins of the j th molecular descriptor, then the total number of cells in which the chemical space is partitioned is obtained from

$$N_{\text{TOT}} = \prod_{j=1}^p n_j$$

where p is the number of descriptors defining the chemical space. This number increases fast with the number of descriptors; for example, the number of cells for 3 and 5 descriptors, each divided into 10 bins, is 10^3 and 10^5 , respectively.

There are different binning schemes; they usually satisfy two simple criteria [Bayley and Willett, 1999]: (1) the maximum and minimum values for each of the descriptors that specify the partition must be set so as to encompass all of the compounds that may need to be processed by the partitioning scheme and (2) it seems appropriate that each molecule be assigned to just a single cell, thus requiring that the cell ranges do not overlap at all.

There are two basic ways in which partition can be generated. The simpler, *descriptor-independent partitioning scheme*, assumes that the bin boundaries used to subdivide the j th descriptor are completely independent of the bin boundaries that have been used to subdivide the preceding $j - 1$ descriptors. Alternatively, a *descriptor-dependent partitioning scheme* generates bin boundaries to subdivide the j th descriptor by taking account of the bin boundaries that have been used to subdivide the preceding $j - 1$ descriptors. Moreover, two simple criteria can be used for controlling the bin width (and hence the occupancy of each bin): each descriptor can be divided into equally sized bins or into equally occupied bins. It is hence possible to identify four types of binning schemes (Figure C1), depending on whether the bin boundaries are, or are not, independent of the preceding bin boundaries and on the occupancy criterion that is used to define each of the bins.

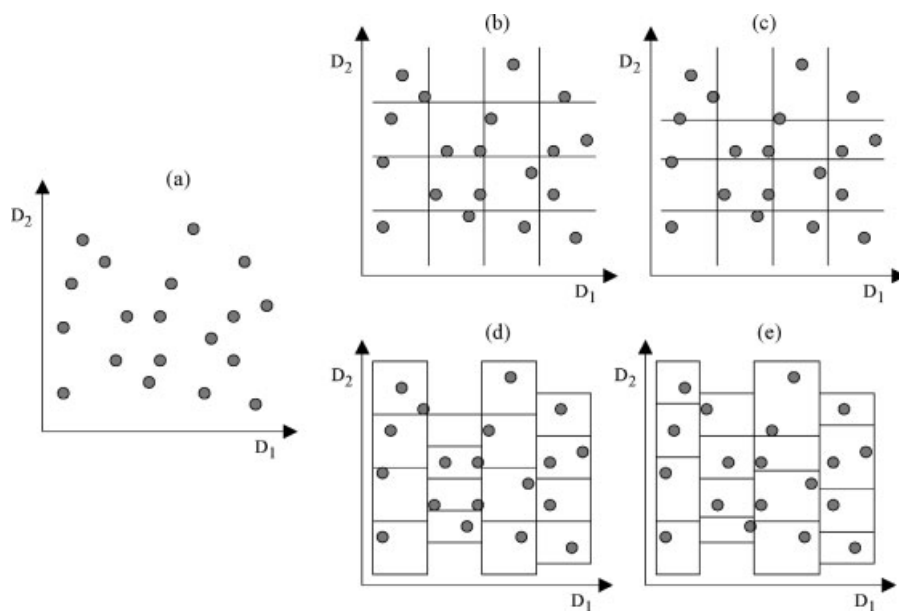


Figure C1 Example of the four binning schemes from [Bayley and Willett, 1999]. (a) Original data represented by $D_1 - D_2$ descriptors; (b) equisized independent binning scheme; (c) equifrequent independent binning scheme; (d) equisized dependent binning scheme; (e) equifrequent dependent binning scheme.

To select a subset of diverse compounds each molecule of the data set is assigned to the cell that matches the set of binned descriptors of the molecule; a structurally diverse subset is then obtained by selecting, for instance, one molecule from each of the cells to obtain the maximal coverage of the chemical space. On the contrary, in lead discovery, all the compounds falling in the same cell as a reference active compound are selected for further evaluation as candidates to be potential drugs.

Cell-based methods are significantly faster than distance-based methods, but are applicable to chemical space with low dimensionality (typically not more than five to six descriptors). In effect, when dimensionality is high, only a small fraction of the cells will be occupied, even with a low bin resolution. Moreover, the results are very sensitive to the grid resolution; indeed, if the bins are too large, the method loses its discriminating power; on the contrary, if the bins are too small, the data are very sparse, local behaviors of the data are highlighted and the general trend is lost. An algorithm based on a fractal approach was proposed to identify the optimal grid resolution by generating random subsets of k molecules from the whole data set, measuring their diversity at several grid resolutions, and identifying the resolution at which the relative variance of the diversity measure over all the random subsets assumes its maximum value [Agrafiotis and Rassokhin, 2002].

An advantage of cell-based methods is that they allow the explicit identification of those regions of the chemical space that are underrepresented, or, even unrepresented (i.e., diversity voids), in a database thus suggesting alternative potential structures to those of the existing chemicals [Pearlman and Smith, 1998].

Several cell-based diversity measures have been proposed in the literature [Pascual, Borrell *et al.*, 2003; Pascual, Mateu *et al.*, 2003]. These are \rightarrow *concentration indices*, such as χ^2 statistics, Gini concentration ratio, and Pratt measure, used to evaluate the distribution of compounds throughout the grid specified by a binning scheme and, thus, sometimes referred to as **occupancy numbers**.

The most natural index is the **cell occupancy ratio**, defined as the ratio of the number of occupied cells N_{OCC} over the total number of cells N_{TOT} :

$$COR = \frac{N_{\text{OCC}}}{N_{\text{TOT}}}$$

Another occupancy measure is the χ^2 statistics defined as

$$\chi^2 = \sum_{k=1}^{N_{\text{TOT}}} \frac{(n_k - n_k^*)^2}{n_k^*}$$

where n_k is the number of compounds in the k th cell and n_k^* is the expected theoretical number of compounds for the k th cell, that is, n/N_{TOT} , n being the total number of compounds in the data set.

Cell-based entropy (I_{cell}) and **cell-based density** (H_{cell}), both based on \rightarrow *information content*, are defined as

$$I_{\text{cell}} = - \sum_{k=1}^{N_{\text{TOT}}} (n_k \cdot \log_2(n_k)) \quad \text{and} \quad H_{\text{cell}} = - \sum_{k=1}^{N_{\text{TOT}}} \left(n_k \cdot \log_2 \left(\frac{n_k}{n_k^{\text{REF}}} \right) \right)$$

where n_k^{REF} is the number of compounds in each k th cell from a reference data set.

To quantify the performance of the partition when used for lead discovery purposes, each compound has to be associated with some activity data and the active cells are defined as those containing at least one active compound. Then, common \rightarrow *classification parameters for two-class problems* can be used. For instance, a partition performance measure is the deviation from the ideal situation evaluated by summing the number of inactive compounds found within the active cells and then dividing the sum by the total number of active molecules, that is, \rightarrow *error rate*.

\rightarrow *Property filters* are a particular implementation of partitioning methods; they are used to select drug-like or lead-like compounds from large chemical libraries. Like the cell-based methods, these filters are based on a partition of the chemical space but each selected molecular descriptor is divided into only two or three subranges of values. While property filters mainly aim at optimizing drug-likeness, cell-based methods at optimizing diversity of chemical libraries.

Three applications of cell-based methods are reported below.

The **Diverse Property-Derived method (DPD method)** is based on the partitioning of six noncorrelated molecular descriptors and physico-chemical properties [Ashton, Jaye *et al.*, 1996]. These are a lipophilicity descriptor (CLOGP), an electrotopological index calculated as normalized sum of the squares of the atomic \rightarrow *electrotopological state indices*, the number of hydrogen-bond acceptors (*HBA*), the number of hydrogen-bond donors (*HBD*), a flexibility index defined as the ratio of the \rightarrow *Kier shape descriptors* $^1\kappa$ over $^2\kappa$, and the aromatic density defined as the number of aromatic rings over the molecular volume (Table C1).

Table C1 Descriptors and ranges used in DPD method.

Descriptor	Bins	Ranges
CLOGP	4	<1.5; 1.5–4.0; >4.0; N^+
Electrotopological index	3	<10; 10–20; >20
HBA	2	0–2; ≥ 3
HBD	2	0–1; ≥ 2
Flexibility index	3	<3.5; 3.5–6.5; >6.5
Aromatic density	4	0; 0.01–3.5; 3.5–6.5; >6.5

Analogously to \rightarrow *Property and Pharmacophoric Features fingerprints (PPF fingerprints)*, **PDR-FP fingerprints** (or *Property Descriptor value Range-derived FingerPrints*) encode value ranges of 93 molecular descriptors that were selected on the basis of their potential to adopt class-selective value ranges for different activity classes [Eckert and Bajorath, 2006a]. Molecular descriptors were selected among those implemented in the program MOE, on the basis of a comparison of their value distribution in 26 different classes of activity. To select the descriptors that systematically respond to molecular features related to the different activities, the DynaMAD scoring function [Eckert, Vogt *et al.*, 2006] was used, which relates descriptor scores to the probability p of compounds to map a given activity range:

$$\text{score} = [1 - p(\text{classMin} \leq x \leq \text{classMax})] \cdot 100$$

where *classMin* and *classMax* are the minimum and maximum values within a class, respectively. This scoring function produces scores between zero, corresponding to no selectivity, and 100, corresponding to optimal selectivity of the descriptor.

Value ranges of molecular descriptors are encoded using from two to seven equiprobable nonoverlapping bins, leading to a final vector of 500 bits.

To generate the binary PDR-FP of a compound, its values for the 93 molecular descriptors are calculated, and for each descriptor (represented by n bins), it is determined into which of the predefined n bins the compound descriptor value falls. The associated bit is then set to 1; all other $n - 1$ bits are set to 0. When this scheme is followed, the bit string representation of any compound has exactly 93 bits that are set on.

Applications of PDR-FP fingerprints reported in literature are [Eckert and Bajorath, 2007a; Wang, Eckert *et al.*, 2007; Tovar, Eckert *et al.*, 2008].

The **Joint Entropy-based Diversity Analysis** (JEDA) is a method to select representative subsets of compounds from combinatorial libraries by using a scoring function based on the \rightarrow Shannon's entropy and implemented in a probabilistic search algorithm [Landon and Schaus, 2006].

Unlike other cell-based diversity methods, which select one compound from each cell of the chemical space, JEDA allows selection of more than one compound belonging to the same cell.

Compounds are described by a number of molecular descriptors; these are first normalized and then subjected to the \rightarrow Principal Component Analysis to reduce the dimensionality of the chemical space. The M most significant principal components are successively transformed into binary vectors where each bit corresponds to a single principal component (PC): the bit can be either 0 or 1 depending on whether the PC value is smaller or greater than the median of that component calculated on the whole library [Xue, Godden *et al.*, 2003b].

The median is here calculated as the value at which the entropy H of a molecular descriptor is maximal for the considered library:

$$H(t) = -\left(\frac{n_b}{n} \cdot \log_2 \frac{n_b}{n} + \frac{n_a}{n} \cdot \log_2 \frac{n_a}{n}\right)$$

where n , n_a , and n_b are the total number of compounds in the library, the number of descriptor values that fall above (a) and below (b) a threshold value t , respectively.

Because the chemical space is defined by M principal components and each component is partitioned into two regions, the chemical space is divided into 2^M cells.

To select the optimal subset of compounds, that is, a set of compounds having the maximal chemical diversity, a probabilistic search algorithm is applied, which consists in selecting a subset of compounds based on a probability assigned to each compound. This algorithm optimizes the \rightarrow joint entropy (JH) of the subset of selected compounds. The task is performed iteratively, assigning each i th compound an initial uniform probability $p_i = 1/n$, then calculating the score s_i that is added to the previous compound probability as

$$p'_i = p_i + s_i, \quad s_i = 1 - p_i^{\left(\frac{JH}{n \cdot T}\right)}$$

and renormalizing, at each step, the probability of the remaining compounds in the library so that the total probability of the library is equal to 1.

The exponent of the score function includes the joint entropy JH of the selected subset, the number of library compounds, and an adjustable parameter T used to control the speed of the search.

After the probability of compounds has been updated, the process is repeated until the probabilities of the compounds in the selected subset sum to 1.

📖 [Godden, Xue *et al.*, 2002b]

- **cell occupancy ratio** → cell-based methods
- **center distance-based criteria** → center of a graph

■ center of a graph

This is the set of central vertices and edges, whose definitions depend on the approach used to determine them [Bonchev, 1989]. The graph center can be a single vertex, a single edge, or a single group of equivalent vertices. Several graph center definitions are derived from approaches aimed at → *canonical numbering* of vertices. Other ways to identify central vertices are the → *pruning of the graph* and the application of → *centric operator* and → *centrocomplexity operator* to → *layer matrices* of the graph [Diudea, Horvath *et al.*, 1992].

According to the most popular definition, the central vertices in a graph are those vertices having the smallest → *atom eccentricity*. In acyclic graphs the center coincides with a single vertex (i.e., a central vertex) or two adjacent vertices (i.e., a single central edge), while in cyclic-containing graphs, it usually coincides with a group of vertices. Other local vertex invariants give information useful to distinguish between terminal and central vertices. → *Centric indices* are molecular descriptors that quantify the degree of compactness of molecules based on the recognition of the graph center.

Specifically applied to study general networks but valid also for molecular graphs, some simple descriptors of vertex centrality are the so-called **centrality measures**. The concept of centrality is related to the ability of a vertex to communicate with other vertices or to its closeness to many other vertices or to the number of pairs of vertices that need a specific vertex as intermediary in their communications [Freeman, 1977, 1979; Estrada, 2006b]. The two simplest centrality measures are the → *vertex degree* and the **degree centrality** DC_i , that is, the number of paths starting/ending at a vertex i [Albert, Jeong *et al.*, 1999].

The **betweenness centrality** BC_i characterizes the degree of influence a vertex has in communicating between vertex pairs and is defined as the fraction of shortest paths going through a given vertex i as [Freeman, 1977; Newman, 2005]

$$BC_i = \sum_{k=1}^{A-1} \sum_{j=k+1}^A \frac{\min P_{kj}(i)}{\min P_{kj}} \quad k, j \neq i$$

where $\min P_{kj}$ is the number of shortest paths connecting vertices k and j , and $\min P_{kj}(i)$ is the number of these shortest paths that pass through the vertex i . Moreover, a relative measure of betweenness centrality BC'_i is obtained by dividing the betweenness centrality BC_i by the maximal value relative to the central vertex of the corresponding → *star graph* as

$$BC'_i = \frac{2 \cdot BC_i}{A^2 - 3 \cdot A + 2}$$

where A is the number of vertices in the graph. From the relative betweenness centrality, a measure of dominance of the most central vertex is defined as

$$BC' = \frac{\sum_{i=1}^A [BC^* - BC'_i]}{A - 1}$$

where BC^* is the maximal centrality value for any vertex in the graph, that is, $\max(BC'_i)$.

The **closeness centrality** CC_i of the i th vertex is defined as [Freeman, 1979; Albert, Jeong *et al.*, 1999]

$$CC_i = \frac{A-1}{\sum_{j=1}^A d_{ij}} = \frac{A-1}{\sigma_i}$$

where A is the number of vertices in the graph and σ_i the \rightarrow *vertex distance degree*, that is, the sum of all distances from the i th vertex; the quantity σ_i , in the network context, is called **farness**.

The **eigenvector centrality** EC_i of a vertex i is derived from the leading eigenvector of the \rightarrow *adjacency matrix* \mathbf{A} representing a connected subgraph or component of the network [Bonacich, 1972, 2007]. It is defined as the i th component of the eigenvector associated to the largest eigenvalue of \mathbf{A} :

$$EC_i = \ell_{i1}$$

A vertex has high value of EC either if it is connected to many other vertices or if it is connected to others that themselves have high EC ; in effect, unlike degree centrality, which weights every neighbor equally, the eigenvector weights connections with others according to their centralities.

The **information centrality** IC_i is based on the information that can be transmitted between any two vertices in a connected network [Stephenson and Zelen, 1989]. It is defined as follows:

$$IC_i = \left[\frac{1}{A} \cdot \sum_{j=1}^A \frac{1}{[\mathbf{I}]_{ij}} \right]^{-1} \quad [\mathbf{I}]_{ij} = \begin{cases} (c_{ii} + c_{jj} - c_{ij})^{-1} & \text{if } i \neq j \\ \infty & \text{if } i = j \end{cases}$$

where c_{ij} are the elements of the matrix \mathbf{C} obtained by inverting the matrix \mathbf{B} , that is strictly related to the \rightarrow *Laplacian matrix* and defined as

$$\mathbf{B} = \mathbf{V} - \mathbf{A} + \mathbf{U}$$

\mathbf{A} is the adjacency matrix, \mathbf{V} the \rightarrow *vertex degree matrix*, that is, the diagonal matrix of the vertex degrees, and \mathbf{U} is the \rightarrow *unit matrix* with all its elements equal to one.

The \rightarrow *subgraph centrality* $C_S(i)$ accounts for the weighted participation of vertices in all subgraphs of the network and is defined as [Estrada and Rodríguez-Velázquez, 2005b]

$$C_S(i) = \sum_{j=1}^A (\ell_{ij})^2 \cdot e^{\lambda_j}$$

where ℓ_{ij} is the i th component of the eigenvector associated to the j th eigenvalue λ_j of the adjacency matrix. This index counts the times that a vertex takes part in the different connected subgraphs of the network, with smaller subgraphs having higher importance.

A **generalized graph center** concept is obtained by a hierarchy of criteria applied recursively so as to reduce the number of vertices qualifying as central vertices [Bonchev, Balaban *et al.*, 1980, 1981].

The graph **center distance-based criteria** 1D–4D for a vertex v_i to belong to the graph center are: *Criterion 1D* \equiv minimum \rightarrow *atom eccentricity* η_i (i.e., the largest distance from the i th vertex):

$$\min_i(\eta_i)$$

Criterion 2D \equiv for the vertices satisfying the first criterion, minimum \rightarrow vertex distance degree σ_i :

$$\min_i(\sigma_i)$$

Criterion 3D \equiv for the vertices satisfying the previous criteria, minimum number of occurrences of the largest distance in the \rightarrow vertex distance code:

$$\min_i({}^{\eta_i}f_i)$$

where ${}^{\eta_i}f_i$ is the frequency of the maximum distance η_i from the vertex v_i to any other vertex, that is, atom eccentricity. If the largest distance occurs the same number of times for two or more vertices, the frequency of the next largest distance $\eta_i - 1$ is considered and so on.

The graph vertices qualified as central according to the first three criteria constitute a smaller graph called **pseudocenter** (or **graph kernel**).

Criterion 4D \equiv iterative process of the first three criteria 1D–3D applied to the pseudocenter instead of the whole graph.

If more than one central vertex results from distance-based criteria, the graph center is called **polycenter**.

To discriminate even further among the vertices of the polycenter, the graph **center path-based criteria** 1P–4P can be applied to the polycenter:

Criterion 1P \equiv minimum \rightarrow vertex path eccentricity ${}^{\Delta}\eta_i$ (i.e., the largest distance from the i th vertex in the detour matrix):

$$\min_i({}^{\Delta}\eta_i)$$

Criterion 2P \equiv for vertices satisfying the first criterion, minimum \rightarrow vertex path sum π_i (i.e., the sum of the lengths m of all paths starting from the considered vertex v_i):

$$\min_i(\pi_i)$$

Criterion 3P \equiv for the vertices satisfying the previous criteria, minimum number of occurrences of the largest order path in the \rightarrow vertex path code:

$$\min_i({}^{\Delta}\eta_i P_i)$$

where ${}^{\Delta}\eta_i P_i$ is the number of paths of maximal length starting from vertex v_i to any other vertex. If the longest path occurs the same number of times for two or more vertices, the path count of the next largest order is considered and so on.

Criterion 4P \equiv iterative process of the first three criteria 1P–3P applied to the small set of vertices selected according to the above described procedure.

The central vertices resulting from the criteria 1P–4P are called **oligocenter**. To further discriminate among the vertices of the oligocenter, analogous graph **center self-returning walk-based criteria** 1W–4W can be applied.

All the criteria defined above can also be applied to search for central edges in the graph using information provided from the \rightarrow edge distance matrix. For example, the center distance-based criteria 1D–4D are defined as

Criterion 1D \equiv minimum \rightarrow bond eccentricity ${}^b\eta_i$:

$$\min_i({}^b\eta_i)$$

Criterion 2D \equiv for the edges satisfying the first criterion, minimum \rightarrow edge distance degree ${}^E\sigma_i$:

$$\min_i({}^E\sigma_i)$$

Criterion 3D \equiv for the edges satisfying the previous criteria, minimum number of occurrences of the longest distance in the \rightarrow edge distance code:

$$\min_i({}^b\eta_i f_i)$$

where ${}^b\eta_i f_i$ is the frequency of the maximum distance ${}^b\eta_i$ from the edge e_i to any other edge, that is, the bond eccentricity. If the longest distance occurs the same number of times for two or more edges, the frequency of the next longest distance ${}^b\eta_i - 1$ is considered and so on.

Criterion 4D \equiv iterative process of the first three criteria 1D–3D applied to the pseudocenter instead of the whole graph.

Moreover, the application of the center distance-based criteria on simultaneously both the vertex distance matrix **D** and the edge distance matrix ${}^E\mathbf{D}$ resulted in a new algorithm, called **Iterative Vertex and Edge Centricity algorithm** (IVEC), for graph center definition and vertex \rightarrow canonical numbering [Bonchev, Mekenyan *et al.*, 1989]. The graph center is selected through the sequential centric ordering of the graph vertices and edges, on the basis of their metric properties and incidence.

In the initial step, the distance-based criteria 1D–4D are applied to graph vertices to order them into equivalence classes identified by ranks 1, 2, 3, ... Rank 1 is assigned to the polycenter and the maximum rank to the most external vertices. Then the same procedure is applied to the edges on the basis of the edge distance matrix.

Additional discrimination within the vertex equivalence classes is obtained by summing the ranks of the edges incident to each vertex of the considered class. New ranks are assigned to the vertices on this basis: lower ranks are assigned to vertices with smaller sum. If the same rank sum is obtained for two or more vertices, the vertex for which the addendum in the sum is smaller is assigned the lower rank. The same operation is performed for the graph edges by summing the new ranks of their incident vertices.

The algorithm continues iteratively until the same vertex and edge equivalence classes are obtained in two consecutive iterations. The center of the graph then includes the vertices and edges of lowest rank.

A modified IVEC algorithm was also proposed to search for the center of graphs where multiple bonds are present [Balaban, Bonchev *et al.*, 1993].

📖 [Bonchev and Balaban, 1981, 1993; Bonchev, 1983; Barysz, Bonchev *et al.*, 1986]

■ center of a molecule (\equiv molecule center)

Molecule centers are reference points used to calculate distributional properties of the molecule and, mathematically speaking, are the first-order moments of property distributions. Arithmetic mean and weighted arithmetic mean are the common way to calculate centers.

For example, the **geometric center** of a molecule is defined as the average value of atom coordinates calculated separately for each axis:

$$\bar{x} = \frac{1}{A} \cdot \sum_{i=1}^A x_i \quad \bar{y} = \frac{1}{A} \cdot \sum_{i=1}^A y_i \quad \bar{z} = \frac{1}{A} \cdot \sum_{i=1}^A z_i$$

where A is the number of atoms in the molecule.

The weighted center is analogously defined, but each i th atom coordinate is weighted by w_i , which represents an atomic property:

$$\bar{x} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot x_i \quad \bar{y} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot y_i \quad \bar{z} = \frac{1}{W} \cdot \sum_{i=1}^A w_i \cdot z_i$$

where W is the sum of the weights over all atoms in the molecule. For example, if the \rightarrow *weighting scheme* w is based on the atomic masses m , the **center of mass** (or **barycenter**) of the molecule is obtained.

Geometric and mass centers of a molecule are not molecular descriptors, but they are commonly used as the reference origin in the calculation of several geometric descriptors to obtain invariance to translation and rotation of molecules (i.e., \rightarrow *TRI descriptors*).

By weighting atoms by atomic charges, the first-order moment of charges is the \rightarrow *dipole moment* in neutral molecules. Moreover, for molecules with zero net charge and non-vanishing dipole moment, the **center-of-dipole** was defined as the appropriate molecule center for multipolar expansions to obtain rotational invariance [Silverman and Platt, 1996].

Another definition of molecule center is obtained by applying the concept of *leverage* to 3D \rightarrow *molecular geometry*; the **atom leverage-based center** is defined as the set of atoms with the minimum value of the diagonal elements h_i of the \rightarrow *molecular influence matrix* \mathbf{H} derived from the centered spatial coordinates of the atoms in a molecule [Todeschini and Consonni, 2000], that is,

$$\{a_i : h_i = \min_i(h_i)\}$$

The molecular influence matrix is calculated as

$$\mathbf{H} = \mathbf{M} \times (\mathbf{M}^T \times \mathbf{M})^{-1} \times \mathbf{M}^T$$

where \mathbf{M} is the rectangular matrix of dimension $A \times 3$ of the atom spatial coordinates (x, y, z) , that is, the \rightarrow *molecular matrix*. The diagonal values h_i are always between 0 and 1.

- **center-of-dipole** \rightarrow center of a molecule
- **center of mass** \rightarrow center of a molecule
- **center path-based criteria** \rightarrow center of a graph
- **center self-returning walk-based criteria** \rightarrow center of a graph
- **central edges** \rightarrow graph
- **centrality measures** \rightarrow center of a graph
- **central moments** \equiv *moments about the mean* \rightarrow statistical indices (\odot moment statistical functions)
- **centralization** \rightarrow distance matrix
- **central vertices** \rightarrow graph

■ centric indices

\rightarrow *Molecular descriptors* proposed to quantify the degree of compactness of molecules by distinguishing between molecular structures organized differently with respect to their centers. Based on the recognition of the \rightarrow *graph center*, these indices are mainly defined by the

information theory concepts applied to a partition of the graph vertices made according to their positions relative to the center. Moreover, \rightarrow *centric operator* and \rightarrow *centrocomplexity operator* have been proposed to calculate \rightarrow *local vertex invariants* from \rightarrow *layer matrices* and corresponding molecular descriptors, which account for molecular centrality.

The main centric indices are listed below; they have been divided into two main groups, one containing the indices proposed by Balaban and the other indices proposed by Bonchev.

- **Balaban centric index (B)**

A topological index defined for acyclic graphs based on the **pruning of the graph**, a stepwise procedure for removing all the \rightarrow *terminal vertices*, that is, vertices with a \rightarrow *vertex degree* of one ($\delta_i = 1$), and the corresponding incident edges from the \rightarrow *H-depleted molecular graph*. The vertices and edges removed at the k th step are n_k and the total number of steps to remove all vertices is R [Balaban, 1979].

The **pruning partition** of the graph is the reversed sequence of numbers n_k provided by the pruning procedure:

$$\{n_R, n_{R-1}, \dots, n_1\}$$

The pruning partition is related to \rightarrow *molecular branching* and the reversed order of numbers n_k is due to the fact that the number of branches cannot decrease when starting from the center of the tree. Moreover, the first entry is always equal to one (center) or two (bicenter) and the pruning partition is a partition of A , that is, the number of graph vertices; this means that

$$\sum_{k=1}^R n_k = A$$

The Balaban centric index is calculated from the pruning partition in analogy with the \rightarrow *first Zagreb index* M_1 as

$$B = \sum_{k=1}^R n_k^2$$

This index provides a measure of molecular branching: the higher the value of B , the more branched is the tree. It is called centric index because it reflects the topology of the tree as viewed from the center.

The **normalized centric index** C is derived by normalization, that is, imposing the same lower bound equal to zero for the least branched (linear) trees, on all the graphs. It is defined as

$$C = \frac{B - 2 \cdot A + U}{2}$$

where A is the number of graph vertices. The term U is defined as

$$U = \frac{1 - (-1)^A}{2} = \begin{cases} 0 & \text{if } A \text{ is even} \\ 1 & \text{if } A \text{ is odd} \end{cases}$$

The **binormalized centric index** C' is derived by a binormalization, that is, imposing on all the graphs the same lower bound and an upper bound equal to one for star graphs. In practice, it is

obtained from the normalized centric index C dividing it by the corresponding value of the star graph:

$$C' = \frac{B - 2 \cdot A + U}{(A - 2)^2 - 2 + U}$$

The binormalized centric index provides information on the topological shape of trees in a similar way to the \rightarrow *binormalized quadratic index* Q' .

• **lopping centric information index (\bar{I}_B)**

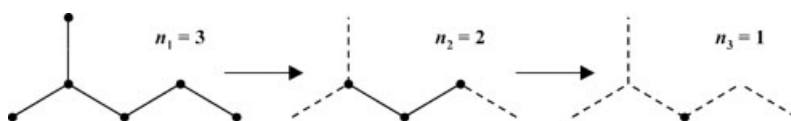
An index defined as the \rightarrow *mean information content* derived from the pruning partition of a graph:

$$\bar{I}_B = - \sum_{k=1}^R \frac{n_k}{A} \cdot \log_2 \frac{n_k}{A}$$

where n_k is the number of terminal vertices removed at the k th step, A the number of graph vertices, and R the number of steps to remove all graph vertices [Balaban, 1979].

Example C2

Pruning partition of 2-methylpentane and some centric indices.



Pruning partition: {1, 2, 3}

$$B = n_1^2 + n_2^2 + n_3^2 = 3^2 + 2^2 + 1^2 = 14 \quad \bar{I}_B = -\frac{1}{6} \cdot \log_2 \frac{1}{6} - \frac{2}{6} \cdot \log_2 \frac{2}{6} - \frac{3}{6} \cdot \log_2 \frac{3}{6} = 1.459$$

$$C = \frac{B - 2 \cdot A + U}{2} = \frac{14 - 2 \cdot 6 + 0}{2} = 1 \quad C' = \frac{B - 2 \cdot A + U}{(A - 2)^2 - 2 + U} = \frac{14 - 2 \cdot 6 + 0}{(6 - 2)^2 - 2 + 0} = 0.143$$

• **information content based on center (IBC)**

Defined only for acyclic graphs and substituents, it is calculated as the \rightarrow *total information content* based on the shells around the center of the graph:

$$IBC = 2W \cdot \log_2 2W - \sum_k q_k \cdot \log_2 q_k$$

where W is the \rightarrow *Wiener index*, that is, the sum of all the distances in the graph and q_k is the sum of the \rightarrow *vertex distance degree* (i.e., the sum of all distances from a vertex) of the vertices located at a \rightarrow *topological distance* equal to k from the center [Balaban, Bertelsen *et al.*, 1994].

- **average information content based on center (AIBC)**

Defined only for acyclic graphs and substituents, it is the average of the information content based on center *IBC*, that is,

$$AIBC = \frac{IBC}{W}$$

IBC is divided by the \rightarrow *Wiener index* *W* rather than $2W$ to have *AIBC* values higher than one [Balaban, Bertelsen *et al.*, 1994].

Bonchev centric information indices are centric indices derived from the vertex \rightarrow *distance matrix* *D* and the \rightarrow *edge distance matrix* *^ED*, based on the concept of graph center and calculated as \rightarrow *mean information content* [Bonchev, Balaban *et al.*, 1980; Bonchev, 1983, 1989].

For **vertex centric indices**, the number of equivalent vertices in each equivalence class is calculated applying the \rightarrow *center distance-based criteria* 1D–4D to the graph vertices, that is, the subsequent application of these criteria increases the discrimination of graph vertices. Analogously, for **edge centric indices**, the number of equivalent edges in each equivalence class is calculated applying the center distance-based criteria to the graph edges.

Once the \rightarrow *polycenter* of the graph has been found, four other centric information indices, called **generalized centric information indices**, are calculated on the vertex (edge) partition based on the average topological distance between each vertex (edge) and the atoms of the polycenter. An increasing discrimination of the graph vertices (edges) is obtained by subsequently applying the remaining criteria 2D–4D.

Other centric information indices can be calculated by the same formulas on both vertex and edge graph partition based on graph \rightarrow *center path-based criteria* 1P–4P and \rightarrow *center self-returning walk-based criteria* 1W–4W. Moreover, **edge centric indices for multigraphs** have different values from those calculated on the parent graph, the edge distance matrix of the multigraph being different from the edge distance matrix of the parent graph.

The Bonchev centric information indices and the corresponding generalized centric information indices are listed below.

- **radial centric information index (${}^V\bar{I}_{C,R}$)**

It is defined as

$${}^V\bar{I}_{C,R} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same \rightarrow *atom eccentricity*, that is, the maximum distance from a vertex to any other vertex in the graph, *G* the number of different vertex equivalence classes, and *A* the number of graph vertices.

- **distance degree centric index (${}^V\bar{I}_{C,deg}$)**

It is defined as

$${}^V\bar{I}_{C,deg} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having both the same atom eccentricity and the same \rightarrow *vertex distance degree* (i.e., the sum of all distances from a vertex), G the number of different vertex equivalence classes, and A the number of graph vertices.

• **distance code centric index** (${}^V\bar{I}_{C,code}$)

It is defined as

$${}^V\bar{I}_{C,code} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same atom eccentricity, the same vertex distance degree, and the same \rightarrow *vertex distance code* (i.e., occurrence number of distances of different length from a vertex), G the number of different vertex equivalence classes, and A the number of graph vertices.

• **complete centric index** (${}^V\bar{I}_{C,C}$)

It is defined as

$${}^V\bar{I}_{C,C} = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same atom eccentricity, the same vertex \rightarrow *distance degree*, the same \rightarrow *vertex distance code*, but also distinguishing the vertices defining the \rightarrow *pseudocenter*, that is, removing existing degeneracy of pseudocenter vertices. G is the number of different vertex equivalence classes and A is the number of graph vertices.

• **generalized radial centric information index** (${}^V\bar{I}_{C,R}^G$)

It is defined as

$${}^V\bar{I}_{C,R}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same average topological distance to the polycenter, G the number of different vertex equivalence classes, and A the number of graph vertices.

• **generalized distance degree centric index** (${}^V\bar{I}_{C,deg}^G$)

It is defined as

$${}^V\bar{I}_{C,deg}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having both the same average topological distance to the polycenter and the same vertex distance degree, G the number of different vertex equivalence classes, and A the number of graph vertices.

- **generalized distance code centric index** (${}^V\bar{I}_{C,code}^G$)

It is defined as

$${}^V\bar{I}_{C,code}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices having the same average topological distance to the polycenter, the same vertex distance degree, and the same vertex distance code, G is the number of different vertex equivalence classes, and A the number of graph vertices.

- **generalized complete centric index** (${}^V\bar{I}_{C,C}^G$)

It is defined as

$${}^V\bar{I}_{C,C}^G = - \sum_{g=1}^G \frac{n_g}{A} \log_2 \frac{n_g}{A}$$

where n_g is the number of graph vertices contemporarily having the same average topological distance to the polycenter, the same vertex distance degree, the same \rightarrow *vertex distance code*, but also distinguishing the atoms defining the pseudocenter, that is, removing existing degeneracy of pseudocenter atoms. G is the number of different vertex equivalence classes and A is the number of graph vertices.

- **edge radial centric information index** (${}^E\bar{I}_{C,R}$)

It is defined as

$${}^E\bar{I}_{C,R} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same \rightarrow *bond eccentricity* (i.e., the maximum value in each i th row of the edge distance matrix), G the number of different edge equivalence classes, and B the number of edges.

- **edge distance degree centric index** (${}^E\bar{I}_{C,deg}$)

It is defined as

$${}^E\bar{I}_{C,deg} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having both the same \rightarrow *bond eccentricity* and the same \rightarrow *edge distance degree* (i.e., the sum of the i th row entries of the edge distance matrix), G the number of different edge equivalence classes, and B the number of graph edges.

- **edge distance code centric index** (${}^E\bar{I}_{C,code}$)

It is defined as

$${}^E\bar{I}_{C,code} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same \rightarrow *bond eccentricity*, the same \rightarrow *edge distance degree*, and the same \rightarrow *edge distance code* (i.e., the occurrence of edge distance values for each i th edge), G is the number of different edge equivalence classes, and B the number of graph edges.

• **edge complete centric index** (${}^E\bar{I}_{C,C}$)

It is defined as

$${}^E\bar{I}_{C,C} = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same \rightarrow *bond eccentricity*, the same \rightarrow *edge distance degree*, the same \rightarrow *edge distance code*, but also distinguishing the edges defining the \rightarrow *pseudocenter*, that is, removing existing degeneracy of pseudocenter edges. G is the number of different edge classes and B is the number of graph edges.

• **generalized edge radial centric information index** (${}^E\bar{I}_{C,R}^G$)

It is defined as

$${}^E\bar{I}_{C,R}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same average topological distance to the polycenter, G the number of different edge equivalence classes, and B the number of edges.

• **generalized edge distance degree centric index** (${}^E\bar{I}_{C,\text{deg}}^G$)

It is defined as

$${}^E\bar{I}_{C,\text{deg}}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having both the same average topological distance to the \rightarrow *polycenter* and the same edge distance degree, G the number of different edge equivalence classes, and B the number of graph edges.

• **generalized edge distance code centric index** (${}^E\bar{I}_{C,\text{code}}^G$)

It is defined as

$${}^E\bar{I}_{C,\text{code}}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges having the same average topological distance to the \rightarrow *polycenter*, the same \rightarrow *edge distance degree*, and the same \rightarrow *edge distance code*; G is the number of different edge equivalence classes and B is the number of graph edges.

• **generalized edge complete centric index** (${}^E\bar{I}_{C,C}^G$)

It is defined as

$${}^E\bar{I}_{C,C}^G = - \sum_{g=1}^G \frac{n_g}{B} \log_2 \frac{n_g}{B}$$

where n_g is the number of graph edges contemporarily having the same average topological distance to the \rightarrow *polycenter*, the same \rightarrow *edge distance degree*, the same \rightarrow *edge distance code*, but also distinguishing the edges defining the pseudocenter, that is, removing existing degeneracy of pseudocenter edges. G is the number of different equivalence edge classes and B is the number of graph edges.

- **centricity** \equiv *molecular centricity* \rightarrow molecular complexity
- **centric operator** \rightarrow layer matrices
- **centric topological index** \rightarrow layer matrices
- **centrocomplexity operator** \rightarrow layer matrices
- **centrocomplexity topological index** \rightarrow layer matrices
- **CEP matrix** \equiv *weighted electronic connectivity matrix* \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **CFM** \equiv *Compressed Feature Matrix* \rightarrow substructure descriptors (\odot pharmacophore-based descriptors)
- **CGTA-axis system** \rightarrow biodescriptors (\odot DNA sequences)
- **CHAA₁ index** \rightarrow charged partial surface area descriptors
- **CHAA₂ index** \rightarrow charged partial surface area descriptors (\odot CHAA₁ index)

■ **chainlength**

The chainlength is defined as the size of the longest heavy-atom chain in the molecule with none of the constituent atoms of the chain belonging to rings [Feher and Schmidt, 2003].

- **chain subgraph** \rightarrow molecular graph
- **chance correlation** \rightarrow validation techniques
- **characteristic graph** \equiv *Sachs graph* \rightarrow graph
- **characteristic polynomial** \rightarrow algebraic operators

■ **characteristic polynomial-based descriptors**

These are various molecular descriptors derived from characteristic polynomials of the molecular graph G . They were originally used in the framework of the molecular orbital theory to study unsaturated compounds [Živković, Trinajstić *et al.*, 1975; Gutman, 1979, 1983; Graovac, Gutman *et al.*, 1977; Knop and Trinajstić, 1980; Rosenfeld and Gutman, 1989; Trinajstić, 1992; Gutman, Klavžar *et al.*, 2001; Noy, 2003]. Then, they were generalized to study any compound, finding a lot of applications in modeling physico-chemical properties of molecules [Hosoya, 1971, 1988; Trinajstić, 1988; Ivanciuc, Ivanciuc *et al.*, 1999b]. A comprehensive collection of characteristic polynomial-based descriptors with examples of calculation is presented in the reviews of Ivanciuc [Ivanciuc and Balaban, 1999c; Ivanciuc, Ivanciuc *et al.*, 1999a].

The \rightarrow *characteristic polynomial* of the molecular graph is the characteristic polynomial of a \rightarrow *graph-theoretical matrix* \mathbf{M} derived from the graph [Graham and Lovasz, 1978; Diudea, Ivanciuc *et al.*, 1997; Diudea, Gutman *et al.*, 2001; Ivanciuc, 2001c]:

$$\begin{aligned} Ch(\mathbf{M}; w; x) &= \det(x\mathbf{I} - \mathbf{M}) = \sum_{i=0}^n (-1)^i c_i x^{n-i} \\ &= x^n - c_1 x^{n-1} + c_2 x^{n-2} + \dots + (-1)^{n-1} c_{n-1} x + (-1)^n c_n \end{aligned}$$

where “det” denotes the matrix determinant, \mathbf{I} is the identity matrix of dimension $n \times n$, x is a scalar variable, and c_i are the $n + 1$ polynomial coefficients. \mathbf{M} is any square $n \times n$ matrix computed on weighted or unweighted molecular graphs; w is the \rightarrow *weighting scheme* applied to the molecular graph to encode chemical information. Note that $w = 1$ denotes unweighted graphs. If \mathbf{M} is a vertex matrix then n is equal to A , the number of graph vertices, while, if \mathbf{M} is an edge matrix, then n is equal to B , the number of graph edges. Polynomial coefficients are graph invariants and are thus related to the structure of a molecule graph.

A large number of graph polynomials were proposed in the literature, which differ from each other according to the molecular matrix \mathbf{M} they are derived from, and the weighting scheme w used to characterize heteroatoms and bond multiplicity of molecules.

The most known polynomial is the characteristic polynomial of the \rightarrow *adjacency matrix* ($\mathbf{M} = \mathbf{A}$), which is usually referred to as the **graph characteristic polynomial** [Harary, 1969a; Cvetković, Doob *et al.*, 1995; Bonchev and Rouvray, 1991; Trinajstić, 1992]:

$$Ch(\mathbf{A}; 1; x) = \det(x\mathbf{I} - \mathbf{A})$$

For any acyclic graph, two general rules are observed: (i) the power of x decreases by two and (ii) the absolute values of $Ch(\mathbf{A}; 1; x)$ coefficients are equal to the coefficients of the \rightarrow *Z-counting polynomial* $Q(\mathbf{G}; x)$, which are the nonadjacent numbers $a(\mathbf{G}, k)$ of order k , that is, the numbers of k mutually nonincident edges [Nikolić, Plavšić *et al.*, 1992]. Moreover, if two graphs are \rightarrow *isomorphic graphs*, their characteristic polynomials coincide, while the converse is not true, or, in other words, there exist nonisomorphic graphs with identical characteristic polynomials and spectra; these are called \rightarrow *isospectral graphs* [Harary, King *et al.*, 1971; Herndon, 1974a; Randić, Trinajstić *et al.*, 1976].

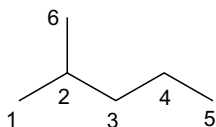
Depending on the elements of the matrix \mathbf{M} , characteristic polynomial can have very large coefficients and, spanning the x axis, often, asymptotic curves are obtained, whose characteristic points are not very representative as graph descriptors. To face with this problem, the characteristic polynomial can be transformed according to some **Hermite-like wave functions** for graphs, as [Gálvez, García-Domenech *et al.*, 2006]

$$\Psi \equiv Ch(\mathbf{M}; w; x) \exp\left(-\frac{x^2}{2}\right)$$

where $Ch(\mathbf{M}; w; x)$ is the characteristic polynomial of a graph. The most significant difference is that the area under the curve becomes finite in this approach, thus allowing the definition of more sound graph invariants, such as the *area under the curve* (AUC), the *maximum Ψ value* (Ψ^{\max}), and the *maximum amplitude* (MA) of the obtained sinusoidal curve.

Example C3

H-depleted molecular graph of 2-methylpentane and its adjacency matrix **A**.



A =

Atom	1	2	3	4	5	6
1	0	1	0	0	0	0
2	1	0	1	0	0	1
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	0
6	0	1	0	0	0	0

The characteristic polynomial of the adjacency matrix of 2-methylpentane is

$$Ch(\mathbf{A}; x) = x^6 - 5 \cdot x^4 + 5 \cdot x^2$$

where coefficients c_1 , c_3 , c_5 , and c_6 are zero. Absolute values of nonzero coefficients are $|c_0| = 1$, which corresponds to the nonadjacent number of zero order, $a(G, 0) = 1$ (by definition); $|c_2| = 5$, which corresponds to the nonadjacent number of first order, $a(G, 1) = 5$ (the number of graph edges); $|c_4| = 5$, which corresponds to the nonadjacent number of second order, $a(G, 2) = 5$ (the number of ways two edges may be selected so that they are nonadjacent).

The **Laplacian polynomial** is the characteristic polynomial of the \rightarrow *Laplacian matrix* **L** of the molecular graph [Ivanciuc, 1993; Gutman, Lee *et al.*, 1994; Trinajstić, Babic *et al.*, 1994; Gutman, Vidović *et al.*, 2002d; Gutman, 2003b; Cash and Gutman, 2004]:

$$Ch(\mathbf{L}; 1; x) = \det(x\mathbf{I} - \mathbf{L})$$

The **distance polynomial** is the characteristic polynomial of the \rightarrow *distance matrix* **D** of the molecular graph [Hosoya, Murakami *et al.*, 1973; Graham, Hoffman *et al.*, 1977; Graham and Lovasz, 1978]:

$$Ch(\mathbf{D}; 1; x) = \det(x\mathbf{I} - \mathbf{D}) = x^n - \sum_{i=1}^n c_i x^{n-1}$$

Note that the coefficients other than c_0 , which is always equal to one, are negative.

The **detour polynomial** is the characteristic polynomial of the \rightarrow *detour matrix* **Δ** of the molecular graph [Nikolić, Trinajstić *et al.*, 1999b]:

$$Ch(\mathbf{\Delta}; 1; x) = \det(x\mathbf{I} - \mathbf{\Delta}) = x^n - \sum_{i=1}^n c_i x^{n-1}$$

As for the distance polynomial, the coefficients other than c_0 , which is always equal to one, are negative.

The **reciprocal distance polynomial** is the characteristic polynomial of the \rightarrow *Harary matrix* **D⁻¹** of the molecular graph [Diudea, Ivanciuc *et al.*, 1997; Ivanciuc, Ivanciuc *et al.*, 1999b]:

$$Ch(\mathbf{D}^{-1}; 1; x) = \det(x\mathbf{I} - \mathbf{D}^{-1})$$

All these polynomials can also be calculated for the corresponding \rightarrow *weighted matrices*, according to different \rightarrow *weighting schemes* *w*.

Example C4

Laplacian, distance, reciprocal distance, and detour polynomials of 2-methylpentane are

$$Ch(\mathbf{L}; 1; x) = x^6 - 10 \cdot x^5 + 35 \cdot x^4 - 52 \cdot x^3 + 32 \cdot x^2 - 6 \cdot x$$

$$Ch(\mathbf{D}; 1; x) \equiv Ch(\mathbf{A}; 1; x) = x^6 - 84 \cdot x^4 - 368 \cdot x^3 - 580 \cdot x^2 - 368 \cdot x - 80$$

$$Ch(\mathbf{D}^{-1}; 1; x) = x^6 - 6.7083 \cdot x^4 - 8.3403 \cdot x^3 - 1.2843 \cdot x^2 + 1.9400 \cdot x + 0.6522$$

Note that, 2-methylpentane being an acyclic molecule, the detour polynomial coincides with the distance polynomial.

By analogy with the \rightarrow Hosoya *Z index* that, for acyclic graphs, can be calculated as the sum of the absolute values of the coefficients of the characteristic polynomial of the adjacency matrix, the **stability index** (or **modified Z index**) is a molecular descriptor calculated for any graph as the sum of the absolute values of the coefficients c_{2i} appearing alternatively in the characteristic polynomial of the adjacency matrix [Hosoya, Hosoi *et al.*, 1975]:

$$\tilde{Z} = \sum_{i=0}^{[A/2]} |c_{2i}|$$

where the square brackets indicate the greatest integer not exceeding $A/2$ and A is the number of graph vertices.

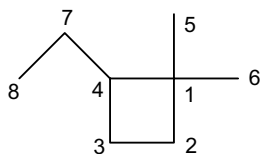
The same approach applied to the distance polynomial led to the definition of the **Hosoya *Z'* index** (or ***Z'* index**) [Hosoya, Murakami *et al.*, 1973]:

$$Z' = \sum_{i=0}^A |c_i|$$

where c_i are the coefficients of the distance polynomial of the molecular graph.

Example C5

H-depleted molecular graph and nonadjacent numbers of 4-ethyl-1,1-dimethylcyclobutane are



$$\begin{aligned} a(G, 0) &= 1 \\ a(G, 1) &= 8 \\ a(G, 2) &= 16 \\ a(G, 3) &= 8 \end{aligned}$$

The Hosoya *Z index* is $Z = 1 + 8 + 16 + 8 = 33$

The graph characteristic polynomial is $Ch(\mathbf{A}; 1; x) = x^8 - 8x^6 + 14x^4 - 6x^2$

The stability index is $\tilde{Z} = 1 + 8 + 14 + 6 = 29$

The distance polynomial is

$$Ch(\mathbf{D}; 1; x) = x^8 - 161x^6 - 1216x^5 - 3728x^4 - 5760x^3 - 4752x^2 - 2048x - 384$$

The *Z'* index is $Z' = 1 + 161 + 1216 + 3728 + 5760 + 4752 + 2048 + 384 = 18048$

An extension of the Z' index are the **Hosoya-type indices** that are defined as the sum of the absolute values of the coefficients of the characteristic polynomial of any square graph-theoretical matrix \mathbf{M} [Ivanciuc, 1999c, 2001c]:

$$\text{Ho}(\mathbf{M}; w) = \sum_{i=0}^n |c_i|$$

where n is the matrix dimension and w the \rightarrow *weighting scheme* applied to compute the matrix \mathbf{M} . The formula for the calculation of Hosoya-type indices was called by Ivanciuc **Hosoya operator**.

For any graph, when \mathbf{M} is the distance matrix of a simple graph, $\text{Ho}(\mathbf{D}; 1) = Z'$, when \mathbf{M} is the adjacency matrix of a simple graph, $\text{Ho}(\mathbf{A}; 1) = \tilde{Z}$; moreover, for acyclic graphs, when \mathbf{M} is the adjacency matrix of a simple graph, $\text{Ho}(\mathbf{A}; 1) = \tilde{Z} = Z$ (Hosoya Z index).

Example C6

Hosoya-type indices derived from adjacency \mathbf{A} , Laplacian \mathbf{L} , distance \mathbf{D} , reciprocal distance \mathbf{D}^{-1} , and detour $\mathbf{\Delta}$ matrices of 2-methylpentane in the case of unweighted molecular graph ($w = 1$).

$$\text{Ho}(\mathbf{A}; 1) \equiv Z = 1 + 5 + 5 = 11$$

$$\text{Ho}(\mathbf{L}; 1) = 1 + 10 + 35 + 52 + 32 + 6 = 136$$

$$\text{Ho}(\mathbf{D}; 1) \equiv \text{Ho}(\mathbf{\Delta}; 1) = 1 + 84 + 368 + 580 + 368 + 80 = 1401$$

$$\text{Ho}(\mathbf{D}^{-1}; 1) = 1 + 6.7083 + 8.3403 + 1.2843 + 1.9400 + 0.6522 = 19.9251$$

Table C2 Some Hosoya-type indices for the data set of 18 octane isomers (Appendix C – Set 1) calculated on the unweighted molecular graph and the following graph-theoretical matrices: \mathbf{A} , adjacency matrix; \mathbf{D} , distance matrix; \mathbf{L} , Laplacian matrix; \mathbf{D}^{-1} , reciprocal distance matrix; χ , χ matrix; \mathbf{G}^{-1} , reciprocal geometry matrix.

C8	Ho(A;1)	Ho(D;1)	Ho(L;1)	Ho(D ⁻¹ ;1)	Ho(χ;1)	Ho(G ⁻¹ ;1)
n-Octane	34	34049	987	53.689	5.281	84663.6
2M	29	31028	932	55.879	4.625	68727.9
3M	31	30513	924	48.781	5.042	69915.5
4M	30	30424	923	44.932	4.958	71521.5
3E	32	29889	915	42.933	5.375	74399.4
22MM	23	26516	848	57.278	3.938	64305.2
23MM	27	27413	868	48.934	4.500	70235.7
24MM	26	27656	872	51.514	4.389	65051.7
25MM	25	28181	880	57.591	4.056	63337.2
33MM	25	25748	836	48.749	4.438	67147.2
34MM	29	26969	861	46.440	4.944	74731.5
2M3E	28	26864	860	44.505	4.833	72374.3
3M3E	28	25049	825	44.242	5.063	72784.0

(Continued)

Table C2 (Continued)

C8	Ho(A;1)	Ho(D;1)	Ho(L;1)	Ho(D ⁻¹ ;1)	Ho(χ;1)	Ho(G ⁻¹ ;1)
223MMM	22	23168	788	51.570	3.917	70745.8
224MMM	19	23897	800	59.762	3.417	63623.3
233MMM	23	22925	784	48.839	4.083	71846.6
234MMM	24	24572	816	51.153	4.074	69364.3
2233MMMM	17	19685	720	55.535	3.125	74424.3

The characteristic polynomial encodes several important properties of the matrix, most notably its eigenvalues, spectral moments, determinant, and trace, which are largely used as molecular descriptors.

Information indices on polynomial coefficients are information indices defined as \rightarrow *total information content* and \rightarrow *mean information content* based on the partition of the coefficients of the characteristic polynomial of the graph. For acyclic molecules they coincide with the \rightarrow *Hosoya total information index* and \rightarrow *Hosoya mean information index*, respectively.

The **graph eigenvalues** λ_i are the roots of the characteristic polynomial of the matrix **M**, that is, the values of the x variable for which

$$Ch(\mathbf{M}; w; \lambda_i) = \det(\lambda_i \mathbf{I} - \mathbf{M}) = 0 \quad i = 1, n$$

where n is the size of the matrix **M**.

The complete set of the n eigenvalues of the matrix **M** is called **spectrum of the graph**:

$$\Lambda(\mathbf{M}; w) = \{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

Two fundamental properties of the characteristic polynomial of a matrix **M** are

$$tr(\mathbf{M}) = \sum_{i=1}^n \lambda_i = -c_1 \quad \det(\mathbf{M}) = \prod_{i=1}^n \lambda_i = (-1)^n \cdot c_n$$

where c_1 and c_n are two coefficients of the characteristic polynomial, “*tr*” denotes the trace of the matrix **M**, that is, the sum of its diagonal elements, and “*det*” the determinant of the matrix **M**. Therefore, it is noteworthy that the coefficient c_1 is always equal to zero for all the graph-theoretical matrices having zero on the main diagonal such as the adjacency, distance, and reciprocal distance matrices of simple graphs. For matrices derived from weighted molecular graphs, the coefficient c_1 , and, accordingly, the sum of the eigenvalues often is equal to the sum of the vertex weights, which are usually being located on the matrix main diagonal. For instance, in the Example C6, $-c_1$ in the characteristic polynomial of the Laplacian matrix **L** of 2-methylpentane is 10, which is the sum of the \rightarrow *vertex degrees* δ_i , that is, the numbers of adjacent vertices: $\delta_1 = 1$, $\delta_2 = 3$, $\delta_3 = 2$, $\delta_4 = 2$, $\delta_5 = 1$, and $\delta_6 = 1$.

A large number of \rightarrow *spectral indices*, which are molecular descriptors based on eigenvalues of molecular matrices, were defined both to study molecular graphs and model physico-chemical properties of molecules.

Spectral moments of the matrix **M**, denoted by $\mu^k(\mathbf{M}; w)$ and calculated with a weighting scheme w , are defined as

$$\mu^k(\mathbf{M}; w) = \sum_{i=1}^n \lambda_i^k = \sum_{i=1}^n [\mathbf{M}^k]_{ii}$$

where $k = 1, \dots, n$ is the order of the spectral moment, λ_i the eigenvalues of the matrix \mathbf{M} , and the last sum goes over the diagonal elements of the k th power of the matrix \mathbf{M} . Note that the spectral moment of order 1 simply is the sum of the matrix eigenvalues, coinciding with the sum of the diagonal elements of the matrix \mathbf{M} and the coefficient c_1 of its characteristic polynomial:

$$\mu^1(\mathbf{M}; w) = \sum_{i=1}^n \lambda_i = \text{tr}(\mathbf{M}) = -c_1$$

If \mathbf{M} is the adjacency matrix \mathbf{A} of a simple graph, each diagonal element of the k th power of the adjacency matrix \mathbf{A}^k is the \rightarrow *atomic self-returning walk count* of order k , that is, the number of self-returning walks of length k of each atom, and, thus, the sum of the atomic self-returning walk counts of all the atoms is the \rightarrow *molecular self-returning walk count* of order k , denoted as srw^k . Therefore, in the case of $\mathbf{M} = \mathbf{A}$, the following relationships hold:

$$\mu^k(\mathbf{A}; 1) = \text{srw}^k = \text{tr}(\mathbf{A}^k)$$

A number of theoretical studies and applications of spectral moments can be found in the literature [Živković, Trinajstić *et al.*, 1975; Jiang, Tang *et al.*, 1984; Hall, 1986; Kiang and Tang, 1986; Jiang and Zhang, 1989, 1990; Poshusta and McHughes, 1989; Marković and Gutman, 1991; Gutman, 1992a; Khadikar, Deshpande *et al.*, 1994; Bonchev and Seitz, 1995; Jiang, Qian *et al.*, 1995; Gutman and Rosenfeld, 1996; Helguera Morales, Cabrera Pérez *et al.*, 2006; Zhou, Gutman *et al.*, 2007].

\rightarrow *Spectral moments of the edge adjacency matrix* and \rightarrow *spectral moments of iterated line graph sequence* were largely investigated in QSAR/QSPR analysis by E. Estrada [Estrada, 1996, 1997, 1998a, 1998b, 1999c; Estrada, Peña *et al.*, 1998; Estrada and Gutierrez, 1999; Marković and Gutman, 1999; Marković, 1999, 2003; Marković, Marković *et al.*, 2001, 2002; Estrada, Paltewicz *et al.*, 2003].

Example C7

Eigenvalues and spectral moments of adjacency \mathbf{A} , Laplacian \mathbf{L} , distance \mathbf{D} , and reciprocal distance \mathbf{D}^{-1} matrices for 2-methylpentane in the case of unweighted molecular graph.

$$\begin{aligned}\Lambda(\mathbf{A}; 1) &= \{1.9021; 1; 1756; 0; 0; -1.1756; -1.9021\} \\ \Lambda(\mathbf{L}; 1) &= \{4.2143; 3; 1.4608; 1; 0.3249; 0\} \\ \Lambda(\mathbf{D}; 1) &= \{11.0588; -0.5115; -0.6730; -1.1726; -2.0000; -6.1717\} \\ \Lambda(\mathbf{D}^{-1}; 1) &= \{3.0788; 0.4827; -0.5000; -0.5714; -1.1271; -1.3631\} \\ \mu(\mathbf{A}; 1) &= \{0; 10; 0; 30; 0; 100\} \\ \mu(\mathbf{L}; 1) &= \{10; 30; 106; 402; 1580; 6341.8\} \\ \mu(\mathbf{D}; 1) &= \{0.53; 166.5; 1107.3; 16425.6; 156413.2; 1884474\} \\ \mu(\mathbf{D}^{-1}; 1) &= \{0; 13.4; 25.0; 95.1; 270.0; 860.2\}\end{aligned}$$

Note that spectral moments of the distance matrix increase very quickly, thus requiring a proper scaling to be used in QSAR/QSPR modeling.

Graph eigenvectors are the eigenvectors associated with the eigenvalues λ_i of the characteristic polynomial of a matrix \mathbf{M} ; for each $i = 1, \dots, n$, the following relationship holds:

$$Ch(\mathbf{M}; w; \lambda) = \det(\mathbf{M} - \lambda_i \mathbf{I}) = 0, \quad i = 1, \dots, n$$

where n being the number of eigenvalues and the size of the matrix \mathbf{M} .

Therefore, for each eigenvalue, $\mathbf{M} - \lambda_i \mathbf{I}$ is singular and, thus, it exists a nonzero n -dimensional vector \mathbf{v} satisfying:

$$\mathbf{M} \cdot \mathbf{v}_i = \lambda_i \cdot \mathbf{v}_i$$

Any vector \mathbf{v} satisfying this relationship is called eigenvector of \mathbf{M} for the eigenvalue λ_i .

Based on the eigenvectors of the adjacency and distance matrices, \rightarrow VEA indices, \rightarrow VRA indices, \rightarrow VED indices, and \rightarrow VRD indices were proposed as molecular descriptors.

Characteristic polynomials belong to a more general class of graph polynomials, which are used to encode some information on molecular graphs. Among these, there are \rightarrow Z-counting polynomial, \rightarrow matching polynomial, and \rightarrow Wiener polynomial.

Characteristic polynomials and Hosoya-type indices were also derived from \rightarrow distance-valency matrices, \rightarrow distance-path matrix, \rightarrow reciprocal distance-path matrix, \rightarrow distance-delta matrix, \rightarrow Szeged matrices [Ivanciuc and Ivanciuc, 1999], \rightarrow layer matrices, and \rightarrow edge adjacency matrix.

Characteristic polynomial, spectrum, spectral moments, eigenvectors, and Hosoya-type indices were also computed on square molecular matrices encoding information about spatial interatomic distances such as the \rightarrow geometry matrix \mathbf{G} and the \rightarrow reciprocal geometry matrix \mathbf{G}^{-1} [Ivanciuc and Balaban, 1999c].

📖 [Balaban and Harary, 1971; Randić, Trinajstić *et al.*, 1976; Balasubramanian, 1982, 1984a 1984b; Balasubramanian and Randić, 1982; Randić, 1982, 1983; Krivka, Jericevic *et al.*, 1985; Barysz, Nikolić *et al.*, 1986; Dias, 1987a, 1987b; Ivanciuc, 1988a, 1992, 1998d, 1998e, 2001b; Trinajstić, 1988; Rosenfeld and Gutman, 1989; Balasubramanian, 1990; Živković, 1990; Shalabi, 1991; Randić, Müller *et al.*, 1997; Cash, 1999; John and Diudea, 2004]

- **characteristic ratio** \rightarrow shape descriptors
- **characteristic root index** \rightarrow spectral indices
- **characteristic sequences** \rightarrow biodescriptors (☉ DNA sequences)
- **charge density matrix** \rightarrow quantum-chemical descriptors

■ charge descriptors

These are \rightarrow electronic descriptors defined in terms of \rightarrow atomic charges and used to describe electronic aspects both of the whole molecule and of particular regions, such as atoms, bonds, molecular fragments, and orbitals. The charges measure the extent of electronic density localization in a molecule: negative q_i values mean that excess electronic charge is at center i while positive values mean that center i is electron-deficient. Electrical charges in the molecule are the driving force of electrostatic interactions and it is well known that local electron densities or charges play a fundamental role in many chemical reactions, physico-chemical properties and receptor-ligand \rightarrow binding affinity.

Charge descriptors are calculated by methods of \rightarrow computational chemistry and are among \rightarrow quantum-chemical descriptors [Lowe, 1978; Streitwieser, 1961]. In the framework of quantum chemistry, \rightarrow population analysis is the basic tool used to calculate atomic charges and the most common approaches are the \rightarrow Mulliken population analysis and \rightarrow Löwdin population

analysis. Moreover, \rightarrow *partial equalization of orbital electronegativity* is the most popular approach for the calculation of partial atomic charges.

Charge descriptors are derived from atomic charges in different ways and a list of the most known descriptors is presented below.

- **maximum positive charge (Q_{\max}^+)**

The maximum positive charge of the atoms in a molecule:

$$Q_{\max}^+ = \max_i(q_i^+)$$

where q^+ are the net atomic positive charges.

- **maximum negative charge (Q_{\max}^-)**

The maximum negative charge of the atoms in a molecule:

$$Q_{\max}^- = \max_i(q_i^-)$$

where q^- are the net atomic negative charges.

- **total positive charge (Q^+)**

The sum of all of the positive charges of the atoms in a molecule:

$$Q^+ = \sum_i q_i^+$$

where q^+ are the net atomic positive charges.

- **total negative charge (Q^-)**

The sum of all of the negative charges of the atoms in a molecule:

$$Q^- = \sum_i q_i^-$$

where q^- are the net atomic negative charges.

- **total absolute atomic charge (Q)**

The sum over all atoms in a molecule of the absolute values of the atomic charges q_i :

$$Q = \sum_i |q_i|$$

This is a measure of molecule polarity, also called **Electronic Charge Index (ECI)**; for example, it has been used to study amino acid side chains [Collantes and Dunn III, 1995].

Moreover, the summation of the partial atomic charges of all the carbon atoms of the molecule was proposed for modeling hydrocarbons and called – quite improperly – **electrotopological descriptor**, denoted as $-\sum q_C$. To model properties of alcohols, partial charges of oxygen atoms were also accounted for, thus resulting into a different charge descriptor, denoted as $-(\sum q_C + \sum q_O)$ [Arupiyoti and Iragavarapu, 1998].

The **charge polarization** is the mean absolute atomic charge in a molecule, defined as

$$P = \frac{\sum_i |q_i|}{A} = \frac{Q}{A}$$

where A is the number of atoms.

Another measure of molecular polarity is obtained from the **total square atomic charge**, defined as

$$Q^2 = \sum_i q_i^2$$

These total charge descriptors can also be calculated restricted to a molecular fragment as well as to a functional group.

- **potential of a charge distribution (ϕ)**

The theoretical potential function of a discrete charge distribution, that is, of atomic point charges q_i , is given at point \mathbf{r} as the following:

$$\phi(\mathbf{r}) = \sum_{i=1}^A \frac{q_i}{r_i}$$

where r_i is the distance from each atom to the point \mathbf{r} .

- **Submolecular Polarity Parameter (SPP; ${}^1\Delta$)**

An electronic descriptor defined as the maximum excess charge difference for a pair of atoms in the molecule [Kaliszan, Osmialowski *et al.*, 1985; Osmialowski, Halkiewicz *et al.*, 1985], that is, calculated from the difference between the atomic maximum positive charge Q_{\max}^+ and the atomic maximum negative charge Q_{\max}^- in a molecule:

$${}^1\Delta = |Q_{\max}^+ - Q_{\max}^-|$$

The **second-order submolecular polarity parameter** ${}^2\Delta$ is determined analogously, and is the second largest difference of excess charges [Luco, Yamin *et al.*, 1995].

The interatomic distance r_{\pm} between the two atoms bearing the maximum positive and negative charges is used to derive the **DP descriptor** as follows:

$$DP = \frac{|Q_{\max}^+ - Q_{\max}^-|}{r_{\pm}^2} = \frac{{}^1\Delta}{r_{\pm}^2}$$

where the denominator accounts for the decreasing of atom interaction when interatomic distance increases.

- **topographic electronic descriptors (T^E)**

Topographic electronic descriptors are calculated from partial atomic charges q as the following [Osmialowski, Halkiewicz *et al.*, 1985, 1986; Katritzky and Gordeeva, 1993]:

$$T^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} \quad {}^cT^E = \sum_{b=1}^B \left(\frac{|q_i - q_j|}{r_{ij}^2} \right)_b$$

where the first index considers all pairs of atoms (both connected and disconnected) and the second is restricted to all pairs i - j of bonded atoms; r_{ij} are \rightarrow interatomic distances; A and B are the number of atoms and bonds, respectively. These descriptors are calculated in such a way that they reflect, to some extent, differences in size, shape, and constitution, these quantities affecting the electronic charge distribution and interatomic distances of the molecules.

- **partial charge weighted topological electronic index (PCWT^E)**

A molecular electronic descriptor defined as [Osmialowski, Halkiewicz *et al.*, 1986]

$$\text{PCWT}^E = \frac{1}{Q_{\max}^-} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{|q_i - q_j|}{r_{ij}^2} = \frac{T^E}{Q_{\max}^-}$$

where q are the Zefirov partial atomic charges [Zefirov, Kirpichenok *et al.*, 1987] of the atoms i and j , r_{ij} the corresponding interatomic distance, and Q_{\max}^- the maximum negative charge.

- **local dipole index (D)**

A molecular descriptor calculated as the average of the charge differences over all pairs i - j of bonded atoms:

$$D = \frac{\sum_b |q_i - q_j|_b}{B}$$

where B is the number of bonds [Clare and Supuran, 1994; Karelson, Lobanov *et al.*, 1996].

- **electronic-topological descriptors (E^T)**

Proposed by analogy with the \rightarrow connectivity indices, they are calculated for a \rightarrow hydrogen-included molecular graph using absolute values of partial charges q_i as vertex weights instead of \rightarrow vertex degrees δ as [Katritzky and Gordeeva, 1993]

$$\begin{aligned} {}^0E^T &= \sum_{i=1}^A (|q_i|)^{-1/2} & {}^1E^T &= \sum_{b=1}^B (|q_i \cdot q_j|)_b^{-1/2} \\ {}^2E^T &= \sum_{k=1}^{N_2} (|q_i \cdot q_l \cdot q_j|)_k^{-1/2} & {}^3E^T &= \sum_{k=1}^{^3P} (|q_i \cdot q_l \cdot q_h \cdot q_j|)_k^{-1/2} \end{aligned}$$

where A is the number of graph vertices, B the number of edges, N_2 the \rightarrow connection number, and 3P the number of paths of length three. Each term in the summations is the inverse square root of the product of the absolute partial charges of the vertices contained in the considered path.

- **charge-related indices**

They are global molecular descriptors derived from a \rightarrow H-depleted molecular graph where each vertex is weighted by a \rightarrow local vertex invariant called **Atom-in-Structure Invariant Index (ASII)** defined as [Bangov, 1988]

$$\text{ASII}_i = \text{ASII}_i^0 - h_i + q_i$$

where ASII_i^0 is a standard value for the atom-type and hybridization state of each i th atom (Table C3), h_i the number of hydrogen atoms bonded to the i th atom, and q_i its net \rightarrow atomic charge.

Table C3 Standard values of the Atom-in-Structure Invariant Index for different atom-types.

Atom	ASII ⁰	Atom	ASII ⁰
C sp ³	4	O sp ³	23
C sp ²	11	O sp ²	25
C sp ² (ar)	13	S	28
C sp	7	F	32

(Continued)

Table C3 (Continued)

Atom	ASII ⁰	Atom	ASII ⁰
N sp ³	15	Cl	33
N sp ²	18	Br	34
N sp	20	I	35

From *ASII*, the *ASIIg* index and **Charge Topological Index (CTI)** were derived as the following:

$$ASIIg = \frac{10}{\sqrt{\sum_{i=1}^A ASII_i}} \cdot \left[\sum_{b=1}^B (ASII_i \cdot ASII_j)_b \right]^{1/2}$$

$$CTI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A \frac{ASII_i \cdot ASII_j}{d_{ij}}$$

where *A* and *B* are the number of atoms and bonds, respectively, *d_{ij}* the topological distance between atoms *v_i* and *v_j*, and the summation in *CTI* index runs over all the pairs of atoms. The *ASIIg* index was particularly useful in dealing with isomers [Bangov, 1990]; the *CTI* index was proposed as a highly discriminant index with a low degree of degeneracy [Demirev, Dyulgerov *et al.*, 1991].

📖 [Del Re, 1958; Buydens, Massart *et al.*, 1983; Gasteiger, Röse *et al.*, 1988; Abraham and Smith, 1988; Baumer, Sala *et al.*, 1989; Gombar and Enslein, 1990; Dixon and Jurs, 1992; Reynolds, Essex *et al.*, 1992; Palyulin, Baskin *et al.*, 1995; Hannongbua, Lawtrakul *et al.*, 1996a; Payares, Díaz *et al.*, 1997]

■ Charged Partial Surface Area descriptors (≡ *CPSA descriptors*)

These constitute a set of different descriptors [Stanton and Jurs, 1990] that combine shape and electronic information to characterize molecules and, therefore, encode features responsible for polar interactions between molecules. The molecule representation used for deriving *CPSA* descriptors views molecule atoms as hard spheres defined by the → *van der Waals radius*. The → *solvent-accessible surface area* *SASA* is used as molecular surface area; it is calculated using a sphere with a radius of 1.5 Å to approximate the contact surface formed when a water molecule interacts with the considered molecule. Moreover, the contact surface where polar interactions can take place is characterized by a specific electronic distribution obtained by mapping atomic partial charges on the solvent-accessible surface.

Let *SA_a⁺* and *SA_a[−]* be the surface area contributions of the *a*th positive and negative atoms, respectively; *q_a⁺* and *q_a[−]* the partial atomic charges for the *a*th positive and negative atoms; and *Q⁺* and *Q[−]* the total sum of partial positive and negative charges in the molecule, respectively. The *CPSA* descriptors are defined as the following:

• partial negative surface area (*PNSA₁*)

It is the sum of the solvent-accessible surface areas of all negatively charged atoms, that is,

$$PNSA_1 = \sum_{a-} SA_a^-$$

where the sum is restricted to negatively charged atoms a^- .

- **partial positive surface area ($PPSA_1$)**

It is the sum of the solvent-accessible surface areas of all positively charged atoms, that is,

$$PPSA_1 = \sum_{a^+} SA_a^+$$

where the sum is restricted to positively charged atoms a^+ .

- **total charge weighted negative surface area ($PNSA_2$)**

It is the partial negative solvent-accessible surface area multiplied by the \rightarrow total negative charge Q^- , that is,

$$PNSA_2 = Q^- \cdot \sum_{a^-} SA_a^-$$

- **total charge weighted positive surface area ($PPSA_2$)**

It is the partial positive solvent-accessible surface area multiplied by the \rightarrow total positive charge Q^+ , that is,

$$PPSA_2 = Q^+ \cdot \sum_{a^+} SA_a^+$$

- **atomic charge weighted negative surface area ($PNSA_3$)**

It is the sum of the product of atomic solvent-accessible surface area by the partial charge q_a^- over all negatively charged atoms, that is,

$$PNSA_3 = \sum_{a^-} q_a^- SA_a^-$$

- **atomic charge weighted positive surface area ($PPSA_3$)**

It is the sum of the product of atomic solvent-accessible surface area by the partial charge q_a^+ over all positively charged atoms, that is,

$$PPSA_3 = \sum_{a^+} q_a^+ SA_a^+$$

- **difference in charged partial surface area ($DPSA_1$)**

It is the partial positive solvent-accessible surface area minus the partial negative solvent-accessible surface area, that is,

$$DPSA_1 = PPSA_1 - PNSA_1$$

- **difference in total charge weighted surface area ($DPSA_2$)**

It is the total charge weighted positive solvent-accessible surface area minus the total charge weighted negative solvent-accessible surface area, that is,

$$DPSA_2 = PPSA_2 - PNSA_2$$

- **difference in atomic charge weighted surface area ($DPSA_3$)**

It is the atomic charge weighted positive solvent-accessible surface area minus the atomic charge weighted negative solvent-accessible surface area, that is,

$$DPSA_3 = PPSA_3 - PNSA_3$$

- **fractional charged partial negative surface areas ($FNSA_1, FNSA_2, FNSA_3$)**

They are the partial negative surface area ($PNSA_1$), the total charge weighted negative surface area ($PNSA_2$), and the atomic charge weighted negative surface area ($PNSA_3$), divided by the total molecular solvent-accessible surface area ($SASA$), that is,

$$FNSA_1 = \frac{PNSA_1}{SASA} \quad FNSA_2 = \frac{PNSA_2}{SASA} \quad FNSA_3 = \frac{PNSA_3}{SASA}$$

- **fractional charged partial positive surface areas ($FPSA_1, FPSA_2, FPSA_3$)**

They are the partial positive surface area ($PPSA_1$), the total charge weighted positive surface area ($PPSA_2$), and the atomic charge weighted positive surface area ($PPSA_3$), divided by the total molecular solvent-accessible surface area ($SASA$), that is,

$$FPSA_1 = \frac{PPSA_1}{SASA} \quad FPSA_2 = \frac{PPSA_2}{SASA} \quad FPSA_3 = \frac{PPSA_3}{SASA}$$

- **surface weighted charged partial negative surface areas ($WNSA_1, WNSA_2, WNSA_3$)**

They are the partial negative surface area ($PNSA_1$), the total charge weighted negative surface area ($PNSA_2$), and the atomic charge weighted negative surface area ($PNSA_3$), multiplied by the total molecular solvent-accessible surface area ($SASA$) and divided by 1000, that is,

$$WNSA_1 = \frac{PNSA_1 \cdot SASA}{1000} \quad WNSA_2 = \frac{PNSA_2 \cdot SASA}{1000} \quad WNSA_3 = \frac{PNSA_3 \cdot SASA}{1000}$$

- **surface weighted charged partial positive surface areas ($WPSA_1, WPSA_2, WPSA_3$)**

They are the partial positive surface area ($PPSA_1$), the total charge weighted positive surface area ($PPSA_2$), and the atomic charge weighted positive surface area ($PPSA_3$), multiplied by the total molecular solvent-accessible surface area ($SASA$) and divided by 1000, that is,

$$WPSA_1 = \frac{PPSA_1 \cdot SASA}{1000} \quad WPSA_2 = \frac{PPSA_2 \cdot SASA}{1000} \quad WPSA_3 = \frac{PPSA_3 \cdot SASA}{1000}$$

- **relative negative charge ($RNCG$)**

It is the partial charge of the most negative atom divided by the \rightarrow total negative charge, that is,

$$RNCG = \frac{Q_{\max}^-}{Q^-}$$

- **relative positive charge (RPCG)**

It is the partial charge of the most positive atom divided by the \rightarrow total positive charge, that is,

$$RPCG = \frac{Q_{\max}^+}{Q^+}$$

- **relative negative charge surface area (RNCS)**

It is the solvent-accessible surface area of the most negative atom divided by the relative negative charge (RNCG), that is,

$$RNCS = \frac{SA_{\max}^-}{RNCG}$$

- **relative positive charge surface area (RPCS)**

It is the solvent-accessible surface area of the most positive atom divided by the relative positive charge (RPCG), that is,

$$RPCS = \frac{SA_{\max}^+}{RPCG}$$

- **total hydrophobic surface area (TASA)**

It is the sum of solvent-accessible surface areas of atoms with absolute value of partial charges less than 0.2, that is,

$$TASA = \sum_a SA_a \quad \forall a : |q_a| < 0.2$$

- **total polar surface area (TPSA)**

It is the sum of solvent-accessible surface areas of atoms with absolute value of partial charges greater than or equal to 0.2.

$$TPSA = \sum_a SA_a \quad \forall a : |q_a| \geq 0.2$$

- **relative hydrophobic surface area (RASA)**

It is the total hydrophobic surface area (TASA) divided by the total molecular solvent-accessible surface area (SASA), that is,

$$RASA = \frac{TASA}{SASA}$$

- **relative polar surface area (RPSA)**

It is the total polar surface area (TPSA) divided by the total molecular solvent-accessible surface area (SASA), that is,

$$RPSA = \frac{TPSA}{SASA}$$

Six additional CPSA descriptors were later proposed as [Aptula, Kühne *et al.*, 2003]

$$\begin{aligned} PPSA_4 &= \frac{Q^+}{A} \cdot \sum_{a^+} SA_a^+ & PNSA_4 &= \frac{Q^-}{A} \cdot \sum_{a^-} SA_a^- \\ PPSA_5 &= \frac{Q^+}{A^+} \cdot \sum_{a^+} SA_a^+ & PNSA_5 &= \frac{Q^-}{A^-} \cdot \sum_{a^-} SA_a^- \\ SPMX &= Q_{\max}^+ \cdot SA_{\max}^+ & SNMX &= Q_{\max}^- \cdot SA_{\max}^- \end{aligned}$$

where A is the total number of atoms, A^+ and A^- the total number of positively and negatively charged atoms, respectively, and the meaning of the other symbols is the same as above.

The set of CPSA descriptors was further developed to account for any particular type of polar interaction such as hydrogen-bonding. **Hydrogen-Bond Charged Partial Surface Area descriptors** (or **HB-CPSA descriptors**) were proposed in analogy with CPSA descriptors [Stanton, Egolf *et al.*, 1992]. Hydrogen-bond donor groups are considered to be any heteroatoms (i.e., O, S, or N) possessing a proton that can be donated. Other types of functional groups such as the alkynes were also included in the donor class. Acceptor groups include any functional group possessing sufficient electron density to participate in a hydrogen bond. To simplify the calculations the halogens, some double and some aromatic bonds were not included in the HB-CPSA descriptors.

Katritzky, Mu *et al.*, [Katritzky, Mu *et al.*, 1996b] later enlarged this set of hydrogen-bonding descriptors. All the H-bond descriptors are assigned zero if no hydrogen atoms in the molecule can be donated; moreover, hydrogen-bond acceptors are usually restricted to oxygen, nitrogen, and sulfur atoms (e.g., carbonyl oxygen atoms except in $-\text{COOR}$, hydroxy oxygen atoms, amino nitrogen atoms, aromatic nitrogens, and mercapto sulfur atoms).

The two simplest HB-CPSA descriptors are the \rightarrow *hydrogen-bond acceptor number HBA* and the \rightarrow *hydrogen-bond donor number HBD*.

Let SA_d and SA_a be the solvent accessible surface areas of hydrogen-bonding donors (d) and acceptors (a), respectively, $SASA$ the solvent-accessible surface area, and q_d and q_a the corresponding partial atomic charges. The HB-CPSA descriptors are then defined as follows (note that the two different symbols encountered in the literature for some are considered as synonymous).

- **RHTA index**

It is the ratio of the number of donor groups (HBD) over the number of acceptor groups (HBA), that is,

$$RHTA = \frac{HBD}{HBA}$$

- **SSAH index** (\equiv *HDSA index*)

It is the sum of the surface areas of the hydrogens, which can be donated:

$$SSAH \equiv HDSA = \sum_d SA_d$$

- **RSAH index**

It is the average surface area of hydrogens, which can be donated:

$$RSAH = \frac{\sum_d SA_d}{HBD}$$

where HBD is the number of hydrogen-bond donors.

- **RSHM index** (\equiv FHDSA index)

It is the fraction of the total molecular surface area associated with hydrogens, which can be donated:

$$RSHM \equiv FHDSA = \frac{\sum_d SA_d}{SASA}$$

- **SSAA index** (\equiv HASA index)

It is the sum of the surface areas of all H-bond acceptor atoms:

$$SSAA \equiv HASA = \sum_a SA_a$$

The **HASA₂ index** is a variant of the HASA index defined as [Katritzky, Lobanov *et al.*, 1998]

$$HASA_2 = \sum_a \sqrt{SA_a}$$

- **RSAA index**

It is the average surface area of H-bond acceptor groups:

$$RSAA = \frac{\sum_a SA_a}{HBA}$$

where HBA is the number of hydrogen-bond acceptors.

- **RSAM index** (\equiv FHASA index)

It is the fraction of the total molecular surface area associated with H-bond acceptor groups:

$$RSAM \equiv FHASA = \frac{\sum_a SA_a}{SASA}$$

Based on the HASA₂ index, the **FHASA₂ index** is defined as [Katritzky, Sild *et al.*, 1998c]

$$FHASA_2 = \frac{\sum_a \sqrt{SA_a}}{SASA}$$

- **HDCA index**

It is the sum of charged surface areas of hydrogens, which can be donated:

$$HDCA = \sum_d q_d \cdot SA_d$$

The charged surface area of hydrogens atoms, called **CSA2_H index**, and the charged surface area of chlorine atoms, called **CSA2_{Cl} index**, are two other similar H-bond descriptors defined as

$$CSA2_H = \sum_h q_h \cdot \sqrt{SA_h} \quad \text{and} \quad CSA2_{Cl} = \sum_{Cl} q_{Cl} \cdot \sqrt{SA_{Cl}}$$

where q_h , q_{Cl} , and SA_h , SA_{Cl} are partial atomic charge and solvent-accessible surface area of hydrogen and chlorine atoms, respectively [Katritzky, Lobanov *et al.*, 1998].

- **FHDCA index**

It is the charged surface area of hydrogens, which can be donated relative to the total molecular surface area:

$$FHDCA = \frac{\sum_d q_d \cdot SA_d}{SASA}$$

- **HDCA₂ index**

It is a hydrogen-bonding descriptor based on solvent-accessible area of hydrogen-bond donor atoms and corresponding partial charges proposed as variant of *FHDCA* index [Katritzky, Mu *et al.*, 1996a]:

$$HDCA_2 = \frac{\sum_d q_d \cdot \sqrt{SA_d}}{\sqrt{SASA}}$$

The summation is performed over the number of simultaneously possible hydrogen bonding donor and acceptor pairs per solute molecule; also hydrogen atoms attached to carbon atoms connected directly to carbonyl or cyano groups are considered as hydrogen bonding donors. The **HDSA₂ index** is another hydrogen-bonding donor descriptor with a definition similar to the *HDCA₂* index [Katritzky, Mu *et al.*, 1996b]:

$$HDSA_2 = \frac{\sum_d q_d \cdot \sqrt{SA_d}}{SASA}$$

where the summation is performed over all possible hydrogen bonding donor sites in a molecule.

- **HACA index** (\equiv *SCAA₁ index*)

It is the sum of charged surface areas of hydrogen-bond acceptors:

$$HACA \equiv SCAA_1 = \sum_a q_a \cdot SA_a$$

An average charged surface area called the **SCAA₂ index** was also calculated as:

$$SCAA_2 = \frac{\sum_a q_a \cdot SA_a}{HBA}$$

where *HBA* is the number of hydrogen-bond acceptors [Turner, Costello *et al.*, 1998; Mitchell and Jurs, 1998b].

- **FHACA index**

It is the charged surface area of hydrogen-bond acceptors relative to the total molecular surface area:

$$FHACA = \frac{\sum_a q_a \cdot SA_a}{SASA}$$

- **HBSA index**

It is the sum of the surface areas of both hydrogens that can be donated and hydrogen acceptor atoms:

$$HBSA = HDSA + HASA$$

- **FHBSA index**

It is the surface area of both hydrogens that can be donated and hydrogen acceptor atoms relative to the total molecular surface area:

$$FHBSA = \frac{HBSA}{SASA}$$

- **HBCA index**

It is the sum of charged surface areas of both hydrogens that can be donated and hydrogen acceptor atoms:

$$HBCA = HDCA + HACA$$

- **FHBCA index**

It is the charged surface area of both hydrogens that can be donated and hydrogen acceptor atoms relative to the total molecular surface area:

$$FHBCA = \frac{HBCA}{SASA}$$

- **CHAA₁ index**

It is the sum of partial charges on hydrogen-bonding acceptor atoms [Mitchell and Jurs, 1998b]:

$$CHAA_1 = \sum_a q_a$$

The average value of CHAA₁ is called the **CHAA₂ index** and is defined as:

$$CHAA_2 = \frac{\sum_a q_a}{HBA}$$

where HBA is the number of hydrogen-bond acceptors.

- **ACGD index**

It is the average difference in charge between all pairs of H-bonding donors.

- **HRPCG index**

It is the relative positive charge (RPCG) restricted to H-bonding donor atoms.

- **HRNCG index**

It is the relative negative charge (RNCG) restricted to H-bonding acceptor atoms.

- **HRPCS index**

It is the relative positive charged surface area (*RPCS*) restricted to H-bonding donor atoms, that is, the positively charged surface area corresponding to the most positively charged atom that is also a possible hydrogen donor.

- **HRNCS index**

It is the relative negative charged surface area (*RNCS*) restricted to H-bonding acceptor atoms, that is, the negatively charged surface area corresponding to the most negatively charged atom that is also a possible hydrogen acceptor.

- **CHGD index**

It is the maximum difference in charge between a hydrogen that can be donated and its covalently-bonded heteroatom.

📖 [Stanton and Jurs, 1992; Nelson and Jurs, 1994; Mitchell and Jurs, 1998a; Eldred, Weikel *et al.*, 1999; Johnson and Jurs, 1999; Schweitzer and Morris, 1999; Katritzky, Maran *et al.*, 2000; De Rienzo, Grant *et al.*, 2002; Eike, Brennecke *et al.*, 2003; Schüürmann, Aptula *et al.*, 2003; Stanton, Mattioni *et al.*, 2004]

- **charge-matching function** → molecular shape analysis
- **charge polarization** → charge descriptors (⊙ total absolute atomic charge)
- **charge-related indices** → charge descriptors
- **charge term matrix** → topological charge indices
- **charge topological index** → charge descriptors (⊙ charge-related indices)
- **charge transfer constant** → electronic substituent constants
- **charge-transfer indices** ≡ *topological charge indices*
- **charge-weighted vertex connectivity indices** → connectivity indices
- **Charton characteristic volume** ≡ *Charton steric constant* → steric descriptors
- **Charton inductive constants** → electronic substituent constants (⊙ inductive electronic constants)
- **Charton steric constant** → steric descriptors
- **Chebyshev distance** ≡ *Lagrange distance* → similarity/diversity (⊙ Table S7)
- **ChemDiverse pharmacophore descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)

■ ChemGPS descriptors

ChemGPS (Chemical Global Positioning System) is a tool that positions novel structures in drug space via PCA-score prediction, providing a unique mapping device for the drug-like chemical space [Oprea and Gottfries, 2001b, 2001a].

Drug space map coordinates are the t-scores extracted via → *Principal Component Analysis*. PCA was performed on a total set of 423 *satellite* and *core* structures described by 72 descriptors representing size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity.

Selected molecules include a set of “satellite” structures and a set of representative drugs (“core” structures). Satellites, intentionally placed outside drug space, have extreme values in one or several of the desired properties, while containing drug-like chemical fragments.

- **chemical adjacency matrix** \equiv *atomic weight-weighted adjacency matrix* \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **chemical atom eccentricity** \rightarrow weighted matrices (\odot weighted distance matrices)
- **CHEMICALC** \rightarrow lipophilicity descriptors (\odot Suzuki–Kudo hydrophobic fragmental constants)
- **chemical descriptors** \rightarrow molecular descriptors
- **chemical distance** \rightarrow bond order indices (\odot conventional bond order)
- **chemical distance degree** \rightarrow weighted matrices (\odot weighted distance matrices)
- **chemical distance matrix** \rightarrow weighted matrices (\odot weighted distance matrices)
- **chemical extended connectivity** \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **chemical filters** \equiv *functional group filters*
- **chemical formula** \rightarrow molecular descriptors
- **chemical graph** \rightarrow graph
- **chemical hardness** \equiv *absolute hardness* \rightarrow quantum-chemical descriptors (\odot hardness indices)
- **chemical invariance** \rightarrow molecular descriptors (\odot invariance properties of molecular descriptors)
- **Chemically Advanced Template Search descriptors** \equiv *CATS descriptors* \rightarrow substructure descriptors (\odot pharmacophore-based descriptors)
- **chemically intuitive molecular index** \rightarrow spectral indices (\odot Burden eigenvalues)
- **Chemical Shift Sum** \rightarrow spectra descriptors
- **chemical space** \rightarrow Structure/Response Correlations
- **chemodescriptors** \rightarrow molecular descriptors

■ chemoinformatics

As Johann Gasteiger [Gasteiger, 2003b] says in his introduction to the *Handbook of Chemoinformatics*, “*Chemoinformatics is the use of informatics methods to solve chemical problems.*”

Gasteiger continues, “*It is clear that chemistry is a scientific discipline that is largely built on experimental observations and data. The amount of data and information accumulated is, however, enormous, and the size of this mountain . . . is increasing with increasing speed. The problem is, then, to extract knowledge from these data and this information, and use this knowledge to make predictions.*”

The term “chemoinformatics” was coined in 1998–1999 and rapidly gained widespread use and, as F. K. Brown says [Brown, 1998], “*Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization.*”

Actually, chemoinformatics is not only related to drug design, but embraces under a unique umbrella all the classical chemical disciplines such as organic chemistry, analytical chemistry, physical chemistry, theoretical chemistry, medicinal chemistry, environmental chemistry, chemometrics, and so on, and more recent fields such as web-chemistry, chemical database managements, and library searching.

We would like also to highlight that in the framework of the chemoinformatics since the past century a fundamental role has been played by disciplines dealing with the mathematical aspects of chemistry such as graph theory, quantum-chemistry, and chemometrics [Balaban, 1978b; Trinajstić and Gutman, 2002]. Also molecular descriptors are of great relevance: they are indeed the basic tool in several chemoinformatics applications such as QSAR/QSPR modeling, drug discovery, similarity/diversity analysis, and library searching.

Some reviews and relevant lectures are reported below for the basic topics involved in chemoinformatics.

Chemoinformatics: [Gasteiger, 2003b, 2006; Gasteiger and Engel, 2003; Leach and Gillet, 2003; Agrafiotis, Bandyopadhyay *et al.*, 2007].

Molecular representation: [Aires-de-Sousa, 2003; Bangov, 2003; Barnard, 2003; Esposito, Hopfinger *et al.*, 2003; Gasteiger, 2003a; Karabunarliev, Nikolova *et al.*, 2003; Rohde, 2003; Sadowski, 2003; Wisniewski, 2003; Xu, 2003].

Molecular descriptors: [Devillers and Balaban, 1999; Karelson, 2000; Todeschini and Consonni, 2000].

Algorithms, multivariate techniques, quality control, and experimental design: [King, Srinivasan *et al.*, 2001; Booth, Isenhour *et al.*, 2003; Eriksson, Antti *et al.*, 2003; Hemmer, 2003; Leardi, 2003; Marsili, 2003; Rose, 2003; Varmuza, 2003; von Homeyer, 2003; Merkwirth, Mauser *et al.*, 2004].

Similarity/diversity analysis: [Farnum, DesJarlais *et al.*, 2003; Willett, 2003a; Maldonado, Doucet *et al.*, 2006].

QSAR and drug design: [Müller, 1997a; Olsson and Oprea, 2001; Xu and Hagler, 2002; Jurs, 2003; Kubinyi, 2003b; Nicklaus, 2003; Oprea, 2003; Selzer, 2003; Steinbeck, 2003; Sottriffer, Stahl *et al.*, 2003; García-Domenech, Gálvez *et al.*, 2008].

Bioinformatics: [Mewes, 2003; Rost, Liu *et al.*, 2003].

Web-chemistry: [Ertl and Jacob, 1997; Ertl, 1998a, 1998b, 2000; Augen, 2002; Ertl and Selzer, 2003; Steinbeck, Han *et al.*, 2003; Tarkhov, 2003; Tetko, Gasteiger *et al.*, 2005].

Database management and retrieval: [Ertl, 2003; Karabunarliev, Nikolova *et al.*, 2003; Neudert and Davies, 2003; Paris, 2003; Voigt, 2003; von Homeyer and Reitz, 2003; Wiggins, 2003; Zass, 2003; Adams and Schubert, 2004; Ósk Jónsdóttir, Jørgensen *et al.*, 2005].

■ chemometrics

Chemometrics is a discipline that deals with mathematical and statistical tools for analysis of complex chemical data [Brereton, 1990; Devillers and Karcher, 1991; Frank and Todeschini, 1994; van de Waterbeemd, 1995; Massart, Vandeginste *et al.*, 1997, 1998; Legendre and Legendre, 1998].

The main characterizing strategies are the multivariate approach to the problem, searching for relevant information, model validation to generate models with predictive power, comparison of the results obtained by using different methods, definition and use of indices capable of measuring the quality of extracted information and the obtained models.

Chemometrics finds a widespread use in QSAR and QSPR studies, in that it provides the basic tools for data analysis and modeling and a battery of different methods. Moreover, a relevant aspect of the chemometric philosophy is the attention it pays to the prediction power of models (estimated by using \rightarrow *validation techniques*), \rightarrow *model complexity*, and the continuous search for suitable parameters to assess the model qualities, such as \rightarrow *classification parameters* and \rightarrow *regression parameters*.

Chemometrics includes several topics of mathematics and statistics; some are listed below in alphabetic order.

• Artificial Neural Networks (ANN)

A set of mathematical methods, models, and algorithms designed to mimic information processing and knowledge acquisition methods of the human brain. ANNs are especially

suitable for dealing with nonlinear relationships and trends and are proposed for facing a large variety of mathematical problems such as data exploration, pattern recognition, modeling of continuous and categorized responses, multiple response problems, etc. [Livingstone, Manallack *et al.*, 1997; Zupan and Gasteiger, 1999; Anzali, Gasteiger *et al.*, 1998a; Niculescu, 2003; Zupan, 2003].

Some historically important artificial neural networks are *Hopfield Networks*, *Perceptron Networks* and *Adaline Networks*, while the most known are *Backpropagation Artificial Neural Networks* (BP-ANN), \rightarrow *Self-Organizing Maps* (SOM), *Counter-Propagation Networks* (CP-ANN), *Radial Basis Function Networks* (RBFN), *Probabilistic Neural Networks* (PNN), *Generalized Regression Neural Networks* (GRNN), *Learning Vector Quantization Networks* (LVQ), and *Adaptive Bidirectional Associative Memory* (ABAM).

📖 Additional references are collected in the thematic bibliography (see Introduction).

• classification

Classification is assignment of objects to one of some classes based on a classification rule. The *classes* are defined *a priori* by groups of objects in a training set belonging to those classes. The goal is to calculate a *classification rule* and, possibly, *class boundaries* based on the training set objects of known classes, and to apply this rule to assign a class to objects of unknown classes [Hand, 1981, 1997; Frank and Friedman, 1989]. Classification methods are suitable for modeling several QSAR responses, such as, for example, active/nonactive compounds, low/medium/high toxic compounds, mutagenic/nonmutagenic compounds.

The most popular classification methods are *Linear Discriminant Analysis* (LDA), *Quadratic Discriminant Analysis* (QDA), *Regularized Discriminant Analysis* (RDA), *Kth Nearest Neighbors* (KNN), *classification tree methods* (such as CART), *Soft-Independent Modeling of Class Analogy* (SIMCA), *potential function classifiers* (PFC), *Nearest Mean Classifier* (NMC), *Weighted Nearest Mean Classifier* (WNMC), *Support Vector Machine* (SVM), and *Classification And Influence Matrix Analysis* (CAIMAN).

Moreover, several classification methods can be found among the artificial neural networks.

Classification model performance is evaluated by \rightarrow *classification parameters*, both for fitting and predictive purposes.

📖 Additional references are collected in the thematic bibliography (see Introduction).

• cluster analysis

A special case of exploratory data analysis aimed at grouping similar objects in the same cluster and less similar objects in different clusters [Massart and Kaufman, 1983; Willett, 1987]. Cluster analysis is based on the evaluation of the \rightarrow *similarity/diversity* of all the pairs of objects of a data set. This information is collected into the \rightarrow *similarity matrix* or \rightarrow *distance matrix*.

Many different methods were designed for cluster analysis; the most popular are the *hierarchical agglomerative methods* (i.e., *average linkage*, *complete linkage*, *single linkage*, *weighted average linkage*, etc.), which are more widely used than the *hierarchical divisive methods*. Other

very popular methods are *nonhierarchical methods*, such as *k-means method* and the *Jarvis–Patrick method*. Among the artificial neural networks dedicated to clustering, → *Self-Organizing Maps* are the most commonly used.

📖 [Dunn III and Wold, 1980; Dean and Callow, 1987; Nakayama, Shigezumi *et al.*, 1988; Willett, 1988; Lawson and Jurs, 1990; Jurs and Lawson, 1991; Good and Kuntz, 1995; Shemetulskis, Dunbar Jr *et al.*, 1995; Brown and Martin, 1996, 1997, 1998; Nouwen, Lindgren *et al.*, 1996; Dunbar Jr, 1997; Junghans and Pretsch, 1997; McGregor and Pallai, 1997; Reynolds, Druker *et al.*, 1998; Rose and Wood, 1998; Reijmers, Wehrens *et al.*, 2001; Rodriguez, Tomas *et al.*, 2005; Stanforth, Kolossov *et al.*, 2007]

• experimental design

Statistical procedures for planning an experiment, that is, collecting appropriate data that, after analysis by statistical methods, result in valid conclusions. The design includes the selection of experimental units, the specification of the experimental conditions, that is, the specification of factors whose effect will be studied on the outcome of the experiment, the specification of the level of the factors involved and the combination of such factors, the selection of response to be measured, and the choice of statistical model to fit the data [Box, Hunter *et al.*, 1978; Carlson, 1992; Livingstone, 1996; Lewis, Mathieu *et al.*, 1999].

An *experiment* consists of recording the values of a set of variables from a measurement process under a given set of experimental conditions.

The most known experimental designs are *complete factorial designs*, *fractional factorial designs*, *Plackett–Burman design*, *Dohelert design*, *composite designs*, and *optimal designs*.

📖 [Borth and McKay, 1985; Bonelli, Cechetti *et al.*, 1991; Pastor and Alvarez-Builla, 1991, 1994; Norinder, 1992; Norinder and Hogberg, 1992; Baroni, Clementi *et al.*, 1993; Baroni, Costantino *et al.*, 1993b; Marsili and Saller, 1993; Cruciani and Clementi, 1994; Rovero, Riganelli *et al.*, 1994; Austel, 1995; Sjöström and Eriksson, 1995; van de Waterbeemd, Costantino *et al.*, 1995; Borth, 1996; Eriksson and Johansson, 1996; Eriksson, Johansson *et al.*, 1997; Giraud, Luttmann *et al.*, 2000; Linusson, Gottfries *et al.*, 2000; Andersson and Lundstedt, 2002; Carro, Campisi *et al.*, 2002; Heimstad and Andersson, 2002; Eriksson, Arnhold *et al.*, 2004; Barroso and Besalú, 2005]

• exploratory data analysis

Exploratory data analysis is a collection of techniques that search for structure in a data set before calculating any statistic model [Krzanowski, 1988]. Its purpose is to obtain information about the data distribution, the presence of outliers and clusters, to disclose relationships and correlations between objects and/or variables. → *Principal component analysis* and *cluster analysis* are the most known techniques for data exploration [Jolliffe, 1986; Jackson, 1991; Basilevsky, 1994].

📖 [Weiner and Weiner, 1973; Stuper and Jurs, 1978; Wold, 1978; Cramer III, 1980a; Henry and Block, 1980a; Streich, Dove *et al.*, 1980; Alunni, Clementi *et al.*, 1983; McCabe, 1984; Takahashi, Miashita *et al.*, 1985; Dunn III and Wold, 1990; Cosentino, Moro *et al.*, 1992;

Livingstone, Evans *et al.*, 1992; Langer and Hoffmann, 1998a; Morais, Ramos *et al.*, 2001; Mazzatorta, Benfenati *et al.*, 2002; Hajduk, Mendoza *et al.*, 2003; Migliavacca, 2003; Restrepo and Villaveces, 2005]

• optimization

This is the procedure that allows to find the optimal value (minimum or maximum) of a numerical function f , called objective function, with respect to a set of parameters \mathbf{p} , $f(p_1, p_2, \dots, p_p)$. If the values that the parameters can take on are constrained, the procedure is called *constrained optimization*.

The most popular optimization techniques are *Newton–Raphson optimization*, *steepest ascent optimization*, *steepest descent optimization*, *Simplex optimization*, *Genetic Algorithm optimization*, and *simulated annealing*. More recent optimization techniques are *Particle swarm optimization* [Cedeño and Agrafiotis, 2003; Tang, Zhou *et al.*, 2007] and *ant colony optimization* [Izrailev and Agrafiotis, 2001a, 2001b]. Moreover \rightarrow *variable reduction* and \rightarrow *variable selection* are also among the optimization techniques.

📖 [Holland, 1975; Papadopoulos and Dean, 1991; Carlson, 1992; Hall, 1995; Kalivas, 1995; Handschuh, Wagener *et al.*, 1998; Sundaram and Venkatasubramanian, 1998; Wehrens, Pretsch *et al.*, 1998]

• ranking methods

Ranking methods are mathematical approaches that provide an ordering of the elements of a system. Ordering is one of the possible ways to analyze data and get an overview over the elements of a system [Pavan and Todeschini, 2008]. Ranking methods are largely used in **Multicriteria decision making** (MCDM) to take decisions about the studied objects (events, molecules, cases, scenarios, etc.) on the basis of more than one criterion [Hendriks, de Boer *et al.*, 1992; Carlson, 1992].

The different kinds of ranking methods available can be roughly classified as total ranking methods and partial-order ranking methods, according to the specific order they provide. These methods are the ones needed to support and solve (a) decision problems, that is, in defining a rank order of the available options, such as different procedures in analytical chemistry, (b) setting priorities, that is, to point out the most dangerous chemicals in a series of compounds, (c) defining global indices, that is, deriving environmental or material global quality indices from their multivariate characterization, and (d) evaluating characteristics of molecular descriptors, that is, analyzing the ranking of a series of compounds with respect to their branching description.

In the ranking techniques, the objects to be ordered can be any kind of objects, such as, for example, available alternatives, scenarios, chemicals, and molecular descriptors.

Specifically, ranking methods may be used to organize chemical information by harmonizing structural information, experimental knowledge, and other specific characteristics of the problem in analysis, such as environmental or health parameters.

Each i th object is represented by a set of p variables f_{ij} , ($j = 1, p$), that, in this framework, are also called criteria. In **total ranking methods**, these variables are joined into a global index, after some scaling procedure or some arbitrary transformation function, eventually using different weights for each criterion, in such a way that the actual value f_{ij} of each i th object for the j th

criterion assumes a value between 0 (worst case) and 1 (optimality). Then, based on the values of the calculated global index, the objects can be ordered as the following:

$$a \geq b \geq c \geq \dots \geq z$$

Simple additive ranking, desirability functions [Harrington, 1965], utility functions, and dominance functions are among the most used total ranking methods.

In **partial-order ranking methods**, a new relationship, which introduces the concept of noncomparability between two objects, is added to the classical ordering relationships. The partial-order ranking methods do not produce a global index useful for a total ranking, but use directly the original variables characterizing each object. The \rightarrow *Hasse diagram* is an effective graphical tool to represent partial ordering. Partial ordering has been used to describe \rightarrow *DNA sequences*.

Examples of applications of partial-order ranking methods and other theoretical aspects are reported in [Walczak and Massart, 1999; Klein and Bytautas, 2000; Carlsen, Sørensen *et al.*, 2001; Carlsen, Lerche *et al.*, 2002; Lerche, Sørensen *et al.*, 2003; Sørensen, Brüggemann *et al.*, 2003; Carlsen, 2004; Pavan and Todeschini, 2004; Voigt, Brüggemann *et al.*, 2004; Carlsen, 2005; Ivanciuc, Ivanciuc *et al.*, 2005; Pavan, Consonni *et al.*, 2005; Randić, Lerš *et al.*, 2005b; Todeschini, Consonni *et al.*, 2006]. Other applications of ranking methods can be found in [Bangov, 1988; Willett, 1988; Eriksson, Jonsson *et al.*, 1990; Ginn, Turner *et al.*, 1997; Randić, 2001e; Balaban, Mills *et al.*, 2002; Gramatica, Pilutti *et al.*, 2002, 2005; Russom, Breton *et al.*, 2003; Wilton and Willett, 2003; Hemmateenejad, 2004, 2005; Pavan, Mauri *et al.*, 2004; Papa, Battaini *et al.*, 2005; Batista and Bajorath, 2007; Gramatica and Papa, 2007; Todeschini, Ballabio *et al.*, 2007; Vogt, Godden *et al.*, 2007].

DART (*Decision Analysis by Ranking Techniques*) is a free available software implementing both partial-order ranking and several total ranking methods [DART – Milano Chemometrics, 2007].

• regression analysis

A set of statistical methods using a mathematical equation to model the relationship between an observed or measured response and one or more predictor variables. The goal of this analysis is twofold: modeling and predicting. The relationship is described in algebraic form as

$$y = f(x) + e \quad \text{or} \quad y = \mathbf{X} \cdot \mathbf{b}$$

where x denotes the predictor variable(s), y the response variable(s), $f(x)$ the systematic part of the model, and e the random error, also called model error or residual; y and \mathbf{b} are the vectors of the responses and regression coefficients to be estimated, respectively; the matrix \mathbf{X} is usually called *model matrix*, that is, its columns are the independent variables used in the regression model.

The mathematical equation used to describe the relationship between response and predictor variables is called *regression model* [Frank and Friedman, 1993; Wold, 1995; Ryan, 1997; Draper and Smith, 1998].

Regression analysis includes not only the estimation of model \rightarrow *regression parameters*, but also the calculation of \rightarrow *goodness of fit* and \rightarrow *goodness of prediction* statistics, *regression diagnostics*, *residual analysis*, and *influence analysis* [Atkinson, 1985].

In particular, the **leverage matrix** \mathbf{H} , also called *influence matrix*, is an important tool in regression diagnostics containing information on the independent variables on which the model is built.

Let \mathbf{X} be a matrix with n rows and p' columns, where p' is the number of model parameters. The leverage matrix \mathbf{H} is a symmetric $n \times n$ matrix defined as

$$\mathbf{H} = \mathbf{X} \cdot (\mathbf{X}^T \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^T$$

where the matrix \mathbf{X} is the model matrix. Moreover, a column where all the values are equal to one is added to the model matrix if the model is not constrained in the origin of the independent variables but an offset is allowed. To distinguish the two cases, a parameter c is used; $c = 1$ for the former and $c = 0$ for the latter.

The main properties of the leverage matrix are

$$\begin{aligned} \text{(a)} \quad & \frac{c}{n} \leq h_{ii} \leq 1 \quad \text{(b)} \quad \sum_{i=1}^n h_{ii} = p' \quad \text{(c)} \quad \bar{h} = \frac{\sum_{i=1}^n h_{ii}}{n} = \frac{p'}{n} \\ \text{(d)} \quad & \sum_{j=1}^n h_{ij} = c \quad \forall i \end{aligned}$$

where \bar{h} is the average value of the leverage.

The leverage matrix is related to the response vector \mathbf{y} by the following relationship:


$$\hat{\mathbf{y}} = \mathbf{H} \cdot \mathbf{y}$$

where $\hat{\mathbf{y}}$ is the calculated response vector from the model.

Usually the diagonal elements h_{ii} of the matrix \mathbf{H} are those used for regression diagnostics: the i th object whose diagonal element h_{ii} is greater than two or three time the average value \bar{h} can be considered as having a great influence (leverage) on the model.

Besides the well-known *Ordinary Least Squares regression (OLS)*, *biased regression*, *nonlinear regression*, and *robust regression* models are also important. The most popular biased methods are *Principal Component Regression (PCR)*, *Partial Least Squares regression (PLS)*, *Ridge Regression (RR)*, *Continuum Regression (CR)*, and *StepWise Regression (SWR)*.

Among the nonlinear methods, there are, besides *nonlinear least squares regression*, that is, *polynomial regression*, the *nonlinear PLS* method, *Alternating Conditional Expectations (ACE)*, *SMART*, and *MARS*. Moreover, some Artificial Neural Networks techniques have been specifically designed for nonlinear regression problems, such as the *back-propagation method*.

 Additional references are collected in the thematic bibliography (see Introduction).

- **CHGD index** → charged partial surface area descriptors
- **CHI chirality descriptor** → chirality descriptors
- **CHI index** \equiv *chromatographic hydrophobicity index* → chromatographic descriptors
- **Chi operator** → connectivity indices
- **chiral A_{xi} indices** → spectral indices (\odot A_{xi} eigenvalue indices)
- **chiral connectivity indices** → chirality descriptors (\odot topological chirality descriptors)
- **chiral factors** → weighted matrices (\odot weighted distance matrices)

- **Chirality Codes** → chirality descriptors
- **chirality correction factor** → chirality descriptors (⊙ topological chirality descriptors)

■ chirality descriptors

A n -dimensional object is called *chiral* if it is nonsuperimposable on its mirror image by any rotation in the n -dimensional space. *Chirality* is the property of chiral objects and was perceived by Lord Kelvin in 1884 [Kelvin, 1904]: “*I call any geometrical figure, or group of points, chiral, and say that it has chirality, if its image in a plane mirror, ideally realized, cannot be brought to coincide with itself.*”

If looked in an isolation, → *physico-chemical properties* (and mathematical) of a chiral molecule and its antipodal counterpart all coincide. However, when chiral structures are considered in an environment, their behavior can be different, such as it occurs, for example, when a chiral molecule interacts with a receptor. Thus, chirality descriptors are useful for modeling properties related to interactions involving chiral centers [Aires-de-Sousa, 2003].

Two general classes of chirality measures have been recognized: in the first, the degree of chirality expresses the extent to which a chiral object differs from an achiral reference object, while in the second it expresses the extent to which two enantiomorphs differ from each other [Buda, Auf der Heyde *et al.*, 1992].

Chirality measures of the first class are the **Ruch's chirality functions** [Ruch, 1972], according to which a chiral molecule is represented by an achiral skeleton with attached four “ligands”, for example, chemical groups a , b , c , and d , each of them characterized by a specific parameter λ . Polynomial functions, such as

$$F = (\lambda_a - \lambda_b) \cdot (\lambda_a - \lambda_c) \cdot (\lambda_a - \lambda_d) \cdot (\lambda_b - \lambda_c) \cdot (\lambda_b - \lambda_d) \cdot (\lambda_c - \lambda_d)$$

transform these parameters into a chirality measure, being, for two enantiomers R and S, $F(R) = -F(S)$. A specific application of this approach was proposed by Lukovits and Linert [Lukovits and Linert, 2001] using the → *valence connectivity index* $^1\chi^v$ (→ *Chi chirality descriptor*) to describe chirality.

The **Hausdorff chirality measure** is a chirality measure of the second class [Buda and Mislow, 1992]. Let Q and Q' denote two enantiomorphous, nonempty, and bounded sets of points defined in the geometrical space (x, y, z) . Let $d(q, q')$ denotes the distance between two points: $q \in Q$ and $q' \in Q'$. Then, the Hausdorff distance h between sets Q and Q' is defined as

$$h(Q, Q') = h(Q', Q) = \max[\rho(Q, Q'), \rho(Q', Q)]$$

where ρ is defined as:

$$\rho(Q, Q') = \max_{q \in Q} \{ \min_{q' \in Q'} (d_{qq'}) \} \quad \rho(Q', Q) = \max_{q' \in Q'} \{ \min_{q \in Q} (d_{q'q}) \}$$

The Hausdorff distance $h(Q, Q')$ between Q and Q' corresponds to the smallest number $\delta = h(Q, Q')$ that has the following properties: (a) a spherical ball of radius δ centered at any point of Q contains at least one point of Q' and (b) a spherical ball of radius δ centered at any point of Q' contains at least one point of Q . It is obvious that $h(Q, Q') = 0$ only if $Q = Q'$.

This means that the Hausdorff distance between two sets of points, Q and Q' , representing geometric objects, can be zero only if these two objects are identical, that is, achiral mirror images.

The value of the Hausdorff distance between a geometric object Q and its mirror image Q' depends not only on the shape of these objects but also on their size and their relative

orientations in the geometrical space. By rotating and translating one enantiomorph with respect to the other, one can find the minimal value $h_{\min}(Q, Q')$ corresponding to the optimal overlap.

The Hausdorff chirality measure is finally defined as

$$H(Q) = \frac{h_{\min}(Q, Q')}{d_{\max}(Q)}$$

where $d_{\max}(Q)$ denotes the diameter of Q , that is, the largest distance between any two points of Q . The Hausdorff chirality measure does not depend on the size of Q and Q' and their relative position, and it can be easily shown that it has all the attributes required for a degree of chirality.

The interest in using chirality descriptors in QSAR/QSPR modeling is increasing and hence some chirality descriptors are explained below. \rightarrow *Schultz weighted distance matrices* have been also devised for obtaining the \rightarrow *chiral modification number* that is added to any topological index to discriminate cis/trans isomers. Other interesting methods to quantify chirality are the Kuz'min's method based on the *dissymmetry functions* [Kuz'min, Stel'makh *et al.*, 1992a, 1992b; Kutulya, Kuz'min *et al.*, 1992], the Mezey's method [Mezey, 1997b] based on the \rightarrow *Mezey 3D shape analysis*, \rightarrow *MARCH-INSIDE descriptors*, \rightarrow *chiral TOMOCOMD descriptors*, and the spectral \rightarrow *chiral A_{xi} indices* [Xu, Zhang *et al.*, 2006].

• Seri-Levy chirality coefficients

These are chirality descriptors based on the maximum overlapping of the van der Waals volumes of the two enantiomers R and S in 3D space, which are embedded into a gridded box [Seri-Levy, Salter *et al.*, 1994; Seri-Levy, West *et al.*, 1994]. A shape similarity coefficient S_{RS} is calculated as

$$S_{RS} = \frac{n_{RS}}{\sqrt{n_R \cdot n_S}}$$

where n_{RS} is the number of grid points falling inside both enantiomers, n_R and n_S are the number of grid points included within the volume of enantiomers R and S , respectively.

From the shape similarity coefficient, a shape chirality coefficient was defined as:

$$S'_{RS} = 1 - S_{RS}$$

• Continuous Chirality Measure (CCM)

The continuous chirality measure is an example of first class chirality measure based on the general definition of *continuous symmetry measure*; it is defined as [Zabrodsky and Avnir, 1995]

$$S(G) = \frac{100}{A} \cdot \sum_{i=1}^A \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|^2 \quad 0 < S \leq 100$$

where G is a given symmetry group, \mathbf{p}_i the coordinate vector of the i th atom of the original chiral configuration, $\hat{\mathbf{p}}_i$ the coordinate vector of the corresponding atom in the nearest G -symmetric configuration, and A the number of atoms of the molecule. In practice, since the minimal requirement for an object to be achiral is that it possesses either a reflection mirror (σ), an inversion center (i), or a higher order improper rotation axis (S_{2n}), the function $S(G)$ has to be screened over symmetry groups having these elements. The continuous chirality measure is the total (normalized) distance of the original chiral configuration from the considered G -symmetry configuration, bounded between 0 and 100.

• Randić chirality index

This is a topological index that was proposed to discriminate between a chiral molecule and its mirror image; it is restricted to molecules embedded in 2D space, such as benzenoids, and is based on \rightarrow *periphery codes* [Randić, 1998a, 2001a].

Starting from a selected i th atom on the molecule periphery, the \rightarrow *vertex degrees* δ of the periphery atoms define an ordered code for the molecule; two different codes are obtained whether the clockwise or the anticlockwise direction is chosen. These codes referring to the i th atom are then transformed by making partial sums, adding successively elements of the series; the two obtained codes encode information on the asymmetry of the molecular periphery.

To obtain a single value descriptor D_i for the i th atom, the differences between the corresponding elements in the anticlockwise and clockwise codes are computed and then added:

$$D_i = \sum_{j=1}^A (AD_{ij} - CD_{ij})$$

where AD_{ij} represents the j th element in the i th anticlockwise code and CD_{ij} represents the j th element in the i th clockwise code. This procedure is repeated for all the atoms on the molecule periphery. Finally, the Randić chirality index is defined as

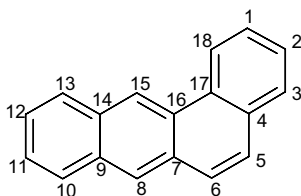
$$CH_R^k = \frac{1}{A^k} \cdot \sum_{i=1}^A D_i^k$$

where k is an odd power exponent and A is the number of atoms. Using different odd powers, a sequence of chirality indices can be calculated to better characterize the molecule and its mirror image.

For achiral molecules, all the chirality indices (and the corresponding vector elements) equal zero.

Example C8

Randić chirality index for benzoanthracene.



clockwise direction

labels	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
atom 1	2	2	2	3	2	2	3	2	3	2	2	2	2	3	2	3	3	2

anticlockwise direction																		
labels	1	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
atom 1	2	2	3	3	2	3	2	2	2	2	3	2	3	2	2	3	2	2

↓

clockwise direction CD_{1j}

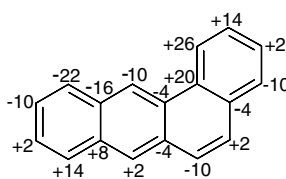
atom 1	2	4	6	9	11	13	16	18	21	23	25	27	29	32	34	37	40	42
--------	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----

anticlockwise direction AD_{1j}

atom 1	2	4	7	10	12	15	17	19	21	23	26	28	31	33	35	38	40	42
--------	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

$$D_1 = \sum_{j=1}^{18} (AD_j - CD_j) = +14$$

↓ repeating the procedure for all the atoms



Randic chirality index CH_R^k					
k	3	5	7	9	11
benzanthracene	+ 2.17284	+ 5.07240	+ 10.97954	+ 23.70935	+ 51.05061
mirror image	- 2.17284	- 5.07240	- 10.97954	- 23.70935	- 51.05061

• Moreau chirality index

This is a molecular descriptor proposed with the aim of quantifying the chirality of a molecule by means of the positive or negative quantification of the chirality of the environment of the atoms in the molecule for any scalar atomic property [Moreau, 1997]. The basic idea is that unsymmetrical environments are not the privilege of atoms in chiral molecules, however they are the most frequent situations. The Moreau chirality index is defined as

$$CH_M = \sum_{i=1}^A AS_i$$

where the summation runs over all atoms in the molecule and AS_i is the measure of the chirality of the environment of the i th atom given by the following expression:

$$AS_i = 10^3 \cdot C_i \cdot \{XYZ\}_i \cdot S_i = 10^3 \cdot \frac{(\lambda_1 - \lambda_2) \cdot (\lambda_2 - \lambda_3) \cdot \lambda_3}{(\sum_k \lambda_k)^3} \cdot \frac{X_i \cdot Y_i \cdot Z_i}{(e_X \cdot e_Y \cdot e_Z)} \cdot (v_1, v_2, v_3)$$

where λ_1 , λ_2 , and λ_3 are the eigenvalues of the square symmetric matrix \mathbf{C} closely related to the covariance matrix of the Cartesian coordinates (x_j, y_j, z_j) of the atoms in the environment of the i th atom and defined as

$$\mathbf{C} = \begin{vmatrix} \sum_j p_j \cdot w_j \cdot x_j^2 & \sum_j p_j \cdot w_j \cdot x_j \cdot y_j & \sum_j p_j \cdot w_j \cdot x_j \cdot z_j \\ \sum_j p_j \cdot w_j \cdot x_j \cdot y_j & \sum_j p_j \cdot w_j \cdot y_j^2 & \sum_j p_j \cdot w_j \cdot y_j \cdot z_j \\ \sum_j p_j \cdot w_j \cdot x_j \cdot z_j & \sum_j p_j \cdot w_j \cdot y_j \cdot z_j & \sum_j p_j \cdot w_j \cdot z_j^2 \end{vmatrix}$$

where summations are over all atoms in the environment of i and p_j and w_j are parameters defined below.

The origin of the coordinates is the barycenter of the environment of the considered atom and the eigenvectors \mathbf{v}_1 , \mathbf{v}_2 , \mathbf{v}_3 associated with the eigenvalues are vectors that define the three principal axes of this environment. The considered i th atom can be included or not in its environment. Each atom of the environment is assigned an atomic property p_j (e.g., unitary property, atomic mass, atomic electronegativity, and atomic van der Waals volume) and a weight w_j , which is a function of the distance of the j th environment atom from the i th atom.

In this approach, the principal planes of the environment are taken as the best approximation of the potential symmetry planes and the asymmetry of the environment as seen from the considered i th atom is defined as proportional to the distance from i to the symmetry plane.

The coefficient C_i in the expression of AS_i accounts for the asymmetry of the environment and is equal to zero when two eigenvalues are equal, or the third eigenvalue equals zero, or the last two eigenvalues equal zero.

The term $\{XYZ\}_i$ is the product of the coordinates of the i th atom with respect to the principal axes normalized by the product of the half-thicknesses e_x, e_y, e_z of the “slab” which approximates the set of weighted environment atoms. This term indicates that asymmetry is positive or negative according to the octant in which the i th atom is.

The term S is the box-product of the three eigenvectors and is equal to $+1$ or -1 depending on the handedness of the principal axis space. It was introduced in the expression of AS to have an intrinsic measure of chirality.

• topological chirality descriptors

The idea of modifying graph invariants to make them chirality sensitive was proposed by Schultz *et al.* in 1992 [Schultz, Schultz *et al.*, 1995], introducing a \rightarrow *chiral factor* equal to $+1$ or -1 for any atom in R- or S-configuration, respectively, and assigning a value of 0 to all the other atoms.

Developing this idea, several series of topological chirality descriptors were introduced by using a **chirality correction factor**, denoted as c , applied to the \rightarrow *vertex degree* of chiral atoms in a \rightarrow *H-depleted molecular graph* [Golbraikh, Bonchev *et al.*, 2001a, 2001b; Golbraikh and Tropsha, 2003]. These descriptors include modified \rightarrow *Zagreb indices*, \rightarrow *connectivity indices*, \rightarrow *extended connectivity indices*, \rightarrow *overall connectivity indices*, and \rightarrow *topological charge indices*.

For each asymmetric atom in R-configuration, the vertex degree δ_i is substituted with $(\delta_i + c)$ and for each atom in S-configuration with $(\delta_i - c)$. This transformation is equivalent to making main diagonal elements a_{ii} of the \rightarrow *adjacency matrix* \mathbf{A} equal to $+c$ or $-c$ for all chiral atoms in R- or S-configuration, respectively. For achiral atom, the chirality correction factor equals zero.

According to this approach, the \rightarrow Randić connectivity index can be transformed into the corresponding chirality index as

$${}^1\chi = \sum_b (\delta_i \cdot \delta_j)_b^{-1/2} \Rightarrow {}^1\chi^{chir} = \sum_b [(\delta_i \pm c) \cdot (\delta_j \pm c)]_b^{-1/2}$$

where summation goes over all edges in the molecular graph and i and j refer to the vertices, which are connected by an edge.

In general, values of $|c| < \delta_i$ were assumed. The chirality correction was also applied to \rightarrow valence vertex degrees from which valence connectivity indices are derived.

Chirality correction can be a real number (chirality descriptors of class I) or an imaginary number (chirality descriptors of class II). In the latter case, chirality descriptors are complex numbers.

The **chiral connectivity indices** were calculated by analogy with the \rightarrow Kier–Hall connectivity indices by using vertex degrees modified by a chirality correction $c = \pm 1$ [Xu, Zhang *et al.*, 2006].

• Chi chirality descriptor (χ^c)

This chirality descriptor is derived from the Ruch's chirality functions applied to the first-order \rightarrow valence connectivity index ${}^1\chi^v$. Separate values of the valence connectivity index are calculated for the four atoms/substituents a , b , c , and d bonded to the chiral atom [Lukovits and Linert, 2001]. The chirality correction χ^c is calculated by the following function F:

$$F \equiv \chi^c = ({}^1\chi_a^v - {}^1\chi_b^v) \cdot ({}^1\chi_a^v - {}^1\chi_c^v) \cdot ({}^1\chi_a^v - {}^1\chi_d^v) \cdot ({}^1\chi_b^v - {}^1\chi_c^v) \cdot ({}^1\chi_b^v - {}^1\chi_d^v) \cdot ({}^1\chi_c^v - {}^1\chi_d^v)$$

that, for two enantiomers R and S, has the following property: $F(R) = -F(S)$.

The chirality descriptor χ^- is then defined as

$$\chi^- = {}^1\chi^v \pm \chi^c$$

where ${}^1\chi^v$ is the valence connectivity index for the whole molecule and χ^c the chirality correction.

A drawback of this index is that index ${}^1\chi^v$ is often degenerate leading to the same value for different substituent groups (e.g., halogens), resulting into a zero correction for chirality. This drawback may be overcome using more discriminant vertex degrees, such as \rightarrow perturbation delta values, \rightarrow Yang vertex degree, or the δ^{CT} proposed by the Authors.

${}^1\chi^v$ values for some substituents are reported in \rightarrow connectivity descriptors (Table C6).

• Chirality Codes (f_{CICC} , f_{CDCC})

The conformational-independent chirality code f_{CICC} is a modification of the \rightarrow radial distribution function code $g(R)$ to account for chirality:

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A w_i \cdot w_j \cdot e^{-\beta \cdot (R-r_{ij})^2} \Rightarrow f_{CICC}(R) = \sum_{i=1}^A \sum_{j=1}^{A-1} \sum_{k=1}^{A-2} \sum_{l=1}^{A-3} S_{ijkl} \cdot e^{-\beta \cdot (R-E_{ijkl})^2}$$

where β is a smoothing factor and A the number of atoms [Aires-de-Sousa and Gasteiger, 2001, 2002; Aires-de-Sousa, 2003]. The term E_{ijkl} was introduced to account for the stereochemical situation of the chiral center; this term considers atoms i , j , k , and l , each of them belonging to a different neighborhood of the four atoms A, B, C, and D that are directly bonded to the chiral

center. This term is defined as

$$E_{ijkl} = \frac{w_i \cdot w_j}{r_{ij}} + \frac{w_i \cdot w_k}{r_{ik}} + \frac{w_i \cdot w_l}{r_{il}} + \frac{w_j \cdot w_k}{r_{jk}} + \frac{w_j \cdot w_l}{r_{jl}} + \frac{w_k \cdot w_l}{r_{kl}}$$

where w is an atomic property and r the distance calculated as the sum of the geometric distances along the shortest path joining two atoms. Further, the chirality signal S_{ijkl} can attain values $+1$ or -1 . For the computation of S_{ijkl} , atoms i, j, k , and l are ranked according to the decreasing atomic property w . When the property of two atoms is the same, the properties of the neighbors (A, B, C, or D) are used for ranking. The (x, y, z) coordinates of A are then used for atom i , those of B for j , those of C for k , and those of D for l . The first three atoms, in the order defined by the ranking define a plane. If they are ordered clockwise and the fourth atom is behind the plane, the chirality signal is set at $+1$; if the geometric arrangement is opposite, $S_{ijkl} = -1$.

The two values, E and S , calculated for all the combinations of the four atoms are then combined to generate the chirality code $f_{\text{ClCC}}(R)$, where the function is calculated at a number of discrete points R with defined intervals to obtain the same number of descriptors, irrespective of the size of the molecule. The actual range of R is chosen according to the range of the studied properties related to the range of observed interatomic distances for the data set molecules.

The number of discrete points determines the resolution of the chirality code.

Moreover, the **Conformational-Dependent Chirality Code** (f_{CDCC}) was defined to account for conformational behavior of molecules, replacing the chirality signal S_{ijkl} by a conformational-dependent geometric parameter C_{ijkl} [Caetano, Aires-de-Sousa *et al.*, 2005]. The f_{CDCC} code is calculated as

$$f_{\text{CDCC}}(R) = \sum_{i=1}^A \sum_{j=1}^{A-1} \sum_{k=1}^{A-2} \sum_{l=1}^{A-3} C_{ijkl} \cdot e^{-\beta \cdot (R - E_{ijkl})^2}$$

where A is the number atoms. The parameter C_{ijkl} takes real values and is defined according to the expression:

$$C_{ijkl} = \frac{x_j \cdot y_k \cdot z_l}{x_j \cdot y_k + x_j \cdot |z_l| + y_k \cdot |z_l|}$$

that is, C_{ijkl} is defined for the combination of four atoms i, j, k , and l and x, y and z are the atomic Cartesian coordinates. The Cartesian coordinates are defined in such a way that atom i is at position $(0, 0, 0)$, atom j lies on the positive side of the x -axis, and atom k on the xy plane and having a positive y coordinate. Therefore, C_{ijkl} will have opposite values for enantiomers, because C_{ijkl} will have either a positive or negative value depending on whether atom l is above ($z_l > 0$) or below ($z_l < 0$) the plane formed by atoms i, j , and k .

• Ursu–Diudea chirality (χ_{1234})

The chirality of an ordered quadruple of atoms numbered 1,2,3,4 is measured in terms of their (x, y, z) Cartesian coordinates, adopting some geometrical constraints, by the sign of the following determinant [Ursu and Diudea, 2005; Ursu, Diudea *et al.*, 2006]:

$$\chi_{1234} = \text{sgn} \left(\det \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{pmatrix} \right)$$

• **Relative Chirality Index (^VRCI)**

The Relative Chirality Index is calculated from a series expansion of the \rightarrow *valence vertex degree*, calculated for the four atoms/groups attached to the chiral carbon, where atom/groups priorities (a, b, c, d) are given according to the Cahn–Ingold–Prelog rule [Natarajan, Basak *et al.*, 2007] (Figure C2).

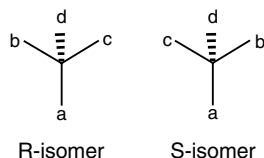


Figure C2 R- and S-configurations.

To calculate the relative chirality index, the least important chemical group (d) is placed at the rear, and the clockwise and anticlockwise arrangement of the other three groups (a, b, c) are used to represent the R- and S-configurations, respectively.

The groups/atoms a, b, c , and d are then assigned \rightarrow *valence vertex degrees* (δ^v). When the group has more than one atom, δ^v for the group a, b , or c is calculated considering the relative proximities of the atoms to the chiral neighbor, and decreasing importance with increasing topological distance was assigned while calculating the contribution of atoms other than hydrogen in a group. The group delta value for any group attached to a chiral carbon is, then, calculated as

$$\delta_i^v = \delta_{n1}^v + \frac{\delta_{n2}^v}{2} + \frac{\delta_{n3}^v}{4} + \frac{\delta_{n4}^v}{8} + \dots$$

where $n1$ is the atom attached directly to the chiral center (nearest neighbor), $n2$ is separated by one atom, $n3$ by two atoms, etc. Relative chirality indices (^VRCI) for a pair of enantiomers (R and S) are calculated as:

$$^V\text{RCI}_R = \delta_a^v \cdot [1 + (1 + \delta_b^v) + (1 + \delta_b^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$^V\text{RCI}_S = \delta_a^v \cdot [1 + (1 + \delta_c^v) + (1 + \delta_c^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

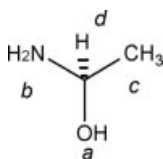
To obtain ^VRCI for molecules containing more than one chiral center, root-mean-square of ^VRCI for all the chiral atoms is calculated as

$$^V\text{RCI} = \sqrt{\frac{1}{n_{\text{chi}}} \sum_{i=1}^{n_{\text{chi}}} (^V\text{RCI}_i)^2}$$

where n_{chi} is the number of chiral centers.

Example C9

Relative chirality indices and chirality correction factors for the two enantiomers shown below, together with valence vertex degrees and first-order valence connectivity index of the functional groups.



$$\begin{aligned}\delta^v(\text{OH}) &= 5 & {}^1\chi^v(\text{OH}) &= 0.25820 \\ \delta^v(\text{NH}_2) &= 3 & {}^1\chi^v(\text{NH}_2) &= 0.33333 \\ \delta^v(\text{CH}_3) &= 1 & {}^1\chi^v(\text{CH}_3) &= 0.57735 \\ \delta^v(\text{H}) &= 0 & {}^1\chi^v(\text{H}) &= 0\end{aligned}$$

$${}^v\text{RCI}_R = \delta_a^v \cdot [1 + (1 + \delta_b^v) + (1 + \delta_b^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$= 5 \cdot [1 + (1 + 3) + (1 + 3 + 3 \cdot 1) + 0] = 60$$

$${}^v\text{RCI}_S = \delta_a^v \cdot [1 + (1 + \delta_c^v) + (1 + \delta_c^v + \delta_b^v \cdot \delta_c^v) + (\delta_b^v \cdot \delta_c^v \cdot \delta_d^v)]$$

$$= 5[1 + (1 + 1) + (1 + 1 + 1 \cdot 3) + 0] = 40$$

The chirality correction factor χ^c for the calculation of the \rightarrow *Chi chirality index* is 0.00029.

📖 [King, 1991; Gilat, 1994; Liang and Mislow, 1994; Flapan, 1995; Franke, Rose *et al.*, 1995; Winberg and Mislow, 1995; De Julián-Ortiz, García-Domenech *et al.*, 1996; Fujita, 1996; Petitjean, 1996; Randić and Mezey, 1996; Randić and Razinger, 1996; Balaban, 1997b; Gutman and Pyka, 1997; Klein and Babic, 1997; Mislow, 1997; De Julián-Ortiz, de Gregorio Alapont *et al.*, 1998; Keinan and Avnir, 1998; Nemba and Balaban, 1998; Randić, 1998a; Golbraikh, Bonchev *et al.*, 2002; González Díaz, Sánchez *et al.*, 2003; Wildman and Crippen, 2003; Kovatcheva, Golbraikh *et al.*, 2004, 2005; Marrero-Ponce, González Díaz *et al.*, 2004; Marrero-Ponce and Castillo-Garit, 2005]

- **chiral modification number** \rightarrow weighted matrices (\odot weighted distance matrices)
- **chiral TOMOCOMD descriptors** \rightarrow TOMOCOMD descriptors
- **CHI-square statistics** \rightarrow statistical indices (\odot concentration indices)
- **chord distance** \rightarrow similarity/diversity (\odot Table S7)

■ chromatic decomposition

A decomposition $\mathcal{A}(V_1, V_2, \dots, V_G)$ of the set V of the graph G vertices into $G \rightarrow$ *equivalence classes* is said **chromatic decomposition** of G (or **vertex chromatic decomposition**) if, for any pair of vertices v_i and v_j belonging to V_g , the edge e_{ij} connecting the considered vertices does not belong to the set of edges E of the graph; it means that two vertices belonging to the same chromatic class V_g cannot be adjacent [Bonchev, 1983].

A decomposition $\mathcal{B}(E_1, E_2, \dots, E_G)$ of the set E of the graph G edges into G equivalence classes is said **edge chromatic decomposition** of G if any pair of edges e_{ij} and e_{kl} belonging to E_g does not belong to the set ${}^2\mathcal{P}$ of the second-order paths of the graph (i.e., the two edges are not adjacent).

A **graph coloring** of vertices (or edges) is an assignment of a minimal number of different colors to the vertices (or edges) of G such that no two adjacent vertices (or edges) have the same color. Graph coloring produces a \rightarrow *chromatic graph*.

The subsets V_g are called **color classes**. The simplest descriptor that can be defined by a vertex chromatic decomposition is called **chromatic number** $k(G)$ (or **vertex chromatic number**, ${}^V k(G)$) and is the smallest number of color equivalence classes (i.e., G). In general, there is not a unique chromatic decomposition of a graph with the smallest number of colors. Analogously, the descriptor obtained by an edge chromatic decomposition is called **edge chromatic number**, denoted as ${}^E k(G)$.

The **chromatic information index** (or **vertex chromatic information index**) [Mowshowitz, 1968d] is the minimum value of the \rightarrow *mean information content* of all possible vertex chromatic decompositions with a number of colors equal to the vertex chromatic number $k(G)$ and is defined as

$${}^V \bar{I}_{CHR} = \min \left(- \sum_{g=1}^{k(G)} \frac{|V_g|}{A} \log_2 \frac{|V_g|}{A} \right)$$

where $|V_g|$ is the number of vertices (i.e., the cardinality of g th set) within the same equivalence class for the decomposition and A is the number of graph vertices.

The **edge chromatic information index** is the minimum value of the mean information content of all possible edge chromatic decompositions having a number of colors equal to the edge chromatic number ${}^E k(G)$ and is defined as

$${}^E \bar{I}_{CHR} = \min \left(- \sum_{g=1}{{}^E k(G)} \frac{|E_g|}{B} \log_2 \frac{|E_g|}{B} \right)$$

where $|E_g|$ is the number of edges within the same equivalence class for the decomposition and B is the number of graph edges.

- **chromatic graph** \rightarrow graph
- **chromatic information index** \rightarrow chromatic decomposition
- **chromatic number** \rightarrow chromatic decomposition

■ chromatographic descriptors

These are experimental quantities derived from chromatographic techniques, that is, from gas chromatography (GC), high-performance liquid chromatography (HPLC), thin-layer chromatography (TLC), and paper chromatography (PC) [Kaliszan, 1987, 1992] or structural indices used to predict experimental chromatographic parameters from molecular structure.

The most important ones are listed below.

• retention time (t_R)

It is the characteristic time it takes to a compound to pass through a chromatographic system (e.g., from the column inlet to the detector) under fixed conditions. The **adjusted retention time** of a compound, denoted by t'_R , is the difference between the total retention time t_R and the retention time of an unretained compound t_M ; the **relative retention time**, denoted as RRT , is the ratio of the adjusted retention time of a compound over that of a reference compound.

Based on the retention times in HPLC systems, a **chromatographic hydrophobicity index (CHI)** was defined aimed at correlating chromatographic retention times with lipophilicity. This is defined as [Valkó, Bevan *et al.*, 1997; Valkó, Plass *et al.*, 1998]

$$CHI = a \cdot t_R + b$$

where the parameters a and b are dependent on the flow rate, column length, gradient time, column, etc. and are experimentally determined; they allow to transfer different chromatographic measurements into a unique scale. Another index derived from HPLC is the **ϕ_0 index**, which is defined as the percentage (by volume) of acetonitrile required to achieve an equal distribution of a compound between the mobile and stationary phases. For most compounds this is a physically attainable volume percent of organic phase with a value between 0 and 100%. By plotting the $\log k'$ values, k' being the \rightarrow *capacity factor*, as a function of the organic solvent concentration, the ϕ_0 value can be obtained from the slope and the intercept of the straight interpolation line as $\phi_0 = -b/a$, where a and b are the slope and intercept of the straight line, respectively.

- **capacity factor (k')**

Also called **phase capacity ratio** or **retention factor**, it is a measure of the degree of retention of a compound in a chromatographic column, as

$$k' = \frac{t_R - t_M}{t_M} = \frac{V_R - V_M}{V_M}$$

where t_R and V_R are the retention time and retention volume, respectively, of the compound; t_M and V_M , named dead time and dead volume, are respectively the retention time and the retention volume of an unretained compound. The quantity $\log k'$ can be considered analogous to the \rightarrow *Bate-Smith-Westall retention index* R_M (see below) and, like this, is related to \rightarrow *partition coefficients*.

In micellar electrokinetic chromatography and microemulsion electrokinetic chromatography, the retention factor k' of a neutral compound is defined as [Muijselaar, Claessens *et al.*, 1994]

$$k' = \frac{t_R - t_M}{t_M} \left(1 - \frac{t_R}{t_m} \right)^{-1}$$

where t_M , t_R , and t_m are the migration times of electroosmotic flow, compound, and microemulsion, respectively.

From this retention factor, the **Migration Index (MI)** for a compound in microemulsion electrokinetic chromatography was defined as [Ishihama, Oda *et al.*, 1996]

$$MI = a \cdot \log \left(\frac{\mu_{aq} - \mu_{eff}}{\mu_{eff} - \mu_{me}} \right) + b = a \cdot \log k' + b$$

where μ_{aq} and μ_{me} are the electrophoretic mobilities of the compound in the aqueous phase and microemulsion phase, respectively, μ_{eff} the effective mobility in the microemulsion solution, and a and b the slope and the intercept of a calibration line relative to $\log k'$ values of reference solutes such as alkyl benzene and their migration indices.

The Migration Index scale can be applied to all neutral compounds that migrate in the range t_M and t_m and this might be independent of the volume of the microemulsion. The Migration

Index scale can be used as a measure of hydrophobicity of compounds for QSAR/QSPR modeling studies [Ishihama, Oda *et al.*, 1996; Fatemi, 2003].

- **Kovats retention index (I_i)**

This is an index characteristic of a gas-chromatographed compound on a given column at a definite temperature defined as

$$I_i = 100 \cdot \frac{\log t'_{R_i} - \log t'_{R(N_C)}}{\log t'_{R(N_C+1)} - \log t'_{R(N_C)}} + 100 \cdot N_C$$

where $t'_{R(N_C)}$ is the \rightarrow *adjusted retention time* of a homologue standard with a number of carbon atoms equal to N_C ; $t'_{R(N_C+1)}$ an analogous parameter for another standard with carbon number $N_C + 1$; t'_{R_i} the adjusted retention time of the i th compound [Kováts, 1968]. The measured total retention time t_R is a sum of two factors $t_R = t_M + t'_R$, where t_M is the initial dead time and t'_R is the adjusted retention time of the compound.

- **Bate-Smith-Westall retention index (R_M)**

This is a retention index derived from thin-layer and paper chromatography defined as [Bate-Smith and Westall, 1950]

$$R_M = \log \left(\frac{1}{R_f} - 1 \right)$$

where R_f is the **retardation factor**, which is the ratio of the migration distance of the compound over that of the solvent front.

The quantity R_M is proportional to the partition coefficient $\rightarrow \log P$ and has been used in its place in many QSAR models.

- **semiempirical topological index (I_{ET})**

The semiempirical topological index (the name is not the best choice) was designed to predict the chromatographic \rightarrow *Kovats retention index* of organic molecules, based on experimental chromatographic measurements [Heinzen, Soares *et al.*, 1999; da Silva Junkes, Amboni *et al.*, 2003b].

This is derived from the \rightarrow *H-depleted molecular graph* of a molecule as

$$I_{ET} = \sum_{i=1}^A (C_i + \delta_i)$$

where A is the number of graph vertices and C_i is the carbon atom contribution of the i th molecular fragment (Table C4); δ_i is the contribution of the carbon atoms bonded to the i th carbon atom, calculated as

$$\delta_i = \sum_{j=1}^A a_{ij} \cdot \log(C_j)$$

where the summation goes over all graph vertices but only contributions from vertices adjacent to the i th vertex are accounted for, a_{ij} being the elements of the \rightarrow *adjacency matrix A*.

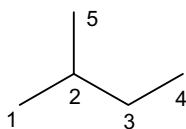
In this approach, the molecular fragment including the functional group and the carbon atom directly attached to the functional group are considered as a single vertex in the molecular graph.

Table C4 Values of the fragmental constant C_i for carbon atoms and functional groups of esters, aldehydes, ketones, and alcohols.

Chemical class	Fragment	Fragment position	C_i
Linear and branched alkanes	$-\text{CH}_3$		1.0
	$-\text{CH}_2-$		0.9
	$-\text{CH}<$		0.8
	$>\text{C}<$		0.7
Alkenes	$=\text{CH}_2; =\text{CH}-$	1C	0.8975
	$=\text{CH}-\text{trans}$	2C	0.895
	$=\text{CH}-\text{cis}$	2C	0.910
	$=\text{CH}-\text{trans}$	3C	0.875
	$=\text{CH}-\text{cis}$	3C	0.885
	$=\text{CH}-\text{trans}$	4C	0.865
	$=\text{CH}-\text{cis}$	4C	0.870
	$=\text{CH}-\text{trans}$	5C	0.865
	$=\text{CH}-\text{cis}$	5C	0.855
	$=\text{CH}-\text{trans}$	6C	0.860
	$=\text{CH}-\text{cis}$	6C	0.850
	$=\text{CH}-\text{trans}$	7C	0.8575
	$=\text{CH}-\text{cis}$	7C	0.845
Alcohols	$-\text{CH}_2-\text{OH}$		2.63
	$>\text{CH}-\text{OH}$	2nd position	1.79
	$>\text{CH}-\text{OH}$	3rd position	1.78
	$>\text{CH}-\text{OH}$	Middle	1.68
	$-\text{CH}<$	α OH	0.75
	$-\text{CH}<$	β OH	0.73
	$>\text{C}<$	α OH	0.61
	$>\text{C}<$	β OH	0.63
Aldehydes and ketones	$\text{HC}=\text{O}$	Aldehyde	2.094
	$\text{C}=\text{O}$	2nd position	1.71
	$\text{C}=\text{O}$	3rd position	1.69
	$\text{C}=\text{O}$	Middle	1.60
	$-\text{CH}<$	α $\text{C}=\text{O}$	0.73
	$-\text{CH}<$	β $\text{C}=\text{O}$	0.70
	$-\text{CH}<$	γ $\text{C}=\text{O}$	0.765
	$>\text{C}<$	α or β $\text{C}=\text{O}$	0.61

Example C10

Semiempirical topological index for 2-methylbutane.



$$\begin{aligned}\delta_1 &= \log 0.8 = -0.0969 \\ \delta_2 &= \log 0.9 + \log 1.0 + \log 1.0 = -0.0458 \\ \delta_3 &= \log 0.8 + \log 1.0 = -0.0969 \\ \delta_4 &= \log 0.9 = -0.0458 \\ \delta_5 &= \log 0.8 = -0.0969\end{aligned}$$

$$\begin{aligned}I_{ET} &= (1.0 - 0.0969) + (0.8 - 0.0458) + (0.9 - 0.0969) + (1.0 - 0.0458) + (1.0 - 0.0969) \\ &= 4.3177\end{aligned}$$

📖 [Amboni, da Silva Junkes *et al.*, 2002b, 2002a; da Silva Junkes, Amboni *et al.*, 2002, 2003a, 2004, 2007; da Silva Junkes, Silva Arruda *et al.*, 2005]

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **chromatographic hydrophobicity index** → chromatographic descriptors (⊙ retention time)
- **CID'/CID index** → bond order indices (⊙ graphical bond order)
- **CIM index** ≡ *Chemically Intuitive Molecular index* → spectral indices (⊙ Burden eigenvalues)
- **circuit** ≡ *cyclic path* → graph
- **circular substructure descriptors** → substructure descriptors
- **CIRD indices** → distance matrix
- **CIRS indices** → distance matrix
- **CIRS' indices** → distance matrix
- **cis/trans binary factor** → *cis/trans* descriptors

■ cis/trans descriptors

cis/trans isomerism is usually easily distinguished by using → *geometrical descriptors*, that is, descriptors derived from 3D molecular structures or structures embedded in a 3D space [Randić, Jerman-Blazic *et al.*, 1990]. Otherwise, the simplest way to distinguish cis/trans isomers is the **cis/trans binary factor**, which takes value -1 for *cis*-isomers and $+1$ for *trans*-isomers [Lekishvili, 1997].

However, when molecular descriptors are derived from molecular graphs, cis/trans isomerism is not usually recognized and some molecular descriptors were proposed to discriminate between cis/trans isomers, such as the → *corrected electron charge density connectivity index*, and → *periphery codes*. → *Weighted matrices* were also devised for obtaining the → *geometric modification number* that is added to any topological index to discriminate cis/trans isomers.

Another topological descriptor specifically proposed for cis/trans isomerism is the **Pogliani cis/trans connectivity index** χ_{CT} defined in terms of the → *Randić connectivity index* χ as

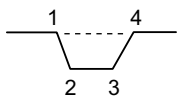
$$\chi_{CT} = \chi - \chi_{CIS} = \chi - \sum_k (\delta_1^r \cdot \delta_2 \cdot \delta_3 \cdot \delta_4^r)_k^{-1.5}$$

where the summation runs over all the cis-butadienic or cis-2-butenic fragments in the graph; δ is the → *vertex degree* and δ^r is the raised vertex degree obtained by joining the two *cis* vertices by a

virtual bond, forming a virtual four-membered ring [Pogliani, 1994b] (Example C11). For all-trans molecular graphs, $\chi_{CT} = \chi$.

Example C11

Calculation of the cis correction factor for the Pogliani cis/trans connectivity index.



$$\delta_1 = 2 + 1; \quad \delta_2 = 2; \quad \delta_3 = 2; \quad \delta_4 = 2 + 1$$

$$\chi_{CIS} = 0.046$$

📖 [Balaban, 1976b, 1998]

- **city-block distance** \equiv *Manhattan distance* \rightarrow similarity/diversity (☉ Table S7)
- **Ciubotariu shape indices** \rightarrow shape descriptors
- **Clark distance** \rightarrow similarity/diversity (☉ Table S7)
- **classical QSAR** \rightarrow structure/response correlations
- **classification** \rightarrow chemometrics

■ classification parameters

Statistical indices used to evaluate the performance of classification models [Frank and Todeschini, 1994]. They are derived from two kinds of statistics, called \rightarrow *goodness of fit* and \rightarrow *goodness of prediction*, thus distinguishing if the classification results are obtained as true predictions or not.

All the classification parameters can be derived from the **confusion matrix**, where the rows represent the known true classes and the columns the classes assigned by the classification method. It is a non-symmetric matrix of size $G \times G$, where G is the number of classes. For example, for a three-class problem ($G = 3$), the confusion matrix is:

		assigned classes			
true classes	class	A'	B'	C'	n_g
	A	c_{11}	c_{12}	c_{13}	n_a
	B	c_{21}	c_{22}	c_{23}	n_b
	C	c_{31}	c_{32}	c_{33}	n_c
	n'_g	n'_a	n'_b	n'_c	n

where A, B, and C represent labels for the true classes and A', B', and C' labels for the assigned classes. n_g represents the total number of objects effectively belonging to the g -th class and n'_g the total number of objects assigned by the classification model to the g -th class. The diagonal elements c_{gg} represent the correctly classified objects, while the off-diagonal elements c_{gk} represent the objects erroneously classified from class g to class k . Usually, two confusion matrices are obtained, one in the fitting and one after a validation procedure.

From the confusion matrix entries, the following parameters are defined:

Nonerror rate (NER), also called **overall accuracy** or simply *accuracy*, is the simplest measure of the quality of a classification model; usually expressed as a percentage, it is defined as

$$NER\% = \frac{\sum_g c_{gg}}{n} \times 100$$

where c_{gg} are the diagonal elements of the confusion matrix and n the total number of objects.

The complementary quantity is called **error rate (ER)** and is defined as

$$ER\% = \frac{n - \sum_g c_{gg}}{n} \times 100 = 100 - NER\%$$

To evaluate the efficiency of a classification model the error rate can be compared with the *no-model error rate (NOMER)*, that represents the error rate without a classification model and is calculated considering all the objects of smaller classes as erroneously classified in the largest class containing n_M objects:

$$NOMER\% = \frac{n - n_M}{n} \times 100$$

Another classification reference parameter is the *random classification error*, which is the error rate obtained if the objects are randomly assigned to the classes. It is defined as:

$$RER\% = \frac{1}{n} \cdot \left[\sum_{g=1}^G (n - n_g) \cdot p_g \right] \times 100$$

where n_g is the number of objects belonging to the g -th class, p_g is g -th class a-priori probability and n the total number of objects. $RER\%$ is equal to $NOMER\%$ if all the classes contain the same number of objects, i.e. $RER\% = NOMER\% = (1 - 1/G) \times 100$, which also corresponds to the error rate obtained by a complete random assignment.

Misclassification risk (MR). This is defined as:

$$MR\% = \sum_g \frac{(\sum_k L_{gk} \cdot c_{gk}) \cdot p_g}{n_g} \times 100$$

where p_g is the *prior class probability*, defined *a priori*, usually as $p_g = 1/G$ or $p_g = n_g/n$, where G is the total number of classes and n_g is the number of objects belonging to the g th class. L'_{gk} are elements of the *loss matrix* L , which is a user-defined nonsymmetric penalty matrix for classification errors, whose diagonal elements are zero, that is, no penalty is applied for correct classification, and the off-diagonal elements are the costs of the classification errors.

Sensitivity (Sn). A parameter that characterizes the ability of a classifier to correctly catch objects of the g th class, defined as

$$Sn_g = \frac{c_{gg}}{n_g} \times 100$$

Specificity (Sp). A parameter which characterizes the ability of the g -th class to reject objects of the other classes after the application of a classifier and defined as:

$$Sp_g = \left(1 - \frac{n'_g - c_{gg}}{n - n_g} \right) \times 100$$

where n'_g is the number of objects assigned to the g -th class.

Precision (Pr). It is a parameter which characterizes the purity of a class after the application of a classifier. It can be simply measured as the ratio of the number of objects assigned to the estimated g -th class and correctly classified (c_{gg}) over the total number of objects assigned to that class:

$$Pr_g = \frac{c_{gg}}{n'_g} \times 100$$

The degree of purity of a class can also be measured by the \rightarrow *Shannon's entropy* and the \rightarrow *Gini index*. In particular, using the Shannon's entropy, the **information gain in classification**, denoted as IG and usually expressed as percent, is calculated as:

$$IG\% = \frac{1}{H_0} \cdot \left[H_0 - \sum_{k=1}^G \frac{n'_k}{n} \cdot \sum_{g=1}^G \left(-\frac{n_{gk}}{n'_k} \cdot \log_2 \frac{n_{gk}}{n'_k} \right) \right] \times 100 = \frac{1}{H_0} \cdot \left[H_0 - \sum_{k=1}^G \frac{n'_k}{n} \cdot H'_k \right] \times 100$$

where n is the total number of objects, n'_k/n the proportion of objects in each final class (or in a node, for classification tree algorithms), n_{gk} the number of objects of the g th class present in the k th class (or node), H'_k the final entropy in each class (or node), and the summation over all the classes is the residual entropy [A-Razzak and Glen, 1992]. H_0 is the initial entropy, that is, the entropy before the classification:

$$H_0 = \sum_{g=1}^G \left[-\frac{n_g}{n} \cdot \log_2 \left(\frac{n_g}{n} \right) \right]$$

In the case of a perfect classification, the residual entropy, that is, the second term in the IG expression, is equal to zero and $IG\% = 100\%$.

Example C12

In a three-class problem ($G = 3$) of 30 objects ($n = 30$), after the application of a classification algorithm, the following confusion matrix is obtained:

Class	1'	2'	3'	n_g
1	9	1	0	10
2	2	8	2	12
3	1	2	5	8
n_k	12	11	7	30

The objects in the three classes are the following: $n_1 = 10$, $n_2 = 12$, and $n_3 = 8$. Moreover, a unitary loss matrix is assumed and the *a-priori* probabilities p_g of each class are assumed equal to $1/G$.

$$RER\% = \frac{\left[\left(\frac{30-10}{30} \right) \cdot 10 \right] + \left[\left(\frac{30-12}{30} \right) \cdot 12 \right] + \left[\left(\frac{30-8}{30} \right) \cdot 8 \right]}{30} \times 100 = 65.8\%$$

$$NOMER\% = \frac{30-12}{30} \times 100 = 60.0\%$$

$$NER\% = \frac{9+8+5}{30} \times 100 = 73.3\% \quad ER\% = 100 - 73.3 = 26.7\%$$

$$MR\% = \left[\frac{(0 \times 9 + 1 \times 1 + 1 \times 0) \cdot \frac{1}{3}}{10} + \frac{(1 \times 2 + 0 \times 8 + 1 \times 2) \cdot \frac{1}{3}}{12} + \frac{(1 \times 1 + 1 \times 2 + 0 \times 5) \cdot \frac{1}{3}}{8} \right] \times 100$$

$$= (0.033 + 0.111 + 0.125) \times 100 = 26.9\%$$

$$Sn_1 = \frac{9}{10} = 0.900 \quad Sn_2 = \frac{8}{12} = 0.667 \quad Sn_3 = \frac{5}{8} = 0.625$$

$$Sp_1 = \frac{17}{30-10} = 0.850 \quad Sp_2 = \frac{15}{30-12} = 0.833 \quad Sp_3 = \frac{20}{30-8} = 0.909$$

$$Pr_1 = \frac{9}{12} = 0.750 \quad Pr_2 = \frac{8}{11} = 0.727 \quad Pr_3 = \frac{5}{7} = 0.714$$

$$H_0 = -\frac{10}{30} \cdot \log_2 \left(\frac{10}{30} \right) - \frac{12}{30} \cdot \log_2 \left(\frac{12}{30} \right) - \frac{8}{30} \cdot \log_2 \left(\frac{8}{30} \right) = 1.566$$

$$H'_1 = -\frac{9}{12} \cdot \log_2 \left(\frac{9}{12} \right) - \frac{2}{12} \cdot \log_2 \left(\frac{2}{12} \right) - \frac{1}{12} \cdot \log_2 \left(\frac{1}{12} \right) = 1.041$$

$$H'_2 = -\frac{1}{11} \cdot \log_2 \left(\frac{1}{11} \right) - \frac{8}{11} \cdot \log_2 \left(\frac{8}{11} \right) - \frac{2}{11} \cdot \log_2 \left(\frac{2}{11} \right) = 1.096$$

$$H'_3 = -0 - \frac{2}{7} \cdot \log_2 \left(\frac{2}{7} \right) - \frac{5}{7} \cdot \log_2 \left(\frac{5}{7} \right) = 0.863$$

$$IG\% = \frac{1.566 - \left[1.041 \cdot \frac{12}{30} + 1.096 \cdot \frac{11}{30} + 0.863 \cdot \frac{7}{30} \right]}{1.566} \times 100 = 34.9\%$$

The misclassification risk calculated using *a-priori* probabilities defined by the relative frequencies of the classes, that is, $p_1 = 10/30$, $p_2 = 12/30$, and $p_3 = 8/30$, is $MR\% = 17.7\%$.

For two-class problems (the most common ones), classification parameters can be defined using \rightarrow *binary distance measures*, based on the frequencies a , b , c , and d , which in this case may be interpreted as true positive (TP), false negative (FN), false positive (FP), and true negative (TN), respectively.

Let n be the number of objects, P the number of objects belonging to the class P and N the number of objects of the class N, the following frequency table can be constructed:

	Class P'	Class N'	
Class P	TP	FN	P
Class N	FP	TN	N
	P'	N'	n

where P' is the number of objects the classifier assigns to the class P and N' the number of objects assigned to the class N, and the following relationship holds $P + N = P' + N' = n$.

The nonerror rate is then defined as

$$NER\% = \frac{TP + TN}{TP + TN + FP + FN}$$

and the **Pearson coefficient** Φ (also called **Matthews correlation index**, *MCC*, [Matthews, 1975]), which is the most used global binary classification measure, is defined as

$$\Phi \equiv MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TP + FP) \cdot (TN + FN) \cdot (TN + FP)}}$$

(See also binary similarity coefficients in \rightarrow *similarity/diversity*.)

Specific characteristics of binary classifications can also be highlighted by the following parameters:

$$\begin{aligned} Sn \equiv TPR &= \frac{TP}{TP + FN} & FNR &= \frac{FN}{TP + FN} = 1 - Sn & PPV &= \frac{TP}{TP + FP} \\ Sp \equiv TNR &= \frac{TN}{TN + FP} & FPR &= \frac{FP}{TN + FP} = 1 - Sp & NPV &= \frac{TN}{TN + FN} \end{aligned}$$

Sn being the sensitivity (or the **true positive rate**, *TPR* or **recall**), Sp the specificity (or the **true negative rate**, *TNR*), FNR the **false negative rate**, FPR the **false positive rate**, PPV the **positive predictive value** (or precision), and NPV the **negative predictive value**.

Moreover, derived from sensitivity and positive predictive value, the **F-measure** was also proposed as [Cannon, Amini *et al.*, 2007]

$$F\text{-measure} = \frac{2Sn \cdot PPV}{Sn + PPV}$$

The **Receiver Operator Characteristic curve (ROC curve)** is a graphical plot of the sensitivity Sn versus false positive rate FPR for a binary classifier system as its discrimination threshold is varied. The ROC curve can also be represented equivalently by plotting the fraction of true positives (TP) versus the fraction of false positives (FP) (Figure C3). ROC analysis provides tools to select possibly optimal classification models.

For binary classification, **weighted classification accuracy (WCA)** was also defined as [Jensen, Refsgaard *et al.*, 2005]

$$WCA = \frac{4 \cdot Sn + 1 \cdot Sp}{8 \cdot FPR + 2 \cdot FNR + 2 \cdot NCR}$$

where *NCR* stands for the nonclassified rate, defined as

$$NCR = \frac{NCP + NCN}{TP + TN + FP + FN + NCP + NCN}$$

where NCP and NCN are the number of nonclassified objects belonging to the classes P and N , respectively. The coefficients of the *WCA* index were defined to deal with the specific classification purposes of the authors and may be modified depending on the problem; however, for a perfect classification result, *WCA* suffers from singularity. Then, a modified general form of this index [Authors, This book], ranging between 0 and 1 and similar to the \rightarrow *Baroni–Urbani association index* defined for \rightarrow *binary distances*, may be the following:

$$\text{mod-WCA} = \frac{a \cdot Sn + b \cdot Sp}{a \cdot Sn + b \cdot Sp + c \cdot FPR + d \cdot FNR + e \cdot NCR}$$

where a , b , c , d , and e are the weighting coefficients to be estimated or defined depending on each specific problem.

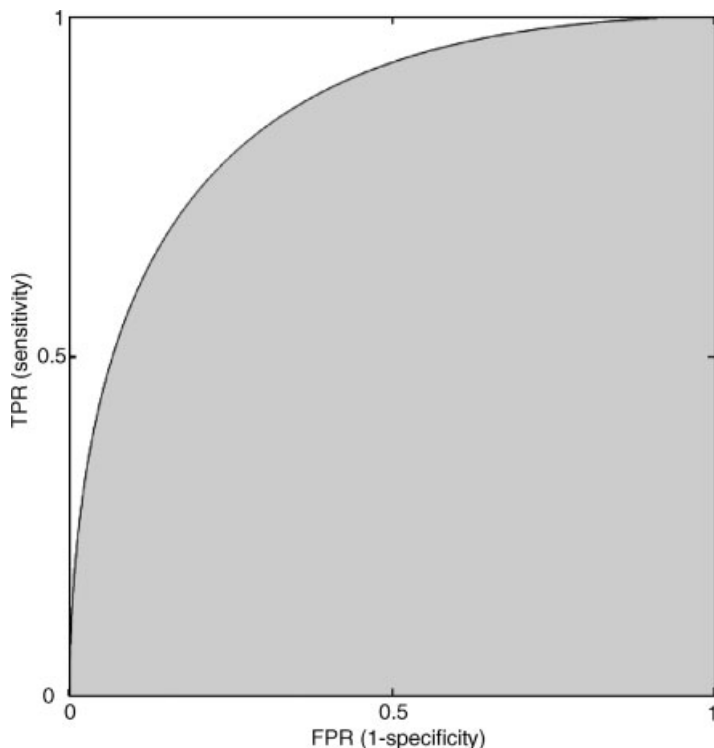


Figure C3 ROC curve.

- **class unfolding** → data set
- **clique** → graph
- **CLOGP** → lipophilicity descriptors (© Leo–Hansch hydrophobic fragmental constants)
- **closeness centrality** → center of a graph
- **cloud point** → technological properties
- **Cluj difference matrix** → Cluj matrices
- **Cluj-detour index** → Cluj matrices
- **Cluj-detour matrix** → Cluj matrices
- **Cluj-distance index** → Cluj matrices
- **Cluj-distance matrix** → Cluj matrices
- **Cluj-Ilmenau index** → Omega polynomial

■ Cluj matrices (CJ)

These are square unsymmetrical matrices $A \times A$ (A being the number of graph vertices), denoted by **UCJ**, defined following the principle of single endpoint characterization of a path; symmetric Cluj matrices, denoted by **SCJ**, are derived from the unsymmetrical Cluj matrices **UCJ** [Diudea, 1996b, 1997a, 1997b]. Several indices can be calculated from Cluj matrices, either directly by the → *orthogonal Wiener operator* from the unsymmetrical matrices or as the half-sum of entries in the symmetric matrices by the → *Wiener operator*.

A Cluj fragment, denoted by $CJ_{ij,p_{ij}}$, collects vertices lying closer to vertex v_i than to vertex v_j , the endpoints of a path p_{ij} . In other words, such a fragment collects the vertex proximity of the i th vertex against any j th vertex, joined by the path p_{ij} , with the distances measured in the subgraph $G-p_{ij}$, obtained by deleting the edges and any internal vertices of the considered path p_{ij} in the \rightarrow *H-depleted molecular graph* G ; the vertices v_i and v_j are not deleted. The Cluj fragment is formally defined as

$$CJ_{ij,p_{ij}} = \{v | v \in V(G-p_{ij}); d_{iv}(G-p_{ij}) < d_{jv}(G-p_{ij})\}$$

that is, the set of vertices closer to v_i than v_j in the component of the subgraph $G-p_{ij}$ containing v_i . The focused vertex v_i is included in the Cluj fragment. $V(G-p_{ij})$ is the set of the subgraph vertices and d is the \rightarrow *topological distance*.

In cycle-containing graphs, more than one path could join the pair v_i and v_j , thus resulting more than one Cluj fragment related to the i th vertex (with respect to the j th vertex and the given path p_{ij}). Therefore, by definition, the off-diagonal entries in the Cluj matrix are taken as the cardinality of the largest Cluj fragment, that is, the fragment with the maximum number of vertices:

$$[UCJ]_{ij} = \begin{cases} \max_{p_{ij}} |CJ_{ij,p_{ij}}| & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

For acyclic graphs, only one Cluj fragment exists for each pair of vertices and, accordingly, the Cluj matrix entry is its cardinality. Moreover, for these graphs, the Cluj fragment cardinality coincides with the number of paths going to the j th vertex through v_i . Diagonal entries are always assumed to be zero.

When the path p_{ij} belongs to the set of topological distances $D(G)$, that is, it is the shortest path connecting vertices v_i and v_j , then, the suffix **D** is added to the matrix symbol, as **UCJD** and **SCJD**, and the matrix is properly called **Cluj-distance matrix**. When the path p_{ij} belongs to the set of detours $\Delta(G)$, that is, it is the longest path connecting vertices v_i and v_j , then, the matrix symbol is **UCJA** or **SCJA**, and the matrix is called **Cluj-detour matrix** [Diudea, Pârv *et al.*, 1997a; Diudea, Katona *et al.*, 1998].

The Cluj matrices are defined for any graph and are, in general, unsymmetrical, except for some symmetric graphs. They can be symmetrized by the \rightarrow *Hadamard matrix product* with their transpose:

$$SCJ = UCJ \otimes UCJ^T$$

where UCJ^T is a transposed unsymmetrical Cluj matrix and **SCJ** is the corresponding symmetric Cluj matrix.

The Cluj matrices defined above, both symmetric and unsymmetrical, can be either **path-Cluj matrices** (UCJ_p and SCJ_p) when all the pairs of vertices of the graph are accounted for in the matrix calculation or **edge-Cluj matrices** (UCJ_e and SCJ_e) if the only nonzero elements correspond to edges, that is, only pairs of adjacent vertices are accounted for. The edge-Cluj matrices can be obtained by the Hadamard product of the path-Cluj matrices and the \rightarrow *adjacency matrix* **A**:

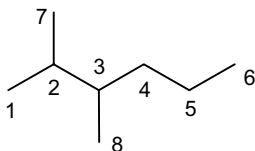
$$SCJ_e = SCJ_p \otimes A \quad UCJ_e = UCJ_p \otimes A$$

In trees, since there exists only one path joining any pair of vertices, Cluj-distance and Cluj-detour matrices coincide; moreover, symmetric Cluj matrices, **SCJD** and **SCJA**, are equal to the \rightarrow *Wiener matrix* **W** ($SCJD_e = SCJA_e = W_e$ and $SCJD_p = SCJA_p = W_p$). For cyclic graphs, Cluj-distance and Cluj-detour matrices are different, while Wiener matrices are not defined.

Moreover, the relationship for edge-matrices $SCJ_e = SZ_e$ holds for any graph, SZ_e being the \rightarrow *Szeged matrix* defined only accounting for edges, while the path-matrices are different, that is, $SCJD_p \neq SZ_p$ and $SCJA_p \neq SZ_p$.

Example C13

Distance matrix **D**, distance-path matrix **D_p**, unsymmetrical path-Cluj-distance matrix **UCJD_p**, symmetric path-Cluj-distance matrix **SCJD_p**, symmetric edge-Cluj-distance matrix **SCJD_e**, and expanded distance unsymmetrical path-Cluj-distance matrix **D_UCJD_p** for 2,3-dimethylhexane. **W_e** and **W_p** are the edge- and path-Wiener matrices, respectively. **VS_i** and **CS_j** are the row and column sums, respectively.



D										D_p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	1	2	3	4	5	2	3	20	1	0	1	3	6	10	15	3	6	44
2	1	0	1	2	3	4	1	2	14	2	1	0	1	3	6	10	1	3	25
3	2	1	0	1	2	3	2	1	12	3	3	1	0	1	3	6	3	1	18
4	3	2	1	0	1	2	3	2	14	4	6	3	1	0	1	3	6	3	23
5	4	3	2	1	0	1	4	3	18	5	10	6	3	1	0	1	10	6	37
6	5	4	3	2	1	0	5	4	24	6	15	10	6	3	1	0	15	10	60
7	2	1	2	3	4	5	0	3	20	7	3	1	3	6	10	15	0	6	44
8	3	2	1	2	3	4	3	0	18	8	6	3	1	3	6	10	6	0	35
CS _j	20	14	12	14	18	24	20	18	140	CS _j	44	25	18	23	37	60	44	35	286

Wiener index (W) = 70

Hyper-distance-path
index (D_p) = 143

UCJD_p										SCJD_p = W_p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	1	1	1	1	1	1	1	7	1	0	7	5	3	2	1	1	1	20
2	7	0	3	3	3	3	7	3	29	2	7	0	15	9	6	3	7	3	50
3	5	5	0	5	5	5	5	7	37	3	5	15	0	15	10	5	5	7	62
4	3	3	3	0	6	6	3	3	27	4	3	9	15	0	12	6	3	3	51
5	2	2	2	2	0	7	2	2	19	5	2	6	10	12	0	7	2	2	41
6	1	1	1	1	1	0	1	1	7	6	1	3	5	6	7	0	1	1	24
7	1	1	1	1	1	1	0	1	7	7	1	7	5	3	2	1	0	1	20
8	1	1	1	1	1	1	1	0	7	8	1	3	7	3	2	1	1	0	18
CS _j	20	14	12	14	18	24	20	18	140	CS _j	20	50	62	51	41	24	20	18	286

Wiener index (W) = 70

Hyper-Wiener index
(WW) = 143Hyper-Cluj-distance
index (C/D_p) = 143

SCJD _e = W _e										D_UCJD _p									
	1	2	3	4	5	6	7	8	VS _i		1	2	3	4	5	6	7	8	VS _i
1	0	7	0	0	0	0	0	0	7	1	0	1	2	3	4	5	2	3	20
2	7	0	15	0	0	0	7	0	29	2	7	0	3	6	9	12	7	6	50
3	0	15	0	15	0	0	0	7	37	3	10	5	0	5	10	15	10	7	62
4	0	0	15	0	12	0	0	0	27	4	9	6	3	0	6	12	9	6	51
5	0	0	0	12	0	7	0	0	19	5	8	6	4	2	0	7	8	6	41
6	0	0	0	0	7	0	0	0	7	6	5	4	3	2	1	0	5	4	24
7	0	7	0	0	0	0	0	0	7	7	2	1	2	3	4	5	0	3	20
8	0	0	7	0	0	0	0	0	7	8	3	2	1	2	3	4	3	0	18
CS _j	7	29	37	27	19	7	7	7	140	CS _j	44	25	18	23	37	60	44	35	286
Wiener index (W)= 70 Cluj-distance index (CJD _e)= 70										D ^U CJD _p = 143									

For acyclic graphs, the main properties of the unsymmetrical path-Cluj matrix \mathbf{UCJ}_p are

- the row sums of \mathbf{UCJ}_p are equal to the corresponding row sums of the \rightarrow *edge-Wiener matrix* \mathbf{W}_e , namely:

$$VS_i(\mathbf{UCJ}_p) = VS_i(\mathbf{W}_e)$$

where VS , which stands for vertex sum, is the \rightarrow *row sum operator*.

- the column sums of \mathbf{UCJ}_p are equal to the corresponding column sums (and row sums) of the \rightarrow *distance matrix* \mathbf{D} , namely:

$$CS_j(\mathbf{UCJ}_p) = CS_j(\mathbf{D}) = VS_i(\mathbf{D}) \quad \text{for } i = j$$

where CS_j is the \rightarrow *column sum operator*.

- from the previous relationships it follows:

$$\sum_{i=1}^A VS_i(\mathbf{UCJ}_p) = \sum_{i=1}^A VS_i(\mathbf{W}_e) = \sum_{i=1}^A VS_i(\mathbf{D}) = \sum_{j=1}^A CS_j(\mathbf{UCJ}_p) = 2 \cdot W$$

where W is the \rightarrow *Wiener index*.

\rightarrow *Expanded distance Cluj matrices* were also proposed [Diudea and Gutman, 1998] as a generalization of the expanded distance matrix and calculated by the \rightarrow *Hadamard matrix product* between the unsymmetrical path-Cluj matrices \mathbf{UCJ}_p and the \rightarrow *distance matrix* \mathbf{D} :

$$\mathbf{D_UCJ}_p = \mathbf{D} \otimes \mathbf{UCJ}_p$$

where \mathbf{UCJ} refers to both Cluj-distance and Cluj-detour matrix. Moreover, the topological distance matrix \mathbf{D} can be replaced by the \rightarrow *geometry matrix* \mathbf{G} , which collects the 3D interatomic distances, to generate expanded geometric distance Cluj matrices accounting for conformational variability and stereoisomers.

In trees, these expanded distance matrices show the two following properties:

$$\begin{aligned} VS_i(\mathbf{D_UCJ}_p) &= VS_i(\mathbf{W}_p) \\ CS_j(\mathbf{D_UCJ}_p) &= CS_j(\mathbf{D}_p) \end{aligned}$$

where VS_i is the \rightarrow row sum operator and CS_j is the \rightarrow column sum operator, \mathbf{W}_p is the \rightarrow path-Wiener matrix and \mathbf{D}_p the \rightarrow distance-path matrix.

Cluj indices are \rightarrow Wiener-type indices calculated either on symmetric (**SCJ**) or unsymmetrical (**UCJ**) Cluj matrices as [Diudea, 1997d; Diudea, Pârv *et al.*, 1997b; Diudea and Gutman, 1998; Katona and Diudea, 2003]

$$\begin{aligned} Wi(\mathbf{SCJ}_e) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SCJ}_e]_{ij} = Wi^\perp(\mathbf{UCJ}_e) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJ}_e]_{ij} \cdot [\mathbf{UCJ}_e]_{ji} \\ Wi(\mathbf{SCJ}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{SCJ}_p]_{ij} = Wi^\perp(\mathbf{UCJ}_p) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJ}_p]_{ij} \cdot [\mathbf{UCJ}_p]_{ji} \end{aligned}$$

where **CJ** refers to both Cluj-distance and Cluj-detour matrix, **CJ_e** and **CJ_p** indicate the corresponding edge-Cluj and path-Cluj matrices, respectively. Wi is the \rightarrow Wiener operator and Wi^\perp the \rightarrow orthogonal Wiener operator. A is the number of graph vertices.

Therefore, $Wi(\mathbf{SCJD}_e) = Wi^\perp(\mathbf{UCJD}_e)$ is the **Cluj-distance index** (CJD_e), $Wi(\mathbf{SCJD}_p) = Wi^\perp(\mathbf{UCJD}_p)$ is the **hyper-Cluj-distance index** (CJD_p), $Wi(\mathbf{SCJ}\Delta_e) = Wi^\perp(\mathbf{UCJ}\Delta_e)$ is the **Cluj-detour index** ($CJ\Delta_e$), and $Wi(\mathbf{SCJ}\Delta_p) = Wi^\perp(\mathbf{UCJ}\Delta_p)$ is the **hyper-Cluj-detour index** ($CJ\Delta_p$).

Note that the \rightarrow Szeged index SZ_e equals the Cluj-distance index CJD_e for any graph. Moreover, for acyclic graphs, the following relationships hold:

$$\begin{aligned} Wi(\mathbf{UCJD}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{UCJD}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{W}_e]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D}]_{ij} = W \\ Wi(\mathbf{D_UCJD}_p) &= \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D_UCJD}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{W}_p]_{ij} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A [\mathbf{D}_p]_{ij} = WW \end{aligned}$$

where \mathbf{UCJD}_p is the unsymmetrical path-Cluj-distance matrix, $\mathbf{D_UCJD}_p$ the corresponding expanded Cluj-distance matrix, \mathbf{W}_e and \mathbf{W}_p the edge-Wiener and path-Wiener matrices, respectively, \mathbf{D} the topological distance matrix, and \mathbf{D}_p the distance-path matrix.

Thus, $Wi(\mathbf{D_UCJD}_p)$ reduces to the \rightarrow hyper-Wiener index WW , calculated as the half sum of entries in the matrix $\mathbf{D_UCJD}_p$. This matrix is a direct proof of the finding that, in acyclic graphs, the sum of all internal paths (given by \mathbf{D}_p) equals the sum of all external paths (given by \mathbf{W}_p) with respect to all pairs (i, j) in the graph [Klein, Lukovits *et al.*, 1995]. The matrix $\mathbf{D_UCJD}_p$ offers an alternative definition of the hyper-Wiener index.

Moreover, from unsymmetrical Cluj matrices **UCJ** other invariants are the \rightarrow matrix sum indices MS calculated as:

$$MS(\mathbf{UCJ}) = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{UCJ}]_{ij}$$

where the summation goes over all the matrix elements.

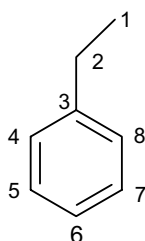
In a bipartite graph, the sum of all edge-counted vertex proximities $MS(\mathbf{UCJD}_e)$ equals the product $A \times B$ of the number of vertices and edges in the graph (e.g., $8 \times 7 = 56$ for 2,3-dimethylhexane in Example C13).

In a tree graph, the sum of all path-counted vertex proximities $MS(\mathbf{UCJD}_p)$ is twice the sum of all distances in the graph or twice the \rightarrow Wiener index W . Moreover, in trees, the \rightarrow PI index,

which represents the edge-counted nonequidistant edges, can be calculated as $MS(UCJD_e)$ from the \rightarrow line graph of the considered molecular graph.

Example C14

Unsymmetrical and symmetric path-Cluj-distance ($UCJD_p$, $SCJD_p$) and path-Cluj-detour ($UCJA_p$, $SCJA_p$) matrices for ethylbenzene. Elements of the corresponding edge-Cluj matrices are highlighted in bold face. VS_i and CS_j are the row sum and column sums, respectively.



$UCJD_p$										$SCJD_p$									
	1	2	3	4	5	6	7	8	VS_i		1	2	3	4	5	6	7	8	VS_j
1	0	1	1	1	1	1	1	1	7	1	0	7	6	4	3	3	3	4	30
2	7	0	2	2	2	2	2	2	19	2	7	0	12	8	6	4	6	6	49
3	6	6	0	5	4	4	4	5	34	3	6	12	0	15	8	8	8	15	72
4	4	4	3	0	5	4	4	2	26	4	4	8	15	0	15	8	8	4	62
5	3	3	2	3	0	5	2	2	20	5	3	6	8	15	0	15	4	8	59
6	3	2	2	2	3	0	3	2	17	6	3	4	8	8	15	0	15	8	61
7	3	3	2	2	2	5	0	3	20	7	3	6	8	8	4	15	0	15	59
8	4	3	3	2	4	4	5	0	25	8	4	6	15	4	8	8	15	0	60
CS_j	30	22	15	17	21	25	21	17	168	CS_j	30	49	72	62	59	61	59	60	452

Cluj-distance index CJD_e :

$$Wi^\perp(UCJD_e) = Wi(SCJD_e) = 109$$

Hyper-Cluj-distance index CJD_p :

$$Wi^\perp(UCJD_p) = Wi(SCJD_p) = 226$$

$UCJA_p$										$SCJA_p$									
	1	2	3	4	5	6	7	8	VS_i		1	2	3	4	5	6	7	8	VS_j
1	0	1	1	1	1	1	1	1	7	1	0	7	6	1	2	3	2	1	22
2	7	0	2	2	2	2	2	2	19	2	7	0	12	2	4	4	4	2	35
3	6	6	0	3	3	4	3	3	28	3	6	12	0	3	3	8	3	3	38
4	1	1	1	0	1	1	4	1	10	4	1	2	3	0	1	1	8	1	17
5	2	2	1	1	0	1	1	2	10	5	2	4	3	1	0	1	1	8	20
6	3	2	2	1	1	0	1	1	11	6	3	4	8	1	1	0	1	1	19
7	2	2	1	2	1	1	0	1	10	7	2	4	3	8	1	1	0	1	20
8	1	1	1	1	4	1	1	0	10	8	1	2	3	1	8	1	1	0	17
CS_j	22	15	9	11	13	11	13	11	105	CS_j	22	35	38	17	20	19	20	17	188

Cluj-detour index CJA_e :

$$Wi^\perp(UCJA_e) = Wi(SCJA_e) = 29$$

Hyper-Cluj-detour index CJA_p :

$$Wi^\perp(UCJA_p) = Wi(SCJA_p) = 94$$

Other Cluj indices can be derived from the Cluj polynomials. The **Cluj polynomials** are \rightarrow counting polynomials defined on the basis of Cluj matrices as [Diudea, 2002a; Diudea, Vizitiu *et al.*, 2007]

$$CJ(G; x) = \sum_k m(G; k) \cdot x^k$$

where the coefficients $m(G; k)$ are calculated from the entries of the Cluj matrices as the frequency of occurrence of each value k . Cluj polynomials can be calculated both from edge-Cluj matrices (i.e., *Cluj-edge polynomial*) and path-Cluj matrices (i.e., *Cluj-path polynomial*).

Derived from Cluj matrices, the **reciprocal Cluj matrices** CJ^{-1} are the matrices whose elements are the reciprocal of the corresponding Cluj matrix elements [Diudea, 1997c; Diudea, Katona *et al.*, 1998]. \rightarrow *Harary indices* and \rightarrow *hyper-Harary indices* are defined for these matrices applying Wiener and orthogonal Wiener operators.

The **Cluj difference matrix**, denoted as CJ_{Δ} , is obtained in the same way as the Cluj matrix CJ , but path contributions are calculated only on paths larger than one [Diudea, 1996b, 1997a; Ivanciuc, Ivanciuc *et al.*, 1997]. The Cluj difference matrix is calculated as difference between path-Cluj and edge-Cluj matrices:

$$CJ_{\Delta} = CJ_p - CJ_e$$

📖 [Diudea and Randić, 1997; Diudea, 1997b; Gutman and Diudea, 1998; Jäntschi, Katona *et al.*, 2000; Ardelan, Katona *et al.*, 2001; Ursu, Don *et al.*, 2004]

- **Cluj polynomials** \rightarrow Cluj matrices
- **cluster analysis** \rightarrow chemometrics
- **cluster analysis feature selection** \rightarrow variable reduction

■ cluster expansion of chemical graphs

Given a \rightarrow molecular graph G , where vertices are labeled by the chemical element of the corresponding atoms, cluster expansion in the additive form is among the \rightarrow group contribution methods expressing a molecular property Φ as a sum of contributions of all the connected subgraphs of G , that is,

$$\Phi(G) = \sum_k \phi_k(G') N_k$$

where ϕ_k is the contribution to the molecular property of the k th fragment and k runs over all the connected subgraphs G' . N_k are called **embedding frequencies** and are the number of times a given substructure (cluster) appears in a chemically isomorphic subgraph within the molecular graph. In practice, embedding frequencies are \rightarrow count descriptors such as atom-type counts, two-atom fragment counts, etc. [Smolenskii, 1964; Gordon and Kennedy, 1973; Essam, Kennedy *et al.*, 1977; Klein, 1986; Schmalz, Klein *et al.*, 1992]. The property contributions of the fragments are estimated by multivariate regression analysis.

Usually this method is used on a \rightarrow *H-depleted molecular graph*, truncated expansions being obtained considering only fragments up to a user-defined size. Some methods for \rightarrow log P estimation are based on cluster expansion. Moreover, a new method for the calculation of embedding frequencies for acyclic trees based on \rightarrow spectral moments of iterated line graph sequence was proposed by [Gutman, Popovic *et al.*, 1998; Estrada, 1999c].

📖 [Schmalz, Živković *et al.*, 1987; Poshusta and McHughes, 1989; McHughes and Poshusta, 1990; Baskin, Skvortsova *et al.*, 1995; Grassy, Trape *et al.*, 1995; Kvasnička and Pospichal, 1995; Klein, Schmalz *et al.*, 1999]

- **clustering coefficient of a vertex** → adjacency matrix
- **cluster significance analysis** → variable selection
- **cluster subgraph** → molecular graph
- **CMC** \equiv *critical micelle concentration* → technological properties
- **CMC index** → similarity/diversity
- **CMD index** → similarity/diversity

■ CODESSA descriptors

Among the several CODESSA descriptors, implemented in the homonymous software CODESSA (*Comprehensive DEscriptors for Structural and Statistical Analysis*) [Katritzky and Gordееva, 1993; CODESSA – Katritzky, Lobanov *et al.*, 1996; Katritzky, Lobanov *et al.*, 1996], are → *molecular weight*, → *molecular volume*, → *count descriptors*, → *topological indices*, → *charge descriptors*, → *shadow indices*, → *charged partial surface area descriptors*, → *quantum-chemical descriptors*, and → *electric polarization descriptors*.

The software CODESSA allows to perform QSAR analysis starting from the calculation of theoretical molecular descriptors up to the evaluation of the best multivariate linear models based on → *variable selection*.

A stepwise selection procedure is adopted to search for QSPR/QSAR models after the preliminary exclusion of → *constant and near-constant variables*. The → *pair correlation cut-off selection* of variables is then performed to avoid highly correlated descriptor variables within the model.

Several molecular properties have been modeled by CODESSA descriptors, such as chromatographic indices [Katritzky, Ignatchenko *et al.*, 1994; Pompe and Novič, 1999], boiling [Katritzky, Mu *et al.*, 1996b; Katritzky, Lobanov *et al.*, 1998; Ivanciuc, Ivanciuc *et al.*, 1998b] and melting points [Katritzky, Maran *et al.*, 1997], critical temperatures [Katritzky, Mu *et al.*, 1998], gas solubilities [Katritzky, Mu *et al.*, 1996a; Huibers and Katritzky, 1998; Katritzky, Wang *et al.*, 1998], critical micelle concentrations [Huibers, Lobanov *et al.*, 1996, 1997], → *solvent polarity scales* [Katritzky, Mu *et al.*, 1997], and mutagenic activities [Maran, Karelson *et al.*, 1999].

📖 Additional references are collected in the thematic bibliography (see Introduction).

- **coefficient of alienation** \equiv *coefficient of nondetermination* → regression parameters
- **coefficient of determination** → regression parameters
- **coefficient of divergence** \equiv *Clark distance* → similarity/diversity (Table S7)
- **coefficient of nondetermination** → regression parameters
- **coefficient of variation** → statistical indices (\odot indices of dispersion)
- **color classes** → chromatic decomposition
- **column sum operator** → algebraic operators
- **column sum vector** → algebraic operators (\odot column sum operator)
- **combinatorial Laplacian matrix** \equiv *Laplacian matrix*
- **combinatorial matrices** → matrices of molecules

■ combined descriptors

These are fixed combinations of selected descriptors accounting for molecular properties of interest. The simplest combined descriptors are the differences and average values of → *basis descriptors* such as → *connectivity indices* or → *path numbers*, and the ratios of different

descriptors defined with the aim of normalization to obtain, for example, size-independent indices. Moreover, optimal linear combinations of highly correlated descriptors are combined descriptors calculated so as to reduce the number of independent variables (e.g., \rightarrow *principal properties*).

Simple sums of different molecular descriptors were proposed as \rightarrow *superindices* to obtain highly discriminant indices; particular superindices were suggested to account for \rightarrow *molecular complexity*.

Examples of combined descriptors are reported below.

Difference indices were proposed as the difference between topological descriptors obtained from the \rightarrow *distance matrix* \mathbf{D} and the \rightarrow *detour matrix* $\mathbf{\Delta}$; they are defined as [Castro, Tueros *et al.*, 2000]

$$\Delta\mathcal{D} = \mathcal{D}(\mathbf{D}) - \mathcal{D}(\mathbf{\Delta})$$

where $\mathcal{D}(\mathbf{D})$ and $\mathcal{D}(\mathbf{\Delta})$ indicate any molecular descriptor obtained from distance and detour matrix, respectively. Difference indices were calculated for \rightarrow *Wiener index*, \rightarrow *Zagreb indices*, and \rightarrow *Schultz molecular topological index*; they equal zero for any acyclic graph, since distance and detour matrices coincide.

A special case of difference indices are the **differential descriptors**, which are \rightarrow *molecular descriptors* or \rightarrow *substituent descriptors* calculated by difference between a compound (or functional group, fragment) and a \rightarrow *reference structure* or a \rightarrow *hyperstructure*. Examples of differential descriptors are those obtained by the \rightarrow *minimal topological difference* (MTD) and \rightarrow *molecular shape analysis* (MSA), as well as some descriptors among the \rightarrow *ETA indices*.

Differential connectivity indices (or **connectivity differences**) are defined as the difference between connectivity indices ${}^m\chi$ and \rightarrow *valence connectivity indices* ${}^m\chi^v$ [Hall and Kier, 1986; Kier and Hall, 1991; Gálvez, García-Domenech *et al.*, 1995; Llacer, Gálvez *et al.*, 2006]:

$${}^m\Delta\chi_t = {}^m\chi_t - {}^m\chi_t^v$$

where the superscript m denotes the order of connectivity indices and the subscript t the type of \rightarrow *molecular subgraph*. These are descriptors proposed to encode electronic information in terms of π and lone pair electrons on that part of the molecule defined by m and t ; moreover, it was found that such descriptors are related to differences in inductive and mesomeric effects [Gálvez, García-Domenech *et al.*, 1994].

Distance measure connectivity indices (DM) are derived from the set of molecular connectivity indices by means of the definition of the \rightarrow *Minkowski distance* [Balaban, Ciubotariu *et al.*, 1990]. They are calculated as

$$DM^k = \sum_{j=1}^{14} [({}^m\chi_t - {}^m\chi_t(R))^k]^{1/k}$$

where the summation goes over all the connectivity indices of different type t up to the sixth order ($m = 6$); k is an integer parameter ranging from 1 to 5 ($k = 1$ is the \rightarrow *Manhattan distance*, $k = 2$ is the \rightarrow *Euclidean distance*); ${}^m\chi_t$ and ${}^m\chi_t(R)$ are the connectivity indices for the considered molecule and a reference molecule R , respectively. DM^k indices can be interpreted as a 14-dimensional measure of the structural diversity of the compound from the reference compound. Methane was proposed as the reference structure, having all the connectivity indices equal to zero.

To account for nondispersive force effects, the **relative valence connectivity indices** to nonpolar compounds were defined as

$$\Delta\chi_{np} = \chi_{np}^v - \chi^v$$

where the nonpolar connectivity index χ_{np}^v is calculated substituting oxygen and nitrogen atoms in the considered molecule by carbon atoms but keeping the number of bonds to all nonhydrogen atoms constant [Bahnick and Doucette, 1988; Schramke, Murphy *et al.*, 1999]. Exceptions to bond constancy were made by replacing the carbonyl group with C—C instead of C=C unless the oxygen atom was directly bonded to an unsaturated ring system (uracils), or the nitrile group with C=C. The difference between connectivity indices of adjacent order was also proposed to model surface tension (${}^2\chi - {}^3\chi$) and critical temperature (${}^1\chi - {}^2\chi$) of alkanes [Randić and Basak, 1999].

A **topological Hammett function** σ_t was also defined by the most significant differences between the \rightarrow *connectivity indices* as

$$\sigma_t = b_0 + b_1 \cdot ({}^4\chi_p - {}^4\chi_p^v) + b_2 \cdot ({}^4\chi_{pc} - {}^4\chi_{pc}^v)$$

where ${}^4\chi_p$, ${}^4\chi_p^v$, ${}^4\chi_{pc}$, and ${}^4\chi_{pc}^v$ are the fourth-order atom and valence \rightarrow *connectivity indices* for path (*p*) and path-cluster (*pc*) graph decompositions; b_j are estimated regression coefficients.

The **L index** was proposed as the molecular descriptor defined as the simple linear combination of molecular \rightarrow *path counts* of order one 1P (the number of bonds), order two 2P (the \rightarrow *connection number* N_2), and order three 3P :

$$L = 2 \cdot {}^1P + {}^2P - {}^3P - 2$$

It was found to correlate the sum of ${}^{13}\text{C}$ atomic chemical shifts in alkanes [Miyashita, Okuyama *et al.*, 1989].

Moreover, path count differences ${}^1P - {}^2P$ and ${}^2P - {}^3P$ [Randić and Trinajstić, 1988] and connectivity differences ${}^1\chi - {}^2\chi$ and ${}^2\chi - {}^3\chi$ are often encountered in QSAR modeling; the following path count combination $P_0 + P_1 + P_2 + P_3$ was also found as the critical parameter in the correlation of carbon-13 \rightarrow *chemical shift sums* in alkanes [Miyashita, Okuyama *et al.*, 1989].

Other examples of combined descriptors are the **connectivity quotients** defined as [Gálvez, García-Domenech *et al.*, 1995; Llacer, Gálvez *et al.*, 2006]

$${}^mC_t = \frac{{}^m\chi_t}{{}^m\chi_t^v}$$

where ${}^m\chi$ are the simple \rightarrow *connectivity indices* and ${}^m\chi^v$ the \rightarrow *valence connectivity indices*. Examples of other connectivity quotients are

$${}^1C = \frac{{}^1\chi}{{}^1\chi^v + 1} \quad {}^4C_p = \frac{{}^4\chi_p}{{}^4\chi_p^v + 1} \quad \chi^{23} = \frac{{}^4\chi_p + {}^3\chi}{2}$$

where the first two were proposed by Gálvez [Gálvez, Gomez-Lechón *et al.*, 1996], and the last one was found to correlate well with the van der Waals area [Randić, 1991g].

Semiempirical molecular connectivity terms *X* are special combinations of \rightarrow *connectivity indices* that make use of empirical parameters, dielectric constants, molar masses, and other

ad hoc related parameters accounting for noncovalent interactions [Pogliani, 1997a, 1999a, 1999c]; an example is

$$X = \frac{{}^1\chi}{{}^2\chi + b \cdot {}^3\chi}$$

where b is a parameter to be optimized. These connectivity terms are derived by a trial-and-error procedure based on connectivity indices of lower order.

Other examples of combined descriptors are ratios of some \rightarrow *count descriptors* used by [Zheng, Luo *et al.*, 2005] to define \rightarrow *property filters* and the following:

$$\frac{W}{Z} \quad \frac{CID}{{}^1\chi} \quad \frac{{}^4\chi_{pc}}{MW} \quad \frac{N_X}{MW}$$

where W is the \rightarrow *Wiener index*, Z the \rightarrow *Hosoya Z index*, CID the \rightarrow *connectivity ID number*, N_X the number of atoms of type X , and MW the molecular weight [Boethling and Sabljic, 1989].

Examples of combined descriptors using products are the contributions $q_a \cdot SA_a$, largely used in \rightarrow *CPSA descriptors*, where q and SA are partial charges and atomic surface areas, respectively [Bakken and Jurs, 1999a].

📖 [Stanton and Jurs, 1992; Randić, 1993a]

- **combined matrices** \rightarrow matrices of molecules
- **CoMFA** \equiv *Comparative Molecular Field Analysis* \rightarrow grid-based QSAR techniques
- **CoMFA descriptors** \rightarrow grid-based QSAR techniques (\odot Comparative Molecular Field Analysis)
- **CoMFA fields** \rightarrow molecular interaction fields
- **CoMFA lattice** \rightarrow grid-based QSAR techniques (\odot Comparative Molecular Field Analysis)
- **CoMMA** \equiv *Comparative Molecular Moment Analysis*
- **CoMMA descriptors** \rightarrow comparative molecular moment analysis
- **common overlap length** \rightarrow molecular shape analysis (\odot common overlap steric volume)
- **common overlap surface** \rightarrow molecular shape analysis (\odot common overlap steric volume)
- **common overlap steric volume** \rightarrow molecular shape analysis
- **compactness** \rightarrow distance matrix
- **Comparative Molecular Field Analysis** \rightarrow grid-based QSAR techniques

■ Comparative Molecular Moment Analysis (CoMMA)

The Comparative Molecular Moment Analysis method based on the 3D \rightarrow *molecular geometry* calculates different molecular moments with respect to the \rightarrow *center of mass*, center of charge, and \rightarrow *center-of-dipole* of the molecule [Silverman and Platt, 1996; Silverman, Pitman *et al.*, 1998].

CoMMA descriptors are the following 14 molecular descriptors:

$$\{MW; I_x, I_y, I_z; \mu; Q; \mu_x, \mu_y, \mu_z; d_x, d_y, d_z; Q_{xx}, Q_{yy}\}$$

The first descriptor MW is the \rightarrow *molecular weight*, that is, the zero-order molecular moment with respect to the center of mass. The three \rightarrow *principal moments of inertia* I are the second-order moments with respect to the center of mass. μ and Q are the magnitudes of \rightarrow *dipole*

moment and \rightarrow *quadrupole moment* that are the first- and the second-order moments with respect to the center of charge, respectively. The dipole moment components μ_x , μ_y , and μ_z and the components of displacement d between the center of mass and the center of dipole are calculated with respect to the \rightarrow *principal inertia axes*. Finally, the quadrupole components Q_{xx} and Q_{yy} are calculated with respect to a translated initial reference frame whose origin coincides with the center-of-dipole (Table C5).

By calculating molecular descriptors based on 3D geometry without a common orientation frame, the Comparative Molecular Moment Analysis overcomes the problems due to the molecule alignment.

To extend the CoMMA approach to account for the lipophilicity of the molecule, the \rightarrow *Leo–Hansch hydrophobic fragmental constants* [Abraham and Leo, 1987] have been proposed as a set of atomic lipophilic weights for the calculation of lipophilic molecular multipole moments, called **hydropoles** [Burden and Winkler, 1999a].

Table C5 Molecular moments of order zero, one, and two. A is the number of atoms, MW the molecular weight, q the atomic charges, μ the total dipole moment, and f the hydrophobic atomic constants.

Moment order	Unit	Mass	Charge	Lipophilicity
0	A	MW	$\sum_i q_i$	$\sum_i f_i$
1	0	0	μ	Lipophilic dipole moment
2	Moments of geometry	Moments of inertia	Electrostatic quadrupole moments	Lipophilic quadrupole moments

📖 [Silverman, Pitman *et al.*, 1998, 1999; Silverman, Platt *et al.*, 1998; Burden and Winkler, 1999a; Silverman, 2000a, 2000b; Pitman, Huber *et al.*, 2001; Kovatcheva, Golbraikh *et al.*, 2004; Can, Dimoglo *et al.*, 2005]

- **Comparative Molecular Similarity Indices Analysis** \rightarrow grid-based QSAR techniques
- **Comparative Molecular Surface Analysis** \rightarrow grid-based QSAR techniques
- **Comparative Receptor Surface Analysis** \equiv CoRSA
- **Comparative Spectral Analysis** \rightarrow spectra descriptors
- **Comparative Structurally Assigned Spectral Analysis** \rightarrow spectra descriptors
- **Compass descriptors** \rightarrow Compass method

■ Compass method

A QSAR method based on the search for the best model predicting compound activity and likely bioactive conformations and alignments from a set of physical properties measured only near the surface of the molecules [Jain, Koile *et al.*, 1994; Jain, Dietterich *et al.*, 1994; Jain, Harris *et al.*, 1995]. The basic assumption is that the enthalpy of ligand-target binding depends on the interactions occurring at the ligand-target interface. Therefore, the main features characterizing the Compass method are the definition of descriptors related to surface properties, an automatic selection of the optimal molecular conformation and alignment, and the use of \rightarrow *artificial neural networks* with back-propagation to take into account also nonlinear structure-activity relationships.

The method is based on three fundamental phases. The first phase consists in the generation of low-energy conformations for each molecule and in the choice of one conformer as the one most likely to be bioactive; all selected conformers are aligned along with the identified pharmacophore or a substructure common to all molecules in the data set. A molecule *pose* is a conformation of the molecule in a particular alignment.

The second phase proceeds iteratively through three steps. (a) For each molecule pose **Compass descriptors** are calculated as \rightarrow *geometric distances* representing the surface shape or polar functionalities of the pose in the proximity of a given point in the space; compass steric descriptors measure distances from sampling points to the van der Waals surface of a molecule, while donor/acceptor ability descriptors measure the distance from a sampling point to the nearest H-bond donor or acceptor group. Few sampling points are scattered on a surface 2.0 \AA outside the average van der Waals envelope of the \rightarrow *hypermolecule* obtained by alignment in an invariant and common reference frame (Figure C4). (b) A neural network model is built relating the structural features (Compass descriptors) of molecule poses to biological activity. The network is trained by the backpropagation algorithm and is constituted by three layers with Gaussian input units and standard sigmoid units in the hidden layer. (c) In the third step the model is used to realign the molecules to find better poses, which are then used to give an improved model until convergence is reached.

The third step predicts the activity and bioactive pose of a new molecule.

With respect to \rightarrow CoMFA, the Compass method effectively reduces the number of descriptors, performing a physico-chemically based \rightarrow *variable reduction* and overcomes the problem of guessing the best conformation and alignment of the molecules.

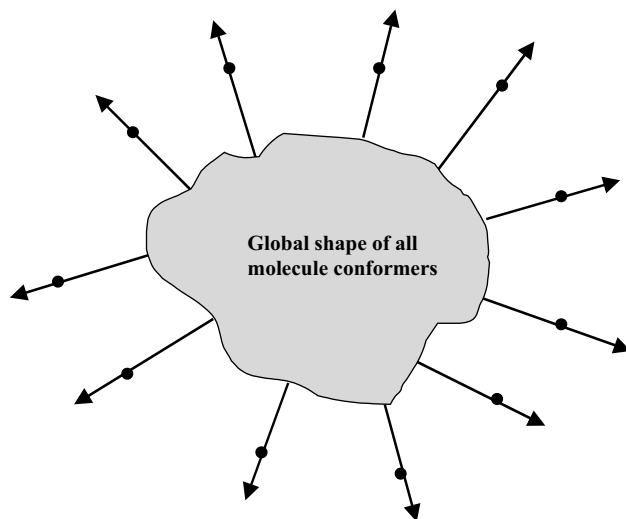


Figure C4 Compass descriptors arising from the molecular surface.

Morphological similarity is a 3D molecular similarity method based on surface shape and charge characteristics of compounds [Jain, 2000]. As in the Compass method, distances to molecular surface from weighted observation points on a uniform grid are calculated. Morphological similarity is defined as a Gaussian function of the differences in molecular

surface distances of two molecules. At each point, a weight is defined in such a way that only grid points that are on the outside of one or another of two molecules contribute to the measure of similarity. Moreover, at each point, for each molecule, in a particular conformation and alignment, three distances are computed: the minimum distance to the van der Waals surface, the minimum distance to a hydrogen-bond acceptor or negatively charged atom, and the minimum distance to a hydrogen-bond donor or positively charged atom. In addition, a directionality term is computed that corresponds to the directional concordance of the vector from an observation point to the polar atom and the atom's favored interaction vector.

- **complementary distance matrix** → distance matrix
- **complementary information content** → indices of neighborhood symmetry
- **complementary Wiener indices** → distance matrix
- **complement Balaban index** → distance matrix
- **complement Barysz distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **complement matrices** → matrices of molecules
- **complement Wiener index** → distance matrix
- **complete centric index** → centric indices
- **complete graph** → graph
- **complexity indices** ≡ *molecular complexity indices* → molecular complexity
- **composite ETA index** → ETA indices
- **composite nuclear potential** → quantum-chemical descriptors
- **composite reference ETA index** → ETA indices
- **composition indices** ≡ *atomic composition indices*
- **Compressed Feature Matrix** → substructure descriptors (⊙ pharmacophore-based descriptors)

■ computational chemistry

In a broad sense, the term computational chemistry includes several fields such as quantum chemistry, statistical molecular mechanics, molecular modeling, approaches based on → *graph invariants*, molecular graphics and visualization, evaluation of experimental data in X-ray crystallography, NMR spectroscopy, and, in general, spectroscopic techniques; moreover, in this broad sense, analysis, exploration, and modeling performed by → *chemometrics* on experimental data, searching for → *structure-response correlations*, information retrieval from chemical databases, and expert chemical systems are also included in computational chemistry, as constitutive parts of → *chemoinformatics*.

Theoretical chemistry and, especially, quantum chemistry constitute the basic core of computational chemistry and their success covers the field of molecular geometries and energies, reactivity, spectroscopic properties, behavior of electrons in atoms and molecules, and various other fundamental chemical topics [Lipkowitz and Boyd, 1990]. Therefore, the term computational chemistry is also used in a more restricted sense to denote the mathematical approaches and their software implementations to the calculation of molecular properties from theoretical chemistry. → *Quantum-chemical descriptors* are derived from computational chemistry in this restricted sense.

Together with the many methods based on quantum chemistry, other important and effective approaches to computational chemistry are those called *Empirical Force-Field methods* (EFF methods), based on a mechanistic view of the molecule in terms of force constants of bonds,

bending, torsion, and other special interaction terms. The set of force constants constitutes a field of empirical parameters used for the calculation of molecular geometries and energies.

Calculations based on computational chemistry methods can be performed by means of software packages, such as AMPAC [AMPAC, 2005], GAMESS [GAMESS, 2005], GAUSSIAN [GAUSSIAN03 – Pople and *et al.*, 1990], JAGUAR [Jaguar – Schrödinger, 1990], MOLPRO [MOLPRO – Werner, Knowles *et al.*, 1991], MOPAC [MOPAC – Air Force Academy, 1999], NWCHEM [NWChem – EMSL, 1990], SPARTAN [SPARTAN, 2005], and TURBOMOLE [TURBOMOLE, 2007].

📖 [Lewis, 1916, 1923; Mulliken, 1928a, 1928b, 1955a; Hückel, 1930, 1932; Pauling, 1932, 1939; Pauling and Wilson, 1935; Coulson, 1939, 1960; Eyring, Walter *et al.*, 1944; Streitweiser, 1961; Dewar, 1969; Murrell and Harget, 1972; Lowe, 1978; Löw and Saller, 1988; Parr and Yang, 1989; Stewart, 1990; Leach, 1996; Szabo and Ostlund, 1996; Jorgensen, Olsen *et al.*, 2000]

- **Computer-Aided Drug Design** → drug design
- **Computer-Aided Molecular Design** → drug design
- **Computer-Aided Molecular modeling** → drug design
- **COMSA** \equiv *Comparative Molecular Surface Analysis* → topological feature maps
- **CoMSIA** \equiv *Comparative Molecular Similarity Indices Analysis* → grid-based QSAR techniques
- **CON index** → statistical indices (⊙ concentration indices)
- **concentration indices** → statistical indices
- **conditional Wiener index** → Wiener index
- **conductance matrix** → resistance matrix
- **Conformational-Dependent Chirality Code** → chirality descriptors (⊙ Chirality Codes)
- **conformational global sensitivity** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **Conformational-Independent Chirality Code** → chirality descriptors (⊙ Chirality Codes)
- **conformational invariance** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **conformational pairwise sensitivity** → molecular descriptors (⊙ invariance properties of molecular descriptors)
- **Conformation Energy Profile** → 4D-Molecular Similarity Analysis
- **confusion matrix** → classification parameters
- **congenericity principle** → Structure/Response Correlations
- **conjugation** → delocalization degree indices
- **connected graph** → graph
- **connectedness index** → Wiener index
- **connection** → edge adjacency matrix
- **connection number** → edge adjacency matrix
- **connection orbital information content** → orbital information indices
- **connective eccentricity index** → eccentricity-based Madan indices (⊙ Table E1)
- **connectivity bond layer matrix** → layer matrices
- **connectivity differences** \equiv *differential connectivity indices* → combined descriptors
- **connectivity ID number** \equiv *Randić connectivity ID number* → ID numbers
- **connectivity index** \equiv *Randić connectivity index* → connectivity indices

■ connectivity indices

Connectivity indices are among the most popular \rightarrow *topological indices* and are calculated from the \rightarrow *vertex degree* δ of the atoms in the \rightarrow *H-depleted molecular graph*. The **Randić connectivity index** was the first connectivity index proposed [Randić, 1975b, 2008; Li and Gutman, 2006]; it is also called **connectivity index** or **branching index**, and is defined as

$$\chi_R \equiv {}^1\chi = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i \cdot \delta_j)^{-1/2} = \sum_{b=1}^B (\delta_{b(1)} \cdot \delta_{b(2)})_b^{-1/2}$$

where the first summation goes over all the pairs of vertices v_i and v_j in the molecular graph, but only contributions from pairs of adjacent vertices are accounted for, a_{ij} being the elements of the \rightarrow *adjacency matrix* A ; the second summation goes over all the edges in the molecular graph. A and B are the total number of vertices and edges in the graph, respectively; δ_i and δ_j are the vertex degrees of the vertices v_i and v_j ; the subscripts $b(1)$ and $b(2)$ represent the two vertices connected by the edge b .

The Randić connectivity index is closely related to the \rightarrow *second Zagreb index* M_2 and was proposed as measure of \rightarrow *molecular branching*.

The term $(\delta_i \cdot \delta_j)^{-1/2}$ for each pair of adjacent vertices is called **edge connectivity** and can be used to characterize edges as a primitive \rightarrow *bond order* accounting for bond accessibility, that is, the accessibility of a bond to encounter another bond in intermolecular interactions, as the reciprocal of the vertex degree δ is the fraction of the total number of nonhydrogen sigma electrons contributed to each bond formed with a particular atom [Kier and Hall, 2000]. This interpretation places emphasis on the bimolecular encounter possibility among molecules, reflecting collective influence of the bond accessibilities of each molecule with other molecules in its immediate environment. Therefore, the Randić connectivity index ${}^1\chi$ can be interpreted as the contribution of one molecule to the bimolecular interaction arising from the encounters of bonds of two identical molecules:

$${}^1\chi = \sqrt{\sum_{b=1}^B \sum_{b'=1}^B (\delta_i \cdot \delta_j)_b^{-1/2} \cdot (\delta_k \cdot \delta_l)_{b'}^{-1/2}}$$

where the two summations run over all the bonds of the molecules and δ are the vertex degrees.

Important papers about characteristics and meaning of the connectivity indices are: [Kier and Hall, 2000, 2002; Hall and Kier, 2001; Randić, 2001g; Estrada, 2002b].

\rightarrow *Information connectivity indices* based on the partition of the edges in the graph according to the equivalence and the magnitude of their edge connectivity values were derived.

The mean Randić connectivity index (or mean Randić branching index) is defined as

$$\bar{\chi}_R = \frac{\chi_R}{B}$$

where B is the number of edges in the molecular graph.

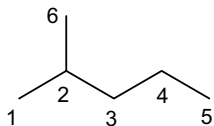
A variant of the Randić connectivity index was also proposed as

$$\chi'_R = A \cdot \chi_R$$

where A is the number of graph vertices [Mihalić, Nikolić *et al.*, 1992].

Example C15

Connectivity indices of order 0, 1, and 2 for 2-methylpentane.



Atoms	1	2	3	4	5	6
δ_i	1	3	2	2	1	1

$${}^0\chi = \delta_1^{-1/2} + \delta_2^{-1/2} + \delta_3^{-1/2} + \delta_4^{-1/2} + \delta_5^{-1/2} + \delta_6^{-1/2} =$$

$$= 1^{-1/2} + 3^{-1/2} + 2^{-1/2} + 2^{-1/2} + 1^{-1/2} + 1^{-1/2} = 4.992$$

$${}^1\chi = (\delta_1 \times \delta_2)^{-1/2} + (\delta_2 \times \delta_3)^{-1/2} + (\delta_3 \times \delta_4)^{-1/2} + (\delta_4 \times \delta_5)^{-1/2} + (\delta_2 \times \delta_6)^{-1/2} =$$

$$= (1 \times 3)^{-1/2} + (3 \times 2)^{-1/2} + (2 \times 2)^{-1/2} + (2 \times 1)^{-1/2} + (3 \times 1)^{-1/2} = 2.770$$

$${}^2\chi = (\delta_1 \times \delta_2 \times \delta_3)^{-1/2} + (\delta_2 \times \delta_3 \times \delta_4)^{-1/2} + (\delta_3 \times \delta_4 \times \delta_5)^{-1/2} + (\delta_1 \times \delta_2 \times \delta_6)^{-1/2} +$$

$$+ (\delta_3 \times \delta_2 \times \delta_6)^{-1/2} =$$

$$= (1 \times 3 \times 2)^{-1/2} + (3 \times 2 \times 2)^{-1/2} + (2 \times 2 \times 1)^{-1/2} + (1 \times 3 \times 1)^{-1/2} + (2 \times 3 \times 1)^{-1/2}$$

$$= 2.183$$

Kier and Hall defined [Kier and Hall, 1986; Kier and Hall, 1977b] a general scheme based on the Randić index to calculate also zero-order and higher order descriptors; these are called **Molecular Connectivity Indices (MCIs)**, also known as **Kier–Hall connectivity indices**. They are calculated by the following:

$${}^0\chi = \sum_{i=1}^A \delta_i^{-1/2} \quad {}^1\chi = \sum_{b=1}^B (\delta_i \cdot \delta_j)_b^{-1/2} \quad {}^2\chi = \sum_{k=1}^{2P} (\delta_i \cdot \delta_l \cdot \delta_j)_k^{-1/2}$$

$${}^m\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i \right)_k^{-1/2}$$

where k runs over all of the m th order subgraphs constituted by n atoms ($n = m + 1$ for acyclic subgraphs); K is the total number of m th order subgraphs present in the molecular graph and in the case of the path subgraphs equals the m th order path count mP . The product is over the simple vertex degrees δ of all the vertices involved in each subgraph. The subscript “ t ” for the connectivity indices refers to the type of \rightarrow molecular subgraph and is “ ch ” for chain or ring, “ pc ” for path-cluster, “ c ” for cluster, and “ p ” for path (that can also be omitted). Obviously, the first-order Kier–Hall connectivity index is the Randić connectivity index.

By replacing the vertex degree δ by the \rightarrow *valence vertex degree* δ^v in the formulas reported above, similar **valence connectivity indices** were proposed [Kier and Hall, 1981, 1983b], denoted by ${}^m\chi_i^v$, able to account for the presence of heteroatoms in the molecule as well as double and triple bonds (Table C6).

Table C6 Values of the first-order Kier–Hall connectivity index for some substituent groups attached to a Carbon atom with a valence vertex degree equal to 3.

Substituent	${}^1\chi^v$	Substituent	${}^1\chi^v$	Substituent	${}^1\chi^v$
–H	0	–COOH	0.7164	–CH ₂ CH ₂ S CH ₃	1.762
–CH ₃	0.5773	–NH ₂	0.3333	–CH ₂ CH ₂ COOH	1.6900
–OH	0.2582	–CH ₂ OH	0.7240	–COH	0.5690

Analogously, **bond order-weighted vertex connectivity indices**, denoted by ${}^m\chi_i^b$, were also defined by using the \rightarrow *bond vertex degree* δ^b instead of the simple vertex degree δ to specifically account for multiplicity in the molecule. To derive these connectivity indices, either the \rightarrow *conventional bond order* or quantum-chemical derived \rightarrow *bond orders* can be used [Estrada and Montero, 1993; Estrada and Molina, 2001a; Jalbout and Li, 2003c]. Moreover, connectivity indices were also calculated using the \rightarrow *Z-delta number* δ^Z and therefore denoted by ${}^m\chi^Z$ [Pogliani, 1999b]. Another set of modified valence connectivity indices was proposed based on the \rightarrow *Li valence vertex degree* δ^{Li} , used in place of the original valence vertex degree [Li, Jalbout *et al.*, 2003].

The inverse-square-root function was selected for the Randić connectivity index, and later used for the most general connectivity indices, because it provided high correlation with properties of isomeric alkane series, thus showing high sensitivity to variation in molecular structure. However, it was observed that it is not a very effective connectivity measure in nonisomeric molecule series as it shows two opposing trends: to increase with molecular size and to decrease with \rightarrow *molecular complexity* [Bonchev, 2001a]. Substituting the inverse-square-root function in favor of the total adjacency function, the \rightarrow *overall connectivity indices* were proposed as a measure of topological complexity.

Table C7 Some connectivity and valence connectivity indices for the data set of phenethylamines (Appendix C– Set 2).

Mol.	X	Y	${}^0\chi$	${}^1\chi$	${}^2\chi$	${}^3\chi$	${}^4\chi$	${}^5\chi$	${}^0\chi^v$	${}^1\chi^v$	${}^2\chi^v$	${}^3\chi^v$	${}^4\chi^v$	${}^5\chi^v$
1	H	H	8.975	5.698	5.005	3.298	2.639	1.702	9.082	4.952	4.246	2.505	1.972	0.825
2	H	F	9.845	6.092	5.627	3.708	2.791	1.914	9.383	5.052	4.387	2.575	1.982	0.824
3	H	Cl	9.845	6.092	5.627	3.708	2.791	1.914	10.139	5.43	4.823	2.827	2.108	0.969
4	H	Br	9.845	6.092	5.627	3.708	2.791	1.914	10.969	5.845	5.302	3.104	2.246	1.129
5	H	I	9.845	6.092	5.627	3.708	2.791	1.914	11.541	6.131	5.632	3.295	2.342	1.239
6	H	Me	9.845	6.092	5.627	3.708	2.791	1.914	10.005	5.363	4.746	2.783	2.086	0.943
7	F	H	9.845	6.092	5.639	3.625	2.934	1.978	9.383	5.052	4.39	2.554	1.991	0.887
8	Cl	H	9.845	6.092	5.639	3.625	2.934	1.978	10.139	5.43	4.827	2.789	2.19	1.128
9	Br	H	9.845	6.092	5.639	3.625	2.934	1.978	10.969	5.845	5.306	3.047	2.408	1.393
10	I	H	9.845	6.092	5.639	3.625	2.934	1.978	11.541	6.131	5.636	3.225	2.559	1.575
11	Me	H	9.845	6.092	5.639	3.625	2.934	1.978	10.005	5.363	4.749	2.747	2.155	1.086

(Continued)

Table C7 (Continued)

Mol.	X	Y	$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^4\chi$	$^5\chi$	$^0\chi^r$	$^1\chi^r$	$^2\chi^r$	$^3\chi^r$	$^4\chi^r$	$^5\chi^r$
12	Cl	F	10.715	6.503	6.135	4.287	3.042	2.161	10.44	5.535	4.916	2.939	2.186	1.122
13	Br	F	10.715	6.503	6.135	4.287	3.042	2.161	11.27	5.95	5.363	3.257	2.394	1.381
14	Me	F	10.715	6.503	6.135	4.287	3.042	2.161	10.306	5.468	4.844	2.887	2.153	1.08
15	Cl	Cl	10.715	6.503	6.135	4.287	3.042	2.161	11.195	5.913	5.323	3.388	2.304	1.258
16	Br	Cl	10.715	6.503	6.135	4.287	3.042	2.161	12.026	6.328	5.77	3.863	2.511	1.517
17	Me	Cl	10.715	6.503	6.135	4.287	3.042	2.161	11.062	5.846	5.251	3.311	2.27	1.216
18	Cl	Br	10.715	6.503	6.135	4.287	3.042	2.161	12.026	6.328	5.77	3.882	2.433	1.407
19	Br	Br	10.715	6.503	6.135	4.287	3.042	2.161	12.856	6.743	6.217	4.529	2.64	1.666
20	Me	Br	10.715	6.503	6.135	4.287	3.042	2.161	11.892	6.261	5.698	3.777	2.399	1.365
21	Me	Me	10.715	6.503	6.135	4.287	3.042	2.161	10.928	5.78	5.179	3.236	2.249	1.192
22	Br	Me	10.715	6.503	6.135	4.287	3.042	2.161	11.892	6.261	5.698	3.756	2.49	1.493

Connectivity-like indices are molecular descriptors calculated applying the same mathematical formula as the connectivity indices, but substituting the vertex degree δ with any \rightarrow *local vertex invariant* (LOVI):

$${}^m\text{Chi}_t(\mathcal{L}) = \sum_{k=1}^K \left(\prod_{i=1}^n \mathcal{L}_i \right)_k^{-1/2}$$

where \mathcal{L}_i is the general symbol for local vertex invariants, the summation goes over all the subgraphs of type t constituted by n atoms and m edges; K is the total number of such m th order subgraphs present in the molecular graph, and each subgraph is weighted by the product of the local invariants associated to the vertices contained in the subgraph. Connectivity-like indices may also be calculated by replacing local vertex invariants \mathcal{L}_i with \rightarrow *atomic properties* P_i .

The general formula for the calculation of connectivity-like indices, which uses the row sums VS_i of a \rightarrow *graph-theoretical matrix* as the local vertex invariants, was called by Ivanciuc **Chi operator** [Ivanciuc, Ivanciuc *et al.*, 1997; Ivanciuc, 2001c]. Specifically, for any square symmetric ($A \times A$) matrix $\mathbf{M}(w)$ representing a molecular graph with A vertices and a \rightarrow *weighting scheme* w , the *Chi operator* is defined as

$${}^m\text{Chi}(\mathbf{M}; w) = \sum_{k=1}^K \left(\prod_{i=1}^n VS_i(\mathbf{M}, w) \right)_k^{-1/2}$$

where VS_i indicates the \rightarrow *row sum operator*.

Randić-like indices are connectivity-like indices defined for graph edges and calculated by using the same mathematical formula as the \rightarrow *Randić connectivity index* ${}^1\chi$, but replacing the vertex degree δ with any \rightarrow *local vertex invariants* \mathcal{L} :

$${}^1\chi(\mathcal{L}) = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\mathcal{L}_i \cdot \mathcal{L}_j)^{-1/2} = \sum_{b=1}^B (\mathcal{L}_{b(1)} \cdot \mathcal{L}_{b(2)})_b^{-1/2}$$

where A and B are the total number of vertices and edges in the graph, respectively; a_{ij} are the elements of the \rightarrow *adjacency matrix* equal to one for pairs of adjacent vertices, and zero otherwise; the subscripts $b(1)$ and $b(2)$ represent the two vertices connected by the edge b . Note

that, in the left expression, the summation goes over all pairs of vertices in the graph but the only nonvanishing contributions are from the pairs of adjacent vertices for which elements a_{ij} equal one. Several mathematical properties of Randić-like indices were investigated [Gutman, 2002a; Li and Gutman, 2006].

Moreover, **generalized connectivity indices** are a generalization of the Kier–Hall connectivity indices in terms of a variable exponent λ as:

$${}^m\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i \right)_k^\lambda$$

where λ is any real exponent. If $\lambda = 1$ and $m = 1$, the \rightarrow *second Zagreb index* M_2 is obtained; values of $\lambda = -1$ and $\lambda = 1/2$ were considered by Altenburg [Altenburg, 1980], and values of $\lambda = -1/3$ and $\lambda = -1/4$ were also investigated [Randić, Hansen *et al.*, 1988; Estrada, 1995c; Amić, Beslo *et al.*, 1998; Ivanciuc, Ivanciuc *et al.*, 2002e].

Related to Randić-like indices are the \rightarrow *Balaban-like indices*, which only differ for the normalization factor.

Some connectivity-like indices are reported below. Other connectivity-like indices reported elsewhere are \rightarrow *JJ indices* derived from the \rightarrow *Wiener matrix*, \rightarrow *electronegativity-based connectivity indices*, \rightarrow *extended edge connectivity indices*, \rightarrow *chiral connectivity indices*, \rightarrow *variable connectivity indices*, \rightarrow *line graph connectivity indices*, and \rightarrow *line graph Randić connectivity index*.

• Evans extended connectivity indices

Two molecular descriptors proposed to generalize the Randić connectivity index, defined as [Evans, Lynch *et al.*, 1978]

$$\chi_2^{ext} = \sum_b [(^2f_i n_i \cdot ^2f_j n_j)_b \cdot \pi_b^*]^{-1/2} \quad \text{and} \quad \chi_3^{ext} = \sum_b [(^3f_i n_i \cdot ^3f_j n_j)_b \cdot \pi_b^*]^{-1/2}$$

where i and j are indices for the two adjacent vertices incident to the edge b , n_i is an integer describing the i th atom type, 2f_i and 3f_i are the second and third order \rightarrow *vertex distance counts* of the i th vertex, that is, the number of the vertices at topological distances 2 and 3 from the i th vertex, respectively. π_b^* is the \rightarrow *conventional bond order* and the summation runs over all edges.

• environment connectivity descriptors

These are Randić-like indices of molecular fragments calculated on fragment atoms and then first neighbors [Jurs, Chou *et al.*, 1979]. The value of the Randić connectivity index for a given fragment represents the immediate surroundings of the substructure as embedded within the molecule. If the fragment is not present in the molecule, zero value is given.

Environment descriptors are closely related to \rightarrow *substructure descriptors*, differing from the latter in using real values in place of binary variables or counts. The set of fragments is defined by the user depending on the data set and the specific problem.

• Fragment Molecular Connectivity indices (FMC)

These are first-order connectivity indices computed for predefined positions on molecular fragments in congeneric series [Takahashi, Miashita *et al.*, 1985]. By superimposition of all congeneric compounds, a template structure is derived whose vertices define the positions for the *FMC* indices; the vertices of the common parent structure are not considered in defining the

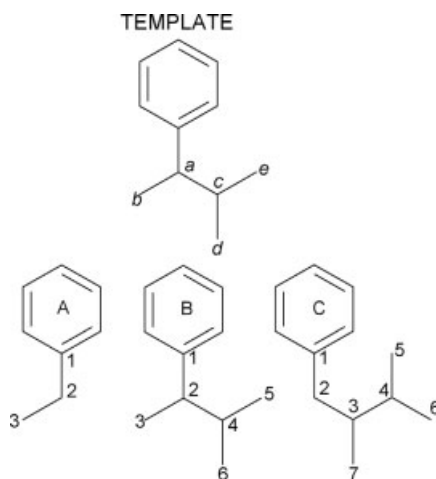
positions. For a k th position the corresponding fragment connectivity index is defined as

$$FMC_k = \sum_i (\delta_k \cdot \delta_i)^{-1/2}$$

where δ can be the simple \rightarrow *vertex degree* or the \rightarrow *valence vertex degree*, and i denotes each vertex joint to the vertex in the k th position (the link vertex of the parent molecule is also considered). By definition, FMC is equal to zero if there is no atom of the substituent in the considered position. Each molecule is finally described by a number of FMC values, corresponding to the number of predefined positions.

Example C16

Fragment molecular connectivity indices.



$$FMC_a(A) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} = (2 \cdot 3)^{-1/2} + (2 \cdot 1)^{-1/2} = 1.115$$

$$FMC_b(A) = (\delta_3 \cdot \delta_2)^{-1/2} = (1 \cdot 2)^{-1/2} = 0.707 \quad FMC_c(A) = FMC_d(A) = FMC_e(A) = 0$$

$$FMC_a(B) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} + (\delta_2 \cdot \delta_4)^{-1/2} = (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} + (3 \cdot 3)^{-1/2} = 1.244$$

$$FMC_b(B) = (\delta_3 \cdot \delta_2)^{-1/2} = (1 \cdot 3)^{-1/2} = 0.577 \quad FMC_d(B) = FMC_e(B) = 0.577$$

$$FMC_c(B) = (\delta_4 \cdot \delta_2)^{-1/2} + (\delta_4 \cdot \delta_5)^{-1/2} + (\delta_4 \cdot \delta_6)^{-1/2} = (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} + (3 \cdot 1)^{-1/2} = 1.488$$

$$FMC_a(C) = (\delta_2 \cdot \delta_1)^{-1/2} + (\delta_2 \cdot \delta_3)^{-1/2} = (2 \cdot 3)^{-1/2} + (2 \cdot 3)^{-1/2} = 0.816$$

$$FMC_c(C) = (\delta_3 \cdot \delta_2)^{-1/2} + (\delta_3 \cdot \delta_4)^{-1/2} + (\delta_3 \cdot \delta_7)^{-1/2} = (3 \cdot 2)^{-1/2} + (3 \cdot 3)^{-1/2} + (3 \cdot 1)^{-1/2} = 1.319$$

Molecule	FMC_a	FMC_b	FMC_c	FMC_d	FMC_e
A	1.115	0.707	0	0	0
B	1.244	0.577	1.488	0.577	0.577
C	0.816	0	1.319	0.577	1.488

• walk connectivity indices

These Randić-like indices are defined by using the \rightarrow atomic walk counts as the local vertex invariants in place of the vertex degrees and applying the Randić-type formula as [Razinger, 1986]

$$\chi^W = \sum_{b=1}^B (awcs_i \cdot awcs_j)_b^{-1/2} \quad \chi^{kW} = \sum_{b=1}^B (awc_i^{(k)} \cdot awc_j^{(k)})_b^{-1/2}$$

where the summations run over all edges in the H-depleted molecular graph and the subscripts i and j refer to the two vertices incident to the considered edge; the first index χ^W is calculated from the \rightarrow atomic walk count sum $awcs$, that is, considering all walks of any length from the vertex, while the second index χ^{kW} is calculated counting only the walks of length k from each vertex ($awc^{(k)}$). In particular, the **longest walk connectivity index** χ^{LW} was proposed as a highly discriminant descriptor, defined as

$$\chi^{LW} = \sum_{b=1}^B (awc_i^{(A-1)} \cdot awc_j^{(A-1)})_b^{-1/2}$$

where only the longest walks of length $A-1$ from each vertex are counted, A being the total number of graph vertices.

The **Randić–Razinger index** χ_i^{kW} is a \rightarrow local vertex invariant, defined as [Diudea, Minailiuc *et al.*, 1997a]

$$\chi_i^{kW} = \sum_{j=1}^A a_{ij} (awc_i^{(k)} \cdot awc_j^{(k)})^{-1/2}$$

where $awc_i^{(k)}$ and $awc_j^{(k)}$ are the \rightarrow atomic walk counts of order k for vertices v_i and v_j ; the summation goes over all the vertices, but accounts only for contributions from vertices adjacent to v_i , a_{ij} being the elements of the adjacency matrix. It can be noted that the sum of these LOVIs over all the vertices corresponds to twice the corresponding walk connectivity index:

$$2 \cdot \chi^{kW} = \sum_{i=1}^A \chi_i^{kW}$$

• Kupchik modified connectivity indices

These are modifications of the Randić connectivity index defined in such a way as to account for the presence of heteroatoms in the molecule [Kupchik, 1986, 1988, 1989]:

$${}^1\chi^r = \sum_{b=1}^B (\delta_i^{\text{het}} \cdot \delta_j^{\text{het}})_b^{-1/2} \quad \text{and} \quad {}^1\chi^b = \sum_{b=1}^B \frac{r_{ij}}{r_{CC}} \cdot (\delta_i \cdot \delta_j)_b^{-1/2}$$

where the summations run over all the edges in the molecular graph and i, j denote the vertices incident with the considered edge; r_{ij} is the bond length and r_{CC} a standard carbon–carbon bond

length (1.54 Å); δ is the simple \rightarrow vertex degree, that is, the number of first neighbors. The \rightarrow Kupchik vertex degree δ^{het} is calculated as

$$\delta_i^{\text{het}} = \frac{R_C}{R_i} \cdot (Z_i^v - h_i)$$

where R_i and R_C are the covalent radius of the i th atom and the carbon atom, respectively; Z_i^v is the atomic number and h_i the number of hydrogen atoms bonded to i th vertex.

The first index was later called **radius-corrected connectivity index** and the second one **bond-length-corrected connectivity index** [Sun, Huang *et al.*, 1996].

These modified connectivity indices were found to be related to the \rightarrow molar refractivity of alkanes, alkylsilanes, and alkylgermanes.

- **perturbation connectivity indices** (${}^m\chi_i^p$)

These are connectivity-like indices based on the \rightarrow perturbation delta value δ^p and defined as [Gombar, Kumar *et al.*, 1987]

$${}^m\chi_t^p = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i^p \right)_k^{-1/2}$$

where

$$\delta_i^p = \delta_i^v + \sum_{j=1}^A a_{ij} \cdot \gamma_{ij} \cdot \delta_j^v$$

is the perturbation delta value. The summation in the first formula goes over all of the m th order subgraphs of type t containing n atoms; K is the total number of m th order subgraphs; a_{ij} are the elements of the \rightarrow adjacency matrix equal to one for adjacent vertices and otherwise zero. Perturbation delta values are obtained from \rightarrow valence vertex degrees δ^v modified by atomic environment.

- **3D-connectivity indices** (${}^m\chi\chi_t$)

These are connectivity-like indices derived from the \rightarrow geometry matrix G ; they are defined using the \rightarrow geometric distance degree ${}^G\sigma$ in place of the topological vertex degree δ [Randić, 1988a, 1988b; Randić, Jerman-Blazic *et al.*, 1990]:

$${}^m\chi\chi_t = \sum_{k=1}^K \left(\prod_{i=1}^n {}^G\sigma_i \right)_k^{-1/2}$$

where k runs over all of the m th order subgraphs constituted by n vertices; K is the total number of m th order subgraphs. The subscript “ t ” refers to the type of molecular subgraph.

- **total structure connectivity index**

This is an extremal connectivity index contemporarily accounting for all the vertices in the graph as [Needham, Wei *et al.*, 1988]

$$\chi_T = \left(\prod_{i=1}^A \delta_i \right)^{-1/2}$$

Note that the total structure connectivity index is the square root of the \rightarrow *simple topological index* proposed by Narumi for measuring molecular branching.

• **local connectivity indices** (${}^m\bar{\chi}_i$) (\equiv *atomic connectivity indices*)

Computed for individual vertices in a graph, they were developed by equally partitioning each term ${}^mw_{ij} = (\delta_i \cdot \delta_k \cdot \dots \cdot \delta_j)^{-1/2}$ of the connectivity indices among all of the vertices along the path i - j of m th order, as [Balaban, Catana *et al.*, 1990]

$${}^m\bar{\chi}_i = \frac{1}{m+1} \cdot \sum_{j=1} {}^mP_{ij} {}^mw_{ij}$$

where ${}^mw_{ij}$ is the \rightarrow *path connectivity*, the index j represents the terminal vertex v_j of a path and the summation runs over all the paths ${}^mP_{ij}$ of length m starting from vertex v_i ; $m+1$ is the number of vertices along each path of length m .

For example, the first-order local connectivity index for the i th vertex is defined as

$${}^1\bar{\chi}_i = \frac{1}{2} \cdot \sum_{j=1}^{\delta_i} {}^1w_{ij}$$

where δ_i is the vertex degree of the i th vertex, that is, the number of edges incident to the i th vertex. The zero-order local connectivity index is simply defined as

$${}^0\bar{\chi}_i = (\delta_i)^{-1/2}$$

By summing these local connectivity indices over all the nonhydrogen atoms, the Kier–Hall connectivity indices are reproduced.

• **H₁ topological index**

This is a Randić-like index defined as [Li and You, 1993a, 1993b; Li Zhang *et al.*, 1995]

$$\begin{aligned} H_1 &= \left(\sum_{b=1}^B \frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)^2 = \\ &= \sum_{b=1}^B \left(\frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)^2 + \sum_{b=1}^{B-1} \sum_{b'=b+1}^B \left(\frac{1}{(1+\Delta_b)\sqrt{\delta_i \cdot \delta_j}} \right)_b \cdot \left(\frac{1}{(1+\Delta_{b'})\sqrt{\delta_i \cdot \delta_j}} \right)_{b'} \end{aligned}$$

where the summations run over all B edges in the \rightarrow *H-filled molecular graph*; δ_i and δ_j are the \rightarrow *vertex degree* of the two vertices incident to the considered b th edge. Δ_b is a bond parameter representing the interaction between the two bonded vertices i and j and calculated as the following:

$$\Delta_b = \alpha \cdot (\text{IP}_i - \text{EA}_j)_b + (1-\alpha) \cdot (\text{IP}_i - \text{EA}_j)_b$$

where IP and EA are the \rightarrow *ionization potential* and \rightarrow *electron affinity*, respectively. The first term in the equation represents the electron transfer interaction from the HOAO (Highest Occupied Atomic Orbital) of the i th atom to the LUAO (Lowest Unoccupied Atomic Orbital) of the j th atom, the second term represents the feedback interaction from the HOAO of the j th

atom to LUAO of the i th atom. The parameter α is used to modulate the importance of the two kinds of interaction; it is generally taken equal to 0.5.

- **modified Randić index** (${}^1\chi_{\text{mod}}$)

This is a molecular descriptor based on atomic properties, accounting for valence electrons and connectivities in the H-depleted molecular graph, calculated by using a Randić-like formula [Lohninger, 1993]:

$${}^1\chi_{\text{mod}} = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \frac{Z_i}{\sqrt{\delta_i \cdot \delta_j}}$$

where the summation goes over all the pairs of vertices in the molecular graph; the only nonvanishing terms are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the \rightarrow adjacency matrix; δ is the \rightarrow vertex degree and Z the atomic number.

- **charge-weighted vertex connectivity indices** (${}^m\Omega_i(q)$)

Very similar to the \rightarrow electronic-topological descriptors, charge-weighted vertex connectivity indices are connectivity-like indices obtained by replacing the simple vertex degree with a charge-related atomic quantity calculated by \rightarrow computational chemistry [Estrada and Montero, 1993; Estrada, 1995d; Estrada and Molina, 2001a]:

$${}^m\Omega_i(q) = \sum_{k=1}^K \left(\prod_{i=1}^n \delta_i(q) \right)_k^{-1/2}$$

where $\delta_i(q) = q_i - h_i$ is the **electron charge density weight**, q_i being the electron charge density on the i th atom and h_i the number of hydrogen atoms bonded to it. Another set of connectivity-like indices, called **corrected charge-weighted vertex connectivity indices**, and denoted by ${}^m\Omega_i^c(q)$, was also defined based on a local vertex invariant corrected for hydrogen atomic charges as

$$\delta_i^c(q) = q_i - \sum_j q_j^H$$

where q_j^H is the electron charge density of j th hydrogen atom bonded to the i th atom. This quantity was called **corrected electron charge density weight**.

- **atomic molecular connectivity index** (χ^c) (\equiv molecular connectivity topochemical index)

This index was designed as an extension of the Randić connectivity index to take into account the relative size of heteroatoms in a H-depleted molecular graph. It is based on the \rightarrow Madan vertex degree derived from the \rightarrow chemical adjacency matrix [Goel and Madan, 1995]:

$$\chi^c = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^c \cdot \delta_j^c)^{-1/2}$$

where the only nonvanishing terms in the summation are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the \rightarrow adjacency matrix; δ^c is the Madan vertex

degree, calculated by summing up relative atomic weights of all the adjacent atoms, assuming the carbon atom weight as the reference.

This index was applied in a structure-activity studies on anti-inflammatory activity of pyrazole carboxylic acid hydrazide analogs. It was demonstrated that, despite the overall classification accuracy was above 80%, this descriptor did not perform better than the popular valence connectivity index.

- **Euclidean connectivity index (χ^E)**

Derived from the \rightarrow geometry matrix G , it is defined by using a Randić-like formula applied to \rightarrow geometric distance degrees ${}^G\sigma$ used in place of the topological vertex degrees δ [Balasubramanian, 1995b]:

$$\chi^E = \sum_{i=1}^{A-1} \sum_{j=i+1}^A ({}^G\sigma_i \cdot {}^G\sigma_j)^{-1/2}$$

This index discriminates the geometrical isomers and can be considered as a measure of the compactness of a molecule in the 3D space. Note that all possible atom pairs are considered instead of the pairs of bonded atoms because in 3D space there exists an Euclidean distance between every pair of atoms.

- **local Balaban index**

This is a local vertex invariant, denoted as J_i and defined by using a Randić-like formula [Diudea, Minailiuc *et al.*, 1997a]:

$$J_i = \sum_{j=1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^{-1/2}$$

where σ_i and σ_j are the \rightarrow vertex distance degrees of vertices v_i and v_j ; the summation goes over all the vertices, but accounts only for contributions from vertices adjacent to v_i , a_{ij} being the elements of the adjacency matrix. The sum of the local Balaban index over all the graph vertices is related to the \rightarrow Balaban distance connectivity index J by the following relation:

$$\frac{2 \cdot J \cdot (C + 1)}{B} = \sum_{i=1}^A J_i$$

where C and B are the \rightarrow cyclomatic number and the number of edges, respectively.

- **solvation connectivity indices (${}^m\chi^s$)**

These are molecular descriptors defined to model solvation entropy and describe dispersion interactions in solution [Zefirov and Palyulin, 2001]. Taking into account the characteristic dimension of the molecules by atomic parameters, they are defined as

$${}^m\chi^s = \frac{1}{2^{m+1}} \cdot \sum_{k=1}^K \frac{\left(\prod_{i=1}^n L_i \right)_k}{\left(\prod_{i=1}^n \delta_i \right)_k^{1/2}}$$

where L is the principal quantum number (2 for C, N, O atoms; 3 for Si, S, Cl, . . .) associated to a vertex in the k th subgraph and δ the corresponding \rightarrow *vertex degree*; K is the total number of m th order subgraphs; n is the number of vertices in the subgraph. The normalization factor $1/(2^{m+1})$ is defined in such a way that the indices ${}^m\chi$ and ${}^m\chi^s$ for compounds containing only second-row atoms coincide.

For example, the first-order solvation connectivity index is

$${}^1\chi^s = \frac{1}{4} \cdot \sum_{b=1}^B \frac{(L_i \cdot L_j)_b}{(\delta_i \cdot \delta_j)_b^{1/2}}$$

where the summation goes over all the B edges; L_i and L_j are the principal quantum numbers of the two vertices incident to the considered edge. This index coincides with the Randić connectivity index ${}^1\chi$ for the hydrocarbons, being $L = 2$ for all the atoms.

These molecular descriptors are defined for a \rightarrow *H-depleted molecular graph*; furthermore, fluorine atoms are not included in the graph, their dimension being very close to that of the hydrogen atom.


• Yang connectivity index (${}^1\chi^Y$)

This is a Randić-like index based on the \rightarrow *Yang vertex degree* [Jiang, Liu *et al.*, 2003]. It is defined as

$${}^1\chi^Y = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\delta_i^Y \cdot \delta_j^Y)^{-1/2}$$

where δ^Y is the vertex degree based on the \rightarrow *Yang's electronegativity force gauge*; the only nonvanishing terms in the summation are those corresponding to pairs of adjacent vertices, a_{ij} being the elements of the \rightarrow *adjacency matrix*;

A widespread use of connectivity descriptors in modeling a lot of molecular properties is recognizable in the literature since 1975.

 Additional references are collected in the thematic bibliography (see Introduction).

- **connectivity-like indices** \rightarrow connectivity indices
- **connectivity matrix** \equiv *atom connectivity matrix* \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **connectivity quotients** \rightarrow combined descriptors
- **connectivity table** \rightarrow molecular geometry
- **connectivity valence layer matrix** \rightarrow layer matrices
- **Connolly surface area** \rightarrow molecular surface (\odot solvent-accessible molecular surface)
- **consecutive AT numbers** \rightarrow vertex degree
- **consensus analysis** \rightarrow structure/response correlations
- **consensus binding free energy** \rightarrow scoring functions
- **consensus fingerprint** \rightarrow substructure descriptors (\odot structural keys)
- **constant interval reciprocal indices** \rightarrow distance matrix
- **constant and near-constant variables** \rightarrow variable reduction

■ constitutional descriptors

These are the most simple and commonly used descriptors, reflecting the molecular composition of a compound without any information about its → *molecular geometry* or topology.

The most common constitutional descriptors are number of atoms (→ *atom number*), number of bonds (→ *bond number*), absolute and relative numbers of specific atom-types (→ *count descriptors*), absolute and relative numbers of single, double, triple, and aromatic bonds, number of rings (→ *cyclomatic number*), number of rings divided by the number of atoms or bonds, number of benzene rings, number of benzene rings divided by the number of atoms, → *molecular weight* and → *average molecular weight*, → *atomic composition indices*, → *information index on size*, etc.

These descriptors are insensitive to any conformational change, do not distinguish among isomers, and are either → *0D-descriptors* or → *1D-descriptors*.

- **constitutional graph** ≡ *molecular graph*
- **contact surface** → molecular surface (⊙ solvent-accessible molecular surface)
- **contingency coefficient** → statistical indices (⊙ concentration indices)
- **Continuous Chirality Measure** → chirality descriptors
- **continuous wavelet transforms** → spectra descriptors
- **contour length** → size descriptors (⊙ Kuhn length)
- **contour profiles** → molecular profiles
- **conventional bond order** → bond order indices
- **conventional bond order ID number** → ID numbers
- **core count** → ETA indices
- **Corey–Pauling–Koltun volume** → volume descriptors
- **corrected charge-weighted vertex connectivity indices** → connectivity indices (⊙ charge-weighted vertex connectivity indices)
- **corrected electron charge density weight** → connectivity indices (⊙ charge-weighted vertex connectivity indices)
- **corrected second moments** ≡ *topological atomic valencies* → self-returning walk counts
- **corrected structure count** ≡ *algebraic structure count* → Kekulé number
- **corrected Taft steric constant** → steric descriptors (⊙ Taft steric constant)
- **correlation distance** → similarity/diversity (Table S7)
- **correlation matrix** → statistical indices (⊙ correlation measures)
- **correlation measures** → statistical indices
- **correlation weights** → variable descriptors
- **Correlation Weights of the Local Invariants of Molecular Graphs** → variable descriptors

■ CoRSA (≡ *Comparative Receptor Surface Analysis*)

CoRSA is a 3D QSAR approach applied to compute structure-activity models whenever the structure of the biological target is not known [Ivanciuc, Ivanciuc *et al.*, 2000a, 2000b; Hirashima, Kuwano *et al.*, 2003]. Using the common steric and electrostatic features of the most active members of a series of compounds, CoRSA generates a virtual receptor model, represented as points on a surface complementary to the van der Waals surface of the set of compounds. The structural descriptors of the model are represented by the total interaction energies between each surface point of the virtual receptor and all atoms in a molecule. These descriptors are used in a Partial Least Squares (PLS) regression to generate a structure-activity model.

The development of a CoRSA model consists of the following seven steps. (1) The geometry of all molecules in the data set is optimized with molecular mechanics or quantum mechanics methods. (2) All optimized molecules are aligned (superimposed) using some pharmacophore hypothesis. The CoRSA model depends on the molecule alignment and errors in this step may lead to models that have a low predictive power. (3) A subset of the most active molecules is selected to generate the virtual receptor model; these compounds form the receptor generation set (RGS) of molecules. The central assumption is the complementarity between the shape and properties of these molecules and the virtual receptor. (4) The virtual receptor is generated using information on the geometry, volume, atomic charges, hydrophobicity, hydrogen-bonding, or other properties of the selected molecules. Unlike real receptors, the virtual receptor is not formed by atoms, but by a three-dimensional receptor surface represented by points having certain properties. The coordinates of these points are generated from the shape field of the RGS molecules.

Two field functions are used to create the shape of the virtual receptor, namely the van der Waals field function and the Wyvill field function. Each field source corresponds to an atom. The van der Waals field function generated by the atom i at distance r_{ij} is

$$V_i^{vdw} = r_{ij} - R_i^{vdw}$$

where r_{ij} is the distance from the atom i to the grid point j and R_i^{vdw} is the van der Waals radius of the atom i . This field function, which is computed for every grid point, has the property that inside the van der Waals volume the value is negative, outside the volume the value is positive, and at the van der Waals surface the value $V(r)$ is zero. If a grid point contains a shape field value computed for a different atom, the smaller of the two values is assigned to that grid point. The **Wyvill function** is a bounded function that decays completely in a finite distance R . The Wyvill field function generated by the atom i at distance r_{ij} is

$$V_i^{Wv} = 1 - \frac{4 \cdot r_{ij}^6}{9 \cdot R^6} + \frac{17 \cdot r_{ij}^4}{9 \cdot R^4} - \frac{22 \cdot r_{ij}^2}{9 \cdot R^2}$$

where r_{ij} is the distance from the atom i to the grid point j . A field value is the sum of the field values contributed by each atom; if a grid point is outside of R , its shape field value is not computed. The value of R depends on the atom type, and usually it is twice the van der Waals radius of the atom i . The Wyvill function has the properties that $V(0) = 1$, $V(R) = 0$, and $V(R/2) = 1/2$.

Using the shape field values the marching cubes isosurface algorithm produces a set of triangulated surface points representing the surface of the virtual receptor. The default grid spacing of 0.5 Å yields an average surface density of 6 points/Å². This gives an average distance between neighboring points (points in the same triangle) of about 0.47 Å.

(5) Each surface point from the virtual receptor contains information about the local properties of the receptor. These properties include electrostatic potential, partial charge, hydrophobicity, and hydrogen-bonding propensity. (6) With the virtual receptor model defined in steps (1)–(5), for each molecule in the data set, a number of molecular descriptors are derived by computing the ligand–receptor interaction energy between each surface point from the virtual receptor and the atoms in the molecule. (7) Finally, the molecular descriptors calculated for all the molecules are processed by PLS algorithm to generate the 3D-QSAR model.

- **CoSA** \equiv Comparative Spectral Analysis \rightarrow spectra descriptors
- **CoSASA** \equiv Comparative Structurally Assigned Spectral Analysis \rightarrow spectra descriptors
- **cosine similarity coefficient** \rightarrow similarity/diversity (Table S9)
- **cospectral graphs** \equiv *isospectral graphs* \rightarrow graph
- **Coulomb potential energy function** \rightarrow molecular interaction fields (\odot electrostatic interaction fields)

■ count descriptors

These are simple molecular descriptors based on counting the defined elements of a compound. The most common chemical count descriptors are \rightarrow *atom number* A , \rightarrow *bond number* B , \rightarrow *cyclomatic number* C , \rightarrow *hydrogen-bond acceptor number* and \rightarrow *hydrogen-bond donor number*, \rightarrow *distance-counting descriptors*, \rightarrow *path counts*, \rightarrow *walk counts*, \rightarrow *atom pairs*, and other related \rightarrow *substructure descriptors*.

When the different chemical nature of atoms is considered, the **atom-type count** is defined as the number of atoms of the same chemical element. $A \rightarrow$ *molecular graph* G can be characterized by a vector of atom-type counts as

$$\{N_C; N_H; N_O; N_N; N_S; N_F; N_{Cl}; N_{Br}; N_I; \dots\}$$

whose entries represent the number of carbon, hydrogen, oxygen, nitrogen, sulfur, fluorine, chlorine, bromine, and iodine atoms, respectively. These descriptors are derived from the chemical formula, that is, they are \rightarrow *OD-descriptors*. The **relative atom-type count** is the ratio between a given atom count and the total number A of atoms, therefore the following vector can be defined:

$$\{\bar{N}_C; \bar{N}_H; \bar{N}_O; \bar{N}_N; \bar{N}_S; \bar{N}_F; \bar{N}_{Cl}; \bar{N}_{Br}; \bar{N}_I; \dots\}$$

where

$$\bar{N}_X = \frac{N_X}{A}$$

The **atomistic topological indices** were proposed by Burden [Burden, 1996] as atom-type counts where each atom is classified by its element and the number of connections, thus also accounting for atom hybridization. In particular, N_p , N_s , N_t , and N_q are the number of primary, secondary, tertiary, and quaternary carbon atoms, respectively; N_{sp^3} , N_{sp^2} , and N_{sp} the numbers of sp^3 , sp^2 , and sp carbon atoms, respectively; N_{AR} the number of aromatic carbon atoms; and N_{Xsp^3} , N_{Xsp^2} , and N_{Xsp} the numbers of sp^3 , sp^2 , and sp heavy atoms, respectively.

Strictly related are the carbon-type counts called **STIMS indices (Simplest Topological Integers from Molecular Structures)** and proposed by Pal *et al.* [Pal, Sengupta *et al.*, 1988, 1989, 1990; Pal, Purkayastha *et al.*, 1992] as a subset of \rightarrow *TAU indices*. These are number of methyl carbon (N_P), number of methylene carbons (N_I), number of tertiary carbons (N_V), number of quaternary carbons (N_X), and number of branched carbons (N_B).

Count descriptors measuring the molecular unsaturation are within the \rightarrow *multiple bond descriptors*, such as the number of double bonds (DB), the number of triple bonds (TB), the number of aromatic bonds (AB), the number of rings (NRG), which is the \rightarrow *cyclomatic number* (denoted as C).

The **functional group count** can be defined considering the well-known *functional chemical groups*, which are groups of atoms having a characteristic and specific reactivity, such as

$$\{\text{NOH}; \text{NCOOH}; \text{NNH}_2; \text{NC=O}; \text{NOCH}_3; \text{NSH}; \text{NH}_2\text{C=}; \text{NBENZ}; \dots\}$$

whose entries represent the number of oxydryl, carboxylic, aminic, carbonilic, methoxy, thyo, methylen, and phenyl functional groups, respectively.

Andrews descriptors are particular atom and functional group counts relative to those groups found to be statistically significant in receptor binding modeling [Andrews, Craik *et al.*, 1984]: CO_2^- , PO_4^- , N^+ , N, OH, C=O, ether and thioether groups, halogens, sp^3 and sp^2 carbon atoms, and the \rightarrow *rotatable bond number*.

Even more general is the definition of **fragment count** as the number of a specific kind of *molecular fragments*, which are arbitrary-selected groups of adjacent atoms in a molecule. A general method for modeling \rightarrow *physico-chemical properties* using fragment counts is the \rightarrow *cluster expansion of chemical graphs*.

The **subgraph count set (SCS)** is a vectorial descriptor, where each entry is the number of times specific subgraphs are obtained by cutting one edge at a time in a \rightarrow *H-depleted molecular graph* [Oberrauch and Mazzanti, 1990]:

$$\{\text{NMETHYL}; \text{NETHYL}; \text{NPROPYL}; \text{NISOPROPYL}; \dots\}$$

The order of counts is not defined *a priori* and a subset of relevant subgraph counts can be used instead of the complete SCS. In chemical terms, these subgraphs are recognized as radicals.

Both the functional group count and the fragment count can be derived from recognized substructures within the molecule, that is, they are \rightarrow *1D-descriptors*; in fact they are also considered specific \rightarrow *substructure descriptors*.

Count descriptors give local chemical information, are insensitive to isomers, to conformational changes and show a high level of degeneracy. However, due to their immediate availability, they are among the most used descriptors.

Examples of count descriptors are reported by Feher and Schmidt [Feher and Schmidt, 2003]. Moreover, count descriptors are usually the basic molecular descriptors used to generate \rightarrow *property filters*. Examples of count descriptors used for property filters are those proposed by [Zheng, Luo *et al.*, 2005] which are listed in Table C8.

Table C8 Some of the descriptors proposed in [Zheng, Luo *et al.*, 2005].

Symbol	Definition
A3	Number of sp^3 hybridized C, O, S, and N atoms
UNC	Number of sp , sp^2 and aromatic carbons
AUH	Number of atoms rather than H and halogens
BDUH	Number of the bonds that do not contain H and halogen atoms
C3p	$\text{N}_{\text{sp}^3} / \text{AUH}$
UNC_C3	$\text{UNC} / \text{N}_{\text{sp}^3}$
A3_C	$\text{A3} / \text{N}_\text{C}$
h_p	Ratio of the number of hydrogen atoms over the total number of nonhalogen heavy atoms
NO_C3	$(\text{N}_\text{N} + \text{N}_\text{O}) / \text{N}_{\text{sp}^3}$

📖 [Chiorboli, Piazza *et al.*, 1993a, 1993b, 1993c, 1996; Tosato, Piazza *et al.*, 1992; Okey and Stensel, 1996; Okey, Stensel *et al.*, 1996; Winkler, Burden *et al.*, 1998; Kaiser and Niculescu, 1999; Tan and Siebert, 2004]

➤ **counter-propagation neural network** → Self-Organizing Maps

■ counting polynomials

The counting polynomial is a description of a graph property in terms of a sequence of numbers, such as the distance degree sequence or the sequence of the number of k independent edge sets [Hosoya, 1988, 1990; Trinajstić, 1992; Diudea, Gutman *et al.*, 2001; Noy, 2003; Diudea, Vizitiu *et al.*, 2007]. The counting polynomial is defined as

$$P(G; x) = \sum_k m(G; k) \cdot x^k$$

where the exponent k represents the extent of the considered graph partitions and the coefficients $m(G; k)$ are related to the frequency of the occurrences of partitions of extent k . Polynomial coefficients are graph invariants and are thus related to the structure of a molecule graph.

Examples of counting polynomials are → *Z-counting polynomial*, → *Wiener polynomial*, → *Cluj polynomials*, → *matching polynomial*, and → *omega polynomial*.

The **Altenburg polynomial** is another example of counting polynomials defined for → *H-depleted molecular graph* G as

$$\alpha(G, a) = \sum_{k=1}^D {}^k f \cdot a_k$$

where the sum runs over all the distances in the graph, D being the → *topological diameter*, that is, the maximum distance in the graph, ${}^k f$ the → *graph distance count* of k th order, that is, the number of distances equal to k in the graph, and a_k the independent variables [Altenburg, 1961]. The Altenburg polynomial is closely related to the → *Wiener index* of a graph: graphs with the same Altenburg polynomials always have just the same Wiener numbers (the contrary does not always hold).

In general, coefficients, roots, and derivatives of counting polynomials can be used for characterization of molecular graphs and as molecular descriptors in QSAR/QSPR modeling.

- **covalent hydrogen-bond acidity** → Theoretical Linear Solvation Energy Relationships
- **covalent hydrogen-bond basicity** → Theoretical Linear Solvation Energy Relationships
- **covariance** → statistical indices (⊙ correlation measures)
- **covariance matrix** → statistical indices (⊙ correlation measures)
- **CPK volume** ≡ *Corey-Pauling-Koltun volume* → volume descriptors
- **CPSA descriptors** ≡ *charged partial surface area descriptors*
- **Craig plot** → Hansch analysis
- **Cramer coefficient** → statistical indices (⊙ concentration indices)
- **critical constants** → physico-chemical properties
- **critical micelle concentration** → technological properties
- **critical packing parameter** → GRID-based QSAR techniques (⊙ VolSurf descriptors)
- **critical pressure** → physico-chemical properties (⊙ critical constants)

- **critical temperature** → physico-chemical properties (⊙ critical constants)
- **critical volume** → physico-chemical properties (⊙ critical constants)
- **crosscorrelation descriptors** → autocorrelation descriptors
- **cross-validated R^2** → regression parameters
- **cross-validation** → validation techniques
- **CSA2_{Cl} index** → charged partial surface area descriptors (⊙ HDCA index)
- **CSA2_H index** → charged partial surface area descriptors (⊙ HDCA index)
- **CT vertex degree** → vertex degree
- **cubic root molecular weight** → physico-chemical properties (⊙ molecular weight)
- **CWLIMG** \equiv *Correlation Weights of the Local Invariants of Molecular Graphs* → variable descriptors
- **cycle** \equiv *cyclic path* → graph
- **cycle-edge incidence matrix** → incidence matrices (⊙ cycle matrices)
- **cycle matrices** → incidence matrices
- **cycle-vertex incidence matrix** → incidence matrices (⊙ cycle matrices)
- **cyclicity** → graph
- **cyclicity index** → detour matrix
- **cyclicity indices** → molecular complexity (⊙ molecular cyclicity)
- **cyclic graph** → graph
- **cyclic path** → graph
- **cyclomatic number** → ring descriptors
- **Czekanowski similarity coefficient** \equiv *Dice similarity coefficient* → similarity/diversity (⊙ Table S9)