G

- **Galvez matrix** → topological charge indices
- **ganglia-augmented atom keys** → substructure descriptors
- **GAO descriptors** → Graph of Atomic Orbitals
- **GAO** ≡ *Graph of Atomic Orbitals*
- **GAP** ≡ *HOMO–LUMO energy gap* → quantum-chemical descriptors
- **gas–solvent partition coefficient** → physico-chemical properties (⊙ partition coefficients)
- **Geary coefficient** → autocorrelation descriptors
- **general a_N-index** → determinant-based descriptors
- **general distance–degree matrix** → distance–degree matrices
- **general free valence index** → quantum-chemical descriptors
- **general graph** → graph
- **General Interaction Properties Function approach** ≡ *GIPF approach*
- **generalized average graph energy** → spectral indices
- **generalized centric information indices** → centric indices
- **generalized cluster significance analysis** → variable selection (⊙ cluster significance analysis)
- **generalized complete centric index** → centric indices
- **generalized connectivity indices** → connectivity indices
- **generalized distance code centric index** → centric indices
- **generalized distance–degree centric index** → centric indices
- **generalized distance matrices** → distance matrix
- **generalized edge complete centric index** → centric indices
- **generalized edge distance code centric index** → centric indices
- **generalized edge distance degree centric index** → centric indices
- **generalized edge radial centric information index** → centric indices
- **generalized expanded Wiener numbers** → expanded distance matrices
- **generalized final prediction error criteria** → regression parameters
- **generalized graph center** → center of a graph
- **generalized graph energy** → spectral indices
- **generalized Hosoya indices** → Hosoya Z index
- **generalized Hosoya Z matrix** → Hosoya Z matrix
- **generalized hyper-Wiener indices** → Wiener matrix
- **generalized matrices** → matrices of molecules
- **generalized molecular-graph matrix** → variable descriptors

Molecular Descriptors for Chemoinformatics, Volume I: Alphabetical Listing
Roberto Todeschini and Viviana Consonni

ISBN: 978-3-527-31852-0

- **generalized radial centric information index** → centric indices
- **generalized reciprocal distance sum** → distance matrix
- **generalized reciprocal matrices** → matrices of molecules
- **Generalized Topological Distance Indices** → distance matrix
- **generalized topological indices** → variable descriptors
- **generalized vertex degree matrix** → vertex degree
- **generalized Wiener indices** → Wiener index
- **generalized Wiener matrix** → Wiener matrix
- **general solubility equation** → property filters (⊙ drug-like indices)
- **Generating Optimal Linear PLS Estimations** ≡ *GOLPE* → variable selection
- **genetic algorithm – variable subset selection** → variable selection
- **genetic function approximation** → variable selection
- **geodesic** → graph
- **geodesic matrix** → algebraic operators (⊙ sparse matrices)

■ **geometrical descriptors**

These are molecular descriptors defined in several different ways but always derived from the three-dimensional structure of the molecule [Ivanciuc, 2001a; Todeschini and Consonni, 2003]. In general, geometrical descriptors are calculated either from some optimized → *molecular geometry* obtained by the methods of the → *computational chemistry* or from crystallographic coordinates.

→ *Topographic indices* constitute a special subset of geometrical descriptors, being calculated on the graph representation of molecules but using the geometric distances between atoms instead of the topological distances.

Examples of geometrical descriptors (Figure G1) are the → *quantum-chemical descriptors*, → *moments of inertia*, → *length-to-breadth ratio*, → *surface areas*, → *volume descriptors*, → *CPSA descriptors*, → *EVA descriptors*, → *WHIM descriptors*, → *GETAWAY descriptors*, → *3D-MoRSE descriptors*, → *interaction energy values*, and → *spectrum-like descriptors*.

🕮 [Mihalić and Trinajstić, 1991; Tvaruzek and Komenda, 1991; Zhu and Klein, 1996; Basak, Gute *et al.*, 1997; Todeschini and Consonni, 2000]

- **geometrical eccentricity** → shape descriptors
- **geometrical representation** → molecular descriptors
- **geometrical shape coefficient** → shape descriptors (⊙ Petitjean shape indices)
- **geometric atom pair descriptors** → substructure descriptors
- **geometric binding property pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **geometric center** → center of a molecule
- **geometric diameter** → molecular geometry
- **geometric distance** → molecular geometry
- **geometric distance degree** → molecular geometry
- **geometric distance–detour distance combined matrix** → matrices of molecules (⊙ Table M3)
- **geometric distance/detour distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **geometric distance matrix** ≡ *geometry matrix* → molecular geometry

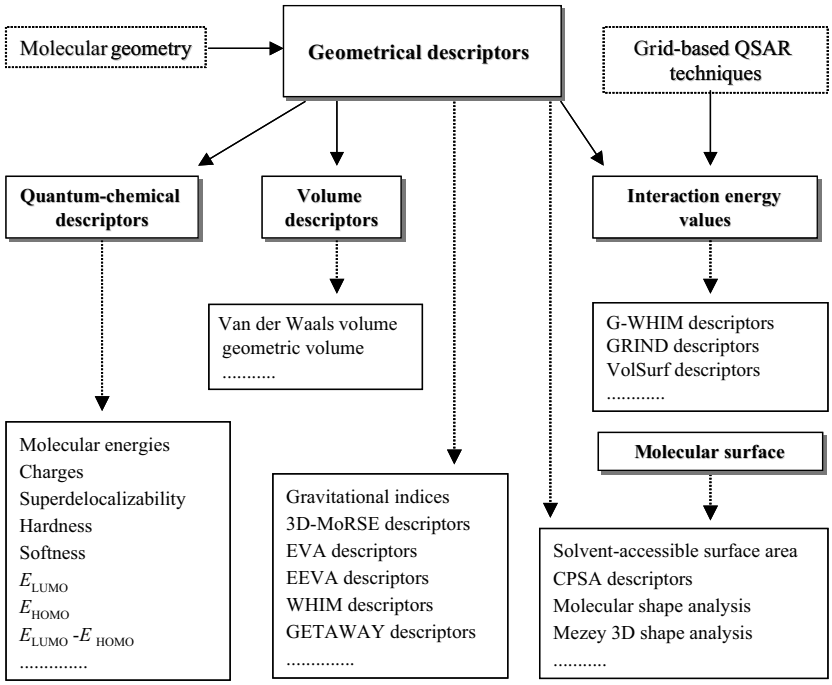


Figure G1 Scheme of the molecular descriptors classified as geometrical descriptors.

- **geometric distance–resistance distance combined matrix** → matrices of molecules (⊙ Table M3)
- **geometric distance/resistance distance quotient matrix** → matrices of molecules (⊙ Table M2)
- **geometric distance–topological distance combined matrix** → molecular geometry
- **geometric distance/topological distance quotient matrix** → molecular geometry
- **geometric eccentricity** → molecular geometry
- **geometric edge distance matrix** → edge distance matrix
- **geometric factors** → weighted matrices (⊙ weighted distance matrices)
- **geometric mean** → statistical indices (⊙ indices of central tendency)
- **geometric mean of the leverage magnitude** → GETAWAY descriptors
- **Geometric Mean Polarizability Effect Index** → electric polarization descriptors (⊙ Polarizability Effect Index)
- **Geometric Mean Polarizability Effect Index of π bond** → electric polarization descriptors (⊙ Polarizability Effect Index)
- **geometric modification number** → weighted matrices (⊙ weighted distance matrices)
- **geometric radius** → molecular geometry
- **geometric sum layer matrix** → layer matrices
- **geometric topological index** → vertex degree
- **geometric volume** → volume descriptors
- **geometry matrix** → molecular geometry
- **George-Foster criterion** → regression parameters (⊙ Table R1)

GETAWAY descriptors

GETAWAY (*GEometry, Topology, and Atom-Weights AssemblY*) descriptors are derived from the **Molecular Influence Matrix** (MIM), that is, a matrix representation of molecules denoted by **H** and defined as the following [Consonni, Todeschini *et al.*, 2002a, 2002b]:

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

where **M** is the → *molecular matrix* consisting of the centered Cartesian coordinates *x*, *y*, *z* of the molecule atoms (hydrogens included) in a chosen conformation. Atomic coordinates are assumed to be calculated with respect to the geometrical center of the molecule to obtain translational invariance. The molecular influence matrix is a symmetric $A \times A$ matrix, where A represents the number of atoms, and shows rotational invariance with respect to the molecule coordinates, thus resulting independent of molecule → *alignment rules*.

The diagonal elements h_{ii} of the molecular influence matrix, called *leverages* being the elements of the → *leverage matrix* defined in statistics, range from 0 to 1 and encode atomic information related to the "*influence*" of each molecule atom in determining the whole shape of the molecule; in effect, mantle atoms always have higher h_{ii} values than atoms near the molecule center. Moreover, the magnitude of the maximum leverage in a molecule depends on the size and shape of the molecule. As derived from the geometry of the molecule, leverage values are effectively sensitive to significant conformational changes and to the bond lengths that account for atom types and bond multiplicity.

Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interaction with the *i*th atom or, in other words, the attitude of the two considered atoms to interact with each other. A negative sign for the off-diagonal elements means that the two atoms occupy opposite molecular regions with respect to the center, hence the degree of their mutual accessibility should be low.

Combining the elements of the molecular influence matrix **H** with those of the → *geometry matrix* **G**, which encodes spatial relationships between pairs of atoms, another symmetric $A \times A$ molecular matrix, called **influence/distance matrix** and denoted by **R**, is derived as the following:

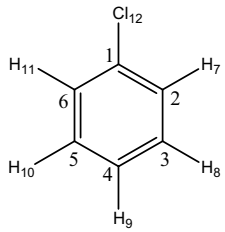
$$[\mathbf{R}]_{ij} \equiv \left[\frac{\sqrt{h_i \cdot h_j}}{r_{ij}}\right]_{ij} \qquad i \neq j$$

where h_i and h_j are the leverages of the atoms *i* and *j*, and r_{ij} is their geometric distance. The diagonal elements of the matrix **R** are zero, while each off-diagonal element *i*–*j*, resembling the single terms in the summation of the → *gravitational indices*, is calculated by the ratio of the geometric mean of the corresponding *i*th and *j*th diagonal elements of the matrix **H** over the interatomic distance r_{ij} provided by the geometry matrix **G**.

The square-root product of the leverages of two atoms is divided by their interatomic distance to make less significant contributions from pairs of atoms far apart, according to the basic idea that interaction between atoms in the molecule decreases as their distance increases. Obviously, the largest values of the matrix elements derive from the most external atoms (i.e., those with high leverages) and simultaneously next to each other in the molecular space (i.e., those having small interatomic distances).

Example G1

Hydrogen-filled molecular graph and molecular influence matrix of chlorobenzene, whose three-dimensional structure was optimized by minimizing the conformational energy.



Molecular Influence Matrix **H**

	C_1	C_2	C_3	C_4	C_5	C_6	H_7	H_8	H_9	H_{10}	H_{11}	Cl_{12}
C_1	**0.065**	0.031	−0.036	−0.070	−0.036	0.031	0.057	−0.063	−0.123	−0.063	0.057	0.148
C_2	0.031	**0.075**	0.042	−0.034	−0.077	−0.044	0.134	0.076	−0.059	−0.136	−0.079	0.071
C_3	−0.036	0.042	**0.079**	0.039	−0.039	−0.077	0.075	0.141	0.068	−0.071	−0.138	−0.082
C_4	−0.070	−0.034	0.039	**0.075**	0.039	−0.034	−0.061	0.067	0.132	0.067	−0.061	−0.159
C_5	−0.036	−0.077	−0.039	0.039	**0.079**	0.042	−0.138	−0.071	0.068	0.141	0.075	−0.082
C_6	0.031	−0.044	−0.077	−0.034	0.042	**0.075**	−0.079	−0.136	−0.059	0.076	0.134	0.071
H_7	0.057	0.134	0.075	−0.061	−0.138	−0.079	**0.242**	0.135	−0.108	−0.246	−0.141	0.130
H_8	−0.063	0.076	0.141	0.067	−0.071	−0.136	0.135	**0.250**	0.118	−0.129	−0.246	−0.143
H_9	−0.123	−0.059	0.068	0.132	0.068	−0.059	−0.108	0.118	**0.232**	0.118	−0.108	−0.280
H_{10}	−0.063	−0.136	−0.071	0.067	0.141	0.076	−0.246	−0.129	0.118	**0.250**	0.135	−0.143
H_{11}	0.057	−0.079	−0.138	−0.061	0.075	0.134	−0.141	−0.246	−0.108	0.135	**0.242**	0.130
Cl_{12}	0.148	0.071	−0.082	−0.159	−0.082	0.071	0.130	−0.143	−0.280	−0.143	0.130	**0.337**

It can be noted that the outer atoms (Cl and hydrogens) have larger leverage values (0.337, 0.242, 0.250, 0.232) than the carbon atoms of the aromatic ring (0.065, 0.075, 0.079). Then, among the outer atoms, the chlorine atom has the largest value (0.337), with its bond length larger than the bond distances of hydrogens. It must also be noted that equal leverage values are obtained for symmetric atoms, such as (C_2, C_6), (C_3, C_5), (H_7, H_{11}), and (H_8, H_{10}). Moreover, the off-diagonal terms give, to some extent, information on the relative spatial position of pairs of atoms. For instance, atoms C_1, C_2, C_6, H_7 and H_{11} have positive off-diagonal values with respect to the chlorine atom and, among these, C_1 has the largest value being the nearest one.

A set of the GETAWAY descriptors (H_{GM}, I_{TH}, I_{SH}, *HIC*, *RARS*, *RCON*, and *REIG*) was derived by applying some traditional matrix operators and concepts of information theory both to the molecular influence matrix **H** and to the influence/distance matrix **R**. Most of these descriptors are simply calculated only by the leverages used as the atomic weightings.

The **geometric mean of the leverage magnitude** (H_{GM}) is defined as

$$H_{GM} = 100 \cdot \left(\prod_{i=1}^{A} h_i \right)^{1/A}$$

where A is the number of atoms and h_i the leverage of the ith atom. It was proposed to encompass information related to molecular shape. It has, in effect, been found that in an isomeric series of hydrocarbons, the H_{GM} index is sensitive to the molecular shape increasing from linear to more branched molecules; it is also inversely related to molecular size, decreasing as the number of atoms in the molecule increases.

The **total information content on the leverage equality** (I_{TH}) and **standardized information content on the leverage equality** (I_{SH}) are defined as

$$I_{TH} = A_0 \cdot \log_2 A_0 - \sum_{g=1}^{G} N_g \cdot \log_2 N_g \qquad I_{SH} = \frac{I_{TH}}{A_0 \cdot \log_2 A_0} = 1 - \frac{\sum_{g=1}^{G} N_g \cdot \log_2 N_g}{A_0 \cdot \log_2 A_0}$$

where A_0 is the number of nonhydrogen atoms, N_g the number of atoms with equal leverage value (within a certain tolerance), and G is the number of equivalence classes.

These descriptors mainly encode information on molecular symmetry; if all the atoms have different leverage values, that is, the molecule does not show any element of symmetry, $I_{TH} = A_0 \log A_0$ and $I_{SH} = 1$; otherwise, if all the atoms have equal leverage values (a perfectly symmetric theoretical case), $I_{TH} = 0$ and $I_{SH} = 0$. The total information content on the leverage equality I_{TH} is more discriminating than I_{SH} because of its dependence on molecular size, and thus it could be thought of as a measure of → *molecular complexity*. These indices were demonstrated to be useful in modeling physico-chemical properties related to entropy and symmetry [Consonni, Todeschini *et al.*, 2002b].

The **mean information content on the leverage magnitude** (*HIC*) is defined as

$$HIC \equiv \bar{I}_H = -\sum_{i=1}^{A} \frac{h_i}{M} \cdot \log_2 \frac{h_i}{M}$$

where M is a constant equal to 1 for linear, 2 for planar, and 3 for nonplanar molecules. This descriptor seems to encompass more information related to molecular complexity than the total and standardized information content on the leverage equality. Unlike I_{TH} and I_{SH}, *HIC* can, for example, recognize the different substituents in a series of monosubstituted benzenes. It is also sensitive to the presence of multiple bonds.

The **average row sum of the influence/distance matrix** (*RARS*) and ***R*-connectivity index** (*RCON*) are defined as

$$RARS = \frac{1}{A} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} = \frac{1}{A} \cdot \sum_{i=1}^{A} VS_i(\mathbf{R})$$

$$RCON = \sum_{i=1}^{A}\sum_{j=1}^{A} a_{ij} \cdot (VS_i(\mathbf{R}) \cdot VS_j(\mathbf{R}))^{1/2}$$

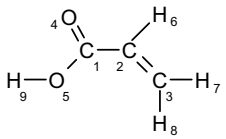
where VS_i is the ith row sum of the influence/distance matrix **R**, A the number of atoms, and elements a_{ij} in *RCON* are equal to 1 for pairs of bonded atoms and zero otherwise. The row sums *VS* encode useful information that may be related to the presence of significant substituents or fragments in the molecule. It was, in effect, observed that larger row sums correspond to terminal atoms that are located very next to other terminal atoms such as those in substituents on a parent structure. Moreover, the *RCON* index is very sensitive to the molecular size as well as to conformational changes and cyclicity.

The ***R*-matrix leading eigenvalue** (*REIG*), in analogy with the → *Lovasz–Pelikan index*, that is, an index of molecular branching calculated as the first eigenvalue of the → *adjacency matrix*, is the largest eigenvalue of the influence/distance matrix **R**.

RARS and *REIG* indices are closely related; their values decrease as the molecular size increases and seem to be a little more sensitive to molecular branching than to cyclicity and conformational changes.

Example G2

Hydrogen-filled molecular graph, molecular influence matrix **H**, and influence/distance matrix **R** for acrylic acid. The matrices were calculated from the *x*, *y*, *z* coordinates of the atoms in the minimum energy conformation optimized by AM1 semiempirical method. Calculation of H_{GM}, I_{TH}, I_{SH}, *HIC*, *RARS*, *RCON*, and *REIG* indices for acrylic acid is here exemplified. VS_i indicates the matrix row sums.



Molecular influence matrix **H**

	C1	C2	C3	O4	O5	H6	H7	H8	H9
C1	0.056	0.004	−0.076	0.130	0.017	0.037	−0.114	−0.110	0.056
C2	0.004	0.054	0.009	0.040	−0.096	0.134	0.049	−0.071	−0.122
C3	−0.076	0.009	0.109	−0.171	−0.048	−0.018	0.170	0.135	−0.109
O4	0.130	0.040	−0.171	0.321	−0.017	0.163	−0.233	−0.293	0.059
O5	0.017	−0.096	−0.048	−0.017	0.179	−0.225	−0.136	0.082	0.243
H6	0.037	0.134	−0.018	0.163	−0.225	0.347	0.061	−0.230	−0.270
H7	−0.114	0.049	0.170	−0.233	−0.136	0.061	0.291	0.157	−0.247
H8	−0.110	−0.071	0.135	−0.293	0.082	−0.230	0.157	0.292	0.038
H9	0.056	−0.122	−0.109	0.059	0.243	−0.270	−0.247	0.038	0.351

Influence/distance matrix **R**

	C1	C2	C3	O4	O5	H6	H7	H8	H9	VS_i
C1	0	0.037	0.031	0.108	0.073	0.064	0.037	0.046	0.073	0.469
C2	0.037	0	0.058	0.054	0.041	0.124	0.059	0.059	0.043	0.475
C3	0.031	0.058	0	0.052	0.050	0.091	0.162	0.162	0.052	0.658
O4	0.108	0.054	0.052	0	0.109	0.125	0.067	0.077	0.150	0.742
O5	0.073	0.041	0.050	0.109	0	0.074	0.059	0.091	0.258	0.755
H6	0.064	0.124	0.091	0.125	0.074	0	0.126	0.102	0.086	0.792
H7	0.037	0.059	0.162	0.067	0.059	0.126	0	0.157	0.066	0.733
H8	0.046	0.059	0.162	0.077	0.091	0.102	0.157	0	0.092	0.786
H9	0.073	0.043	0.052	0.150	0.258	0.086	0.066	0.092	0	0.820

$$H_{GM} = 100 \times \left(\prod_{i=1}^{9} h_i\right)^{1/9} = 100 \times (0.059 \times 0.054 \times 0.109 \times 0.321 \times 0.179 \times 0.347 \times 0.291 \times 0.292 \times 0.351)^{1/9} = 179.8$$

$$I_{TH} = 5 \times \log_2 5 - \sum_{g=1}^{5} N_g \times \log_2 N_g = 11.61 - 5 \times (1 \times \log_2 1) = 11.61$$

$$I_{SH} = \frac{I_{TH}}{5 \times \log_2 5} = \frac{11.61}{11.61} = 1$$

$$HIC = \bar{I}_H = -\sum_{i=1}^{9} \frac{h_i}{2} \times \log_2 \frac{h_i}{2} = -\frac{0.056}{2} \times \log_2 \frac{0.056}{2} - \frac{0.054}{2} \times \log_2 \frac{0.054}{2} - \frac{0.109}{2} \times \log_2 \frac{0.109}{2} - \frac{0.321}{2} \times \log_2 \frac{0.321}{2} - \frac{0.179}{2} \times \log_2 \frac{0.179}{2} - \frac{0.347}{2} \times \log_2 \frac{0.347}{2} - \frac{0.291}{2} \times \log_2 \frac{0.292}{2} - \frac{0.351}{2} \times \log_2 \frac{0.351}{2} = 2.938$$

$$RARS = \frac{1}{9} \times (0.469 + 0.475 + 0.658 + 0.742 + 0.755 + 0.792 + 0.733 + 0.786 + 0.820) = 0.692$$

$$RCON = (0.469 \times 0.475 + 0.469 \times 0.742 + 0.469 \times 0.755 + 0.475 \times 0.658 + 0.475 \times 0.792 + 0.658 \times 0.733 + 0.658 \times 0.786 + 0.755 \times 0.820)^{1/2} = 5.028$$

The set of the eigenvalues of the influence/distance matrix **R** is 0.713, 0.159, 0.022, −0.037, −0.103, −0.149, −0.166, −0.177, −0.263. Therefore, $REIG = 0.713$.

The other set of GETAWAY descriptors consists of autocorrelation vectors obtained by double-weighting the molecule atoms in such way as to account for atomic mass, polarizability, van der Waals volume, and electronegativity together with 3D information encoded by the elements of the molecular influence matrix **H** and influence/distance matrix **R**.

***HATS* indices** are defined by analogy with the → *Moreau–Broto autocorrelation* descriptors *ATS*, weighting each atom of the molecule by its physico-chemical properties combined with the

diagonal elements of the molecular influence matrix **H**, thus also accounting for the 3D features of the molecules:

$$HATS_k(w) = \sum_{i=1}^{A} \sum_{j \geq i}^{A} (w_i \cdot h_i) \cdot (w_j \cdot h_j) \cdot \delta(d_{ij}; k) \quad \text{for} \quad k = 0, 1, 2, \ldots, D$$

where w is an atomic weighting scheme and $\delta(d_{ij}; k)$ a Dirac delta function equal to 1 when the topological distance d_{ij} between atoms i and j is equal to k and zero otherwise. D is the molecule → *topological diameter*, that is, the maximum topological distance in the molecule.

The ***HATS* total index** (*HATS*) is defined as the sum of all the *HATS* indices as

$$HATS(w) = HATS_0(w) + 2 \cdot \sum_{k=1}^{D} HATS_k(w)$$

Example G3

Calculation of *HATS*(m) indices for acrylic acid. This is based on the atomic mass weighting scheme scaled on the Carbon atom: m(C) = 1, m(H) = 0.084, m(O) = 1.332. The molecular influence matrix **H** is in Example G2. Because the topological diameter D is equal to 5, six *HATS* indices ($k = 0, 5$) can be derived. Examples of calculation for $k = 0$ and $k = 3$ are reported. For $k = 0$, the summation goes over the single atoms, then:

$$HATS_0(\mathrm{m}) = \sum_{i=1}^{9} (\mathrm{m}_i \cdot h_i)^2 = 0.003 + 0.003 + 0.012 + 0.183 + 0.057 + 0.001 + 0.001 + 0.001 + 0.001 = 0.262$$

For $k = 3$, the summation goes over all of the atom pairs at topological distance 3:

$$\begin{aligned} HATS_3(\mathrm{m}) = {} & (\mathrm{m}_1 \cdot h_1) \cdot (\mathrm{m}_7 \cdot h_7) + (\mathrm{m}_1 \cdot h_1) \cdot (\mathrm{m}_8 \cdot h_8) + (\mathrm{m}_2 \cdot h_2) \cdot (\mathrm{m}_9 \cdot h_9) + (\mathrm{m}_3 \cdot h_3) \cdot (\mathrm{m}_4 \cdot h_4) \\ & + (\mathrm{m}_3 \cdot h_3) \cdot (\mathrm{m}_5 \cdot h_5) + (\mathrm{m}_4 \cdot h_4) \cdot (\mathrm{m}_9 \cdot h_9) + (\mathrm{m}_4 \cdot h_4) \cdot (\mathrm{m}_6 \cdot h_6) \\ & + (\mathrm{m}_5 \cdot h_5) \cdot (\mathrm{m}_6 \cdot h_6) + (\mathrm{m}_6 \cdot h_6) \cdot (\mathrm{m}_7 \cdot h_7) + (\mathrm{m}_6 \cdot h_6) \cdot (\mathrm{m}_8 \cdot h_8) \\ = {} & 0.001 + 0.001 + 0.002 + 0.047 \\ & + 0.026 + 0.013 + 0.012 + 0.007 + 0.001 + 0.001 = 0.110 \end{aligned}$$

***H* indices** are filtered autocorrelation descriptors defined as

$$H_k(w) = \sum_{i=1}^{A} \sum_{j \geq i}^{A} h_{ij} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; h_{ij}; k) \quad \text{for } k = 0, 1, 2, \ldots, D$$

where h_{ij} are the off-diagonal elements of the molecular influence matrix **H** and the Dirac delta function $\delta(d_{ij}; h_{ij}; k)$ here is defined as

$$\delta(d_{ij}; h_{ij}; k) = \begin{cases} 1 & \text{if } d_{ij} = k \text{ and } h_{ij} > 0 \\ 0 & \text{if } d_{ij} \neq k \text{ or } h_{ij} \leq 0 \end{cases}$$

While the *HATS* indices make use of the diagonal elements of the matrix **H**, the *H* indices exploit the off-diagonal elements, which can be either positive or negative. To emphasize interactions between spatially near atoms, only off-diagonal positive h values are used. In effect,

for a given lag (i.e., topological distance), the product of the atom properties is multiplied by the corresponding h_{ij} value and only those contributions with a positive h_{ij} value are considered. This means that, for a given atom i, only those atoms j at topological distance d_{ij} with a positive h_{ij} value are considered because they may have the chance to interact with the ith atom.

The ***H* total index** (HT) is defined as the sum of all the H indices:

$$HT(w) = H_0(w) + 2 \cdot \sum_{k=1}^{D} H_k(w)$$

Example G4

Calculation of $H(m)$ indices for acrylic acid. Calculation is based on the atomic mass weighting scheme scaled on the Carbon atom: $m(C) = 1$, $m(H) = 0.084$, $m(O) = 1.332$. The molecular influence matrix **H** is in Example G2. Because the topological diameter D is equal to 5, six H indices ($k = 0, 5$) can be derived. Examples of calculations for $k = 0$ and $k = 3$ are reported.

For $k = 0$, the summation goes over the single atoms, then:

$$H_0(m) = \sum_{i=1}^{9} h_i m_i^2 = 0.056 + 0.054 + 0.109 + 0.570 + 0.318 + 0.002 + 0.002 + 0.002 + 0.002 = 1.115$$

For $k = 3$, the summation goes over the atom pairs at topological distance 3, which have a positive h_{ij} value:

$$H_3(m) = (m_4 \times h_4)(m_9 \times h_9) + (m_4 \times h_4) \times (m_6 \times h_6) + (m_6 \times h_6) \times (m_7 \times h_7) = 0.0182 + 0.0066 + 0.0004 = 0.025$$

***R* indices** are defined in the same way as the H indices, by using the off-diagonal elements of the influence/distance matrix **R** instead of the elements of the matrix **H**:

$$R_k(w) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; k) \quad \text{for } k = 1, 2, \ldots, D$$

In this case, no filtering is applied, because geometrical distances r_{ij} act as a smoothing function.

The ***R* total index** (RT) is defined as twice the sum of the R indices:

$$RT(w) = 2 \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j = 2 \cdot \sum_{k=1}^{D} R_k(w)$$

where w is the atomic property and D is the topological diameter. In the case of unitary weights, that is, $w = 1$, the R total index is twice the → *Wiener-type index* derived from the influence–distance matrix as the half sum of all the matrix elements. Moreover, it is strictly related to the → *gravitational index* G1.

To take into account local aspects of the molecule and allow → *reversible decoding*, the **maximal *R* indices** (R^+) were also proposed as

$$R_k^+(w) = \max_{ij}\left(\frac{\sqrt{h_i \cdot h_j}}{r_{ij}} \cdot w_i \cdot w_j \cdot \delta(d_{ij}; k)\right) \quad i \neq j;\ k = 1, 2, \ldots, D$$

where only the maximum property product between atom pairs at a given topological distance (lag) is retained.

The maximum value among the kth order maximal indices $R_k^+(w)$ is called **maximal *R* total index** (RT^+) and defined as

$$RT^+(w) = \max_k(R_k^+(w))$$

Example G5

Calculation of $R(m)$ and $R^+(m)$ indices for acrylic acid. Calculation is based on the atomic mass weighting scheme scaled on the Carbon atom: $m(C) = 1$, $m(H) = 0.084$, $m(O) = 1.332$. The influence/distance matrix **R** is given in Example G2. Because the topological diameter D is equal to 5, five R indices ($k = 1, 5$) can be derived. Example of calculations for $k = 3$ is reported. In this case, the summation goes over all the atom pairs at topological distance 3:

$$\begin{aligned} R_3(m) &= [\mathbf{R}]_{1,7} \cdot m_1 \cdot m_7 + [\mathbf{R}]_{1,8} \cdot m_1 \cdot m_8 + [\mathbf{R}]_{2,9} \cdot m_2 \cdot m_9 + [\mathbf{R}]_{3,4} \cdot m_3 \cdot m_4 + [\mathbf{R}]_{3,5} \cdot m_3 \cdot m_5 \\ &+ [\mathbf{R}]_{4,9} \cdot m_4 \cdot m_9 + [\mathbf{R}]_{4,6} \cdot m_4 \cdot m_6 + [\mathbf{R}]_{5,6} \cdot m_5 \cdot m_6 + [\mathbf{R}]_{6,7} \cdot m_6 \cdot m_7 + [\mathbf{R}]_{6,8} \cdot m_6 \cdot m_8 \\ &= 0.003 + 0.004 + 0.004 + 0.069 + 0.067 \\ &+ 0.017 + 0.014 + 0.008 + 0.001 + 0.001 = 0.188 \end{aligned}$$

$$R_3^+(m) = \max(0.003;0.004;0.004;0.069;0.067;0.017;0.014;0.008;0.001;0.001) = 0.069$$

Note that the $R_3^+(m)$ index identifies the structural fragment $C_3 = C_2 - C_1 = O_4$.

The atomic weighting schemes applied for GETAWAY descriptor calculation are those proposed for the → *WHIM descriptors*, that is, atomic mass (m), → *atomic polarizability* (p), Sanderson → *atomic electronegativity* (e), atomic → *van der Waals volume* (v), and the unit weighting scheme (u).

HATS, *H*, *R*, and maximal *R* indices are → *vectorial descriptors* for structure–property correlations, but they can also be used as molecular profiles suitable for → *similarity/diversity analysis* studies. These descriptors, based on spatial autocorrelation, encode information on structural fragments and therefore seem to be particularly suitable for describing differences in congeneric series of molecules. Unlike the Moreau–Broto autocorrelations, GETAWAYs are geometrical descriptors encoding information on the effective position of substituents and fragments in the molecular space. Moreover, they are independent of molecule alignment and, to some extent, account also for information on molecular size and shape as well as for specific atomic properties.

A joint use of GETAWAY and WHIM descriptors is advised, exploiting both local information of the former and holistic information of the latter set of descriptors. The GETAWAY descriptors have been used for modeling several data sets of pharmacological and environmental interest [Consonni, Todeschini *et al.*, 2002b; Fedorowicz, Singh *et al.*, 2005; Pérez González, Terán *et al.*, 2005b; Saiz-Urra, Pérez González *et al.*, 2007].

Table G1 Some GETAWAY descriptors for the data set comprised of 22 *N*,*N*-dimethyl-α-bromo-phenetylamines (Appendix C – Set 2).

Mol.	*X*	*Y*	I_{TH}	I_{SH}	*HIC*	H_{GM}	*RCON*	*RARS*	*REIG*	*HT*(u)	*HT*(m)	*RT*(u)	*RT*(m)
1	H	H	43.020	1.000	4.428	8.933	17.445	0.744	0.796	17.233	18.411	19.341	8.499
2	H	F	44.106	0.917	4.387	8.935	17.180	0.733	0.780	17.237	27.588	19.071	11.590
3	H	Cl	46.106	0.958	4.384	8.915	16.997	0.726	0.773	17.220	29.869	18.875	13.031
4	H	Br	46.106	0.958	4.381	8.897	16.903	0.722	0.769	17.192	40.542	18.781	17.296
5	H	I	46.106	0.958	4.389	8.936	16.945	0.722	0.768	17.269	58.506	18.783	21.530
6	H	Me	44.106	0.917	4.562	8.199	18.570	0.708	0.752	19.608	24.297	20.542	9.316
7	F	H	46.106	0.958	4.380	8.905	17.123	0.731	0.778	17.179	27.692	19.009	11.965
8	Cl	H	46.106	0.958	4.371	8.846	16.880	0.721	0.768	17.141	30.052	18.739	13.709
9	Br	H	46.106	0.958	4.367	8.825	16.793	0.717	0.763	17.126	42.815	18.636	18.862
10	I	H	46.106	0.958	4.361	8.801	16.701	0.712	0.759	17.095	66.534	18.525	24.200
11	Me	H	48.106	1.000	4.546	8.081	18.754	0.707	0.755	19.323	23.817	20.502	9.407
12	Cl	F	51.303	0.962	4.372	8.836	16.788	0.717	0.764	17.150	33.358	18.648	16.166
13	Br	F	51.303	0.962	4.364	8.797	16.670	0.712	0.760	17.094	47.747	18.523	22.247
14	Me	F	47.303	0.887	4.539	7.975	18.541	0.701	0.751	19.305	24.883	20.331	11.064
15	Cl	Cl	53.303	1.000	4.371	8.822	16.619	0.710	0.757	17.160	36.601	18.468	18.041
16	Br	Cl	53.303	1.000	4.359	8.768	16.462	0.704	0.752	17.040	52.386	18.310	24.842
17	Me	Cl	47.303	0.887	4.533	7.953	18.326	0.694	0.743	19.219	27.406	20.126	12.496
18	Cl	Br	51.303	0.962	4.364	8.792	16.501	0.706	0.753	17.083	49.660	18.354	23.683
19	Br	Br	53.303	1.000	4.355	8.744	16.365	0.701	0.748	17.008	68.303	18.215	32.193
20	Me	Br	49.303	0.925	4.544	7.979	18.348	0.694	0.743	19.321	35.356	20.133	16.136
21	Me	Me	53.303	1.000	4.711	7.439	20.224	0.691	0.736	21.603	22.027	22.096	9.134
22	Br	Me	51.303	0.962	4.555	8.155	18.266	0.696	0.739	19.438	42.620	20.173	17.704

X and *Y* refer to molecule substituents.

🕮 [Consonni and Todeschini, 2001; Grodnitzky and Coats, 2002; Gramatica, Pilutti *et al.*, 2003a, 2004b; Kiralj, Takahata *et al.*, 2003; Pérez González and Helguera, 2003; Farkas, Héberger *et al.*, 2004; Fedorowicz, Zheng *et al.*, 2004; Garkani-Nejad, Karlovits *et al.*, 2004; Gramatica, Battaini *et al.*, 2004; Guha, Serra *et al.*, 2004; Jelcic, 2004; Marrero-Ponce, 2004a; Pérez González, Helguera Morales *et al.*, 2004; Pérez González, Helguera *et al.*, 2004; Pérez González and Moldes Teran, 2004; Schefzik, Kibbey *et al.*, 2004; Kabankin and Gabrielyan, 2005; Kovatcheva, Golbraikh *et al.*, 2004; Deconinck, Hancock *et al.*, 2005; Fedorowicz *et al.*, 2005; Panek, Jezierska *et al.*, 2005; Papa, Battaini *et al.*, 2005; Papa, Villa *et al.*, 2005; Pérez González, Terán *et al.*, 2005a; 2006; Caballero and Fernández, 2006; Li *et al.*, 2006; Li, Maldonado, Doucet *et al.*, 2006; Pis Diez, Duchowicz *et al.*, 2006; Yap, Li *et al.*, 2006; Carlucci, D'Archivio *et al.*, 2007; Cruz-Monteagudo, Borges *et al.*, 2007; Deconinck, Ates *et al.*, 2007; Duchowicz, Pérez González *et al.*, 2007; Zheng, Zheng *et al.*, 2007]

- ➢ **Ghose–Crippen descriptors** → lipophilicity descriptors (⊙ Ghose–Crippen hydrophobic atomic constants)
- ➢ **Ghose–Crippen hydrophobic atomic constants** → lipophilicity descriptors
- ➢ **Gini concentration index** → statistical indices (⊙ concentration indices)
- ➢ **Gini index** → information content

■ **GIPF approach** (≡ *General Interaction Properties Function approach*)
This is a general method, proposed by Politzer and coworkers, to estimate physico-chemical properties depending on noncovalent interactions [Brinck, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1993; Politzer, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1994]. This approach is based on molecular surface area in conjunction with some statistically based quantities related to the → *molecular electrostatic potential* (MEP) at the → *molecular surface*. The → *electron isodensity contour surface* (0.001 a.u. contour of $\rho(\mathbf{r})$) is taken as the molecular surface model.

The general GIPF model for a physico-chemical property Φ is

$$\Phi = f(SA, \Pi, \sigma^2_{tot}, \nu)$$

where SA is the surface area and Π is the → *local polarity index*. The other two molecular surface indices are defined as the following:

$$\sigma^2_{tot} = \sigma^2_+ + \sigma^2_- = \frac{1}{n^+} \cdot \sum_{i=1}^{n^+} [V^+(\mathbf{r}_i) - \bar{V}^+]^2 + \frac{1}{n^-} \cdot \sum_{i=1}^{n^-} [V^-(\mathbf{r}_i) - \bar{V}^-]^2 \quad \text{and}$$

$$\nu = \frac{\sigma^2_+ \cdot \sigma^2_-}{(\sigma^2_{tot})^2}$$

where σ^2_{tot} is the **surface electrostatic potential variance**, which measures the electrostatic interaction tendency of the molecule, σ^2_+ and σ^2_- are the variances of positive and negative regions of the molecular surface potential, V^+ and V^- are the positive and negative values of the MEP at a grid point $\mathbf{r}$ on the molecular surface, $\bar{V}^+$ and $\bar{V}^-$ are their average values, and n^+ and n^- are the numbers of grid points with positive and negative values. ν is the **electrostatic balance term** that reaches a maximum value of 0.25 when σ^2_+ and σ^2_- are equal.

Site-specific molecular quantities can be added to the global molecular descriptors in the GIPF model depending on the physico-chemical property to be estimated. Some of these site-specific descriptors are defined below:

$\bar{I}_{S,\min}$ is the lowest value of the → *average local ionization energy* found on the molecular surface; this reflects the tendency for charge transfer and polarization at any particular molecular site [Haeberlein and Brinck, 1997].

$V_{S,\min}$ and $V_{S,\max}$ are the most negative and positive values of the molecular electrostatic potential on the molecular surface; the maximum reflects the tendency for long-range attraction of nucleophiles at a specific site, whereas the minimum reflects the tendency for long-range attraction of electrophiles at a specific site. $V_{S,\min}$ and $V_{S,\max}$ for a large variety of molecules correlate with hydrogen bond basicity and acidity, respectively [Murray and Politzer, 1998].

SA^+ and SA^- are the portions of the surface area over which $V(\mathbf{r})$ is positive and negative, respectively.

Several properties have been estimated by the GIPF approach such as heat of vaporization, sublimation [Politzer, Murray *et al.*, 1997] and fusion [Murray, Brinck *et al.*, 1996], boiling point and critical constants [Murray, Lane *et al.*, 1993a], surface tension, liquid and solid density [Murray, Brinck *et al.*, 1996], crystal lattice energy [Politzer and Murray, 1998], impact sensitivity [Murray, Lane *et al.*, 1998], diffusion coefficient [Politzer, Murray *et al.*, 1996],

solubility [Politzer, Lane *et al.*, 1992; Murray, Gagarin *et al.*, 1995], aqueous solvation free energy [Murray, Abu-Awwad *et al.*, 1999; Politzer, Murray *et al.*, 2000], → *hydrogen-bonding parameters* [Lowrey, Cramer *et al.*, 1995], and → *lipophilicity*.

🕮 [Murray, Ranganathan *et al.*, 1991; Brinck, Murray *et al.*, 1993; Murray, Lane *et al.*, 1993b; Politzer and Murray, 1994; Beck, Horn *et al.*, 1998]

- **girth** → graph
- **glass transition temperature** → technological properties
- **GLI index** ≡ *Global Leachability Index* → environmental descriptors (⊙ leaching indices)
- **global cyclicity indices** → resistance matrix
- **global flexibility index** → flexibility indices
- **Global Leachability Index** → environmental indices (⊙ leaching indices)
- **global site-property analysis** → Hansch analysis
- **global synthetic invariant** → iterated line graph sequence
- **global topological charge index** → topological charge indices
- **Global Warming Potential** → environmental indices
- **global weighted walk numbers** → walk matrices
- **global WHIM descriptors** → WHIM descriptors
- **globularity** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **globularity factor** → shape descriptors (⊙ ovality index)
- **Gombar hydrophobic model** ≡ *VLOGP* → lipophilicity descriptors
- **GMPEI** ≡ *Geometric Molecular Polarizability Effect Index* → electric polarization descriptors (⊙ polarizability effect index)
- **Golbraikh–Tropsha statistics** → regression parameters
- **GOLPE** → variable selection
- **goodness of fit** → regression parameters
- **goodness of prediction** → regression parameters
- **Gordon–Scantlebury index** ≡ *connection number* → edge adjacency matrix
- **Gordy's bond order** → bond order indices (⊙ bond order–bond length relationships)

graph

A graph is a mathematical object defined within the *graph theory* [Harary, 1964, 1969a, 1969b; Rouvray, 1971, 1990a; Wilson, 1972; Rouvray and Balaban, 1979; Balaban and Harary, 1976; Bonchev and Rouvray, 1991, 1998; Trinajstić, 1992; Ivanciuc and Balaban, 1999c; Marks *et al.*, 2002; Ivanciuc, 2003c; Kruja, Randić, 2003b; Gutman, 2006].

Note. Graph theory is a branch of mathematics that studies the structure of graphs and networks. Graph theory started in 1736 when Euler solved the problem known as the Königsberg bridges problem [Euler, 1741], which was reduced by him to a graph-theoretical problem.

Although the term "graph" was first introduced into literature by mathematician Sylvester [Sylvester, 1877, 1878], who derived it from the contemporary chemical term "graphical notation," used to denote the chemical structure of a molecule, the research field that is nowadays called *chemical graph theory* started some years before when the British mathematician Arthur Cayley published his works about *trees* [Cayley, 1857, 1859] and then the paper "*On the mathematical theory of isomers*" [Cayley, 1874].

The first chemical application of graph theory dates back to 1875 when William Clifford proposed the solution for the counting of alkane isomers. The modern chemical graph theory started with the works of Henze and Blair in 1931 [Henze and Blair, 1931a, 1931b, 1933, 1934] and Pólya in 1936 [Pólya, 1936, 1937a, 1937b; Pólya and Read, 1987].

A graph is usually denoted as $G = (V, E)$, where V is a set of **vertices** and E is a set of elements representing the binary relationship between pairs of vertices; unordered vertex pairs are called **edges**, ordered vertex pairs are called **arcs**, and elements of E that relate a vertex with itself are called **loops**. A graph is described by either **adjacencies**, which refer to adjacent vertex–vertex or edge–edge pairs or **incidences** that refer to adjacent vertex–edge pairs or, more generally, to pairs of mathematical objects of two different kinds.

An edge is a **cut edge** if its removal produces a disconnected graph; it cannot be a part of a cycle; similarly, a vertex is a **cut vertex** if its removal produces a disconnected graph. Of course, each vertex incident to a cut edge is a cut vertex, but a cut vertex can also be a part of a cycle.

If two vertices occur as an unordered pair more than once, they define a **multiple edge**; if two vertices occur as an ordered pair more than once, they define a **multiple arc**. Two edges in a graph G are said to be **independent edges** if they have no common vertex. A collection of k mutually (i.e., pairwise) independent edges in a graph $G (k \geq 2)$ is called a ***k*-matching** of G.

G and G' are called **isomorphic graphs** if a bijective mapping of the vertex and the edge sets exists, that is,

$$V(G) \leftrightarrow V(G') \quad \text{and} \quad E(G) \leftrightarrow E(G')$$

or, in other words, if there exists a one-to-one correspondence between the vertices and the edges, such that adjacency is preserved. A **graph automorphism** is an isomorphic mapping of a graph G onto itself, that is, it is a bijective mapping of the vertex and edge sets onto themselves, which preserves the number of links joining any two vertices:

$$V(G) \leftrightarrow V(G) \quad \text{and} \quad E(G) \leftrightarrow E(G)$$

The set of all automorphisms of a graph forms an **automorphism group**. The occurrence of automorphism depends on the symmetry of the graph; in particular, it depends on the presence of equivalent vertices, which can be mapped automorphically onto each other, that is, they can interchange preserving the adjacency of the graph. The cardinality of the automorphism group of a graph is called **symmetry number** and is considered among the → *symmetry descriptors*.

Topologically equivalent vertices constitute disjoint subsets of vertices called **orbits**.

A graph G' is a **subgraph** of the graph G if the following relationships hold:

$$V(G') \subseteq V(G) \quad \text{and} \quad E(G') \subseteq E(G)$$

Graph components are connected subgraphs or vertices that are not connected to each other.

The → *vertex degree* is the number of edges incident to a given vertex. If two vertices are connected by an arc, two degrees are assigned to each vertex; the **indegree** counts the arcs ending on the vertex, the **outdegree** counts the arcs starting from the vertex. **Terminal vertices** are the vertices of a graph with degree equal to 1; **terminal edges** are the edges incident to terminal vertices. **Central vertices** and **central edges** are the vertices (edges) belonging to the → *graph center*. All vertices with vertex degree equal to zero are called **isolated vertices**. *Branching* of a graph is a fuzzy concept that can be based on the presence in the graph of vertices with degrees equal to 3 or higher. It plays a basic role in assessing the → *molecular complexity*.

A **walk** (or **random walk**) in $\mathcal{G}$ is a sequence of vertices $w = (v_1, v_2, \ldots, v_k)$ such that $(v_i, v_{i+1}) \in \mathcal{E}$ for each $i = 1, k-1$, that is, a sequence of pairwise adjacent edges leading from vertex v_1 to vertex v_k; any vertex or edge can be traversed several times. The **walk length** is the number of edges traversed by the walk.

A **path** (or **self-avoiding walk**) is a walk without any repeated vertices. The **path length** is the number of edges associated with the path. The smallest path between two vertices considered is called **geodesic** and its length corresponds to the → *topological distance*; **elongation** is the longest path between two vertices considered and its length corresponds to the → *detour distance* between the vertices.

A walk closed in itself is called **self-returning walk**, that is, a walk starting and ending on the same vertex.

A self-returning path is called **cyclic path** (or **cycle** or **circuit**), that is, a cycle is a walk with no repeated vertices (i.e., a path) other than its first and last ones ($v_1 = v_k$). The number of independent cycles (or rings) in a graph is the → *cyclomatic number*. **Cyclicity** C^+ is the number of all possible cycles in a graph. A **girth** is the length of the shortest cycle (if any) in a graph $\mathcal{G}$. Acyclic graphs are considered to have infinite girth.

A **trail** is a walk in which vertices can be revisited but edges can be traversed only once; an **Eulerian walk** is a trail in which all vertices of the graph must be encountered.

A **Hamiltonian path** is a path in which all vertices of the graph must be visited once and the beginning and the end are different. A **Hamiltonian circuit** is a path in which all vertices of the graph must be visited once, starting and ending on the same vertex.

A **dissection of a graph** $\mathcal{G}$ is a collection of subgraphs obtained by erasing one vertex at a time from a graph $\mathcal{G}$ and all its so-obtained subgraphs $\mathcal{G}^*$ generated from $\mathcal{G}$, which neither represent isolated vertices nor isolated bonds [Randić, Guo *et al.*, 2000]. The dissection of a graph is determined by a stepwise procedure in which one removes one vertex at a time from the graph considered and continues to do so on all subgraphs that are neither isolated vertices nor isolated bonds. The total number a of subgraphs, which are isolated vertices, and the total number b of subgraphs, which are isolated edges, obtained by the whole graph dissection, can be considered two simple topological descriptors.

A list of graphs of practical interest follows.

- **simple graph** (≡ *normal graph, schlicht graph*)
Graph having no arcs, no multiple edges or loops.

- **planar graph**
Graph that can be drawn so that no edge-crossing appears.

- **cyclic graph**
Graph containing at least one cycle. Each cycle is usually denoted as C_m ($m \geq 3$), where m is the number of vertices in the cycle.

- **digraph** (≡ *directed graph*)
Graph in which all vertex pairs are arcs. If any vertex pair is associated with only one arc, the graph is called **oriented graph**.

- **multigraph** (≡ *multiple graph*)
Graph having no arcs or loops, but including multiple edges between at least a pair of vertices.

- **general graph** (≡ *nonsimple graph*)
Graph containing multiple edges and loops.

- **pseudograph** (≡ *loop-multigraph*)
Graph having no arcs or multiple edges but containing loops.

- **connected graph**
Graph in which for each pair of vertices $\{i,j\} \in V(G)$ at least one path exists. Otherwise G is called **disconnected graph**. The simplest disconnected graph is a graph with an isolated vertex and the vertices not joined by a path belong to different components of the graph. The number of components of a graph is denoted $k(G)$.

- **regular graph**
If all vertices in a graph have the same degree, then the graph is called regular graph, otherwise **irregular graph**.

- **tree** (≡ *acyclic graph*)
Connected graph without cycles, usually denoted as T_A, where A is the number of vertices in the graph. The number of edges B and the number of vertices A are related by the condition $B = A - 1$. A **rooted tree** is a tree having one vertex distinguished from the others; if this vertex is an end point, the graph is called **planted tree**. In chemistry, rooted and planted trees can be used to represent molecular substituents.

- **linear graph** (≡ *path graph*)
A tree without branching; there are exactly two terminal vertices of degree 1 and $A-2$ vertices of degree two.

- **star graph**
A maximally branched tree, that is, a set of vertices joined by a common vertex; there are $A-1$ terminal vertices of degree 1 and one vertex of degree $A-1$. It is usually denoted as S_A, where A is the number of vertices in the graph.

- **complete graph**
Graph in which all vertices and edges are mutually adjacent, that is, all vertices have degree $A-1$. The maximal number of edges in a graph is

$$B = \binom{A}{2} = \frac{A \cdot (A-1)}{2}$$

A complete graph contains the maximal number of cycles and is denoted as K_A, A being the number of vertices.

- **clique**
A maximal complete subgraph in which every vertex is connected to every other vertex and which is not contained in any other larger subgraph with this property.

- **forest**
A set of disjoint trees $\mathcal{F} = \{(\mathcal{V}_1, \mathcal{E}_1), \ldots, (\mathcal{V}_k, \mathcal{E}_k)\}$; a forest does not contain cycles.

- **spanning tree**
A connected acyclic subgraph containing all the vertices of $\mathcal{G}$.

- **minimal spanning tree**
A spanning tree in which the number of edges is minimal.

- **indexed graph**
A graph $\mathcal{G}$ associated to a mapping ϕ such as

$$\phi : \mathcal{V}(\mathcal{G}) \to \{1, 2, \ldots, A\}$$

where $\mathcal{V}(\mathcal{G})$ is the set of A vertices of a graph and the indexing function ϕ assigns an integer number to each vertex of the graph. Univocally defined indexed graphs are obtained by $\rightarrow$ *canonical numbering* of graph vertices.

- **Sachs graph** ($\equiv$ *basic graph, mutation graph, characteristic graph*)
A graph defined as a subgraph of $\mathcal{G}$ whose components are K_2 (complete graphs) or C_m (cycle graphs) or combinations between a K_2 components and b C_m components, under the constraint:

$$a \times 2 + b \times m = A$$

where A is the number of vertices in the Sachs graph.

- **signed graph**
A signed graph is a graph with a sign attached to each edge.

- **weighted graph**
A graph $\mathcal{G}$ in which a weight $w_{ij} \geq 0$ is assigned to each edge $\{i, j\} \in \mathcal{E}(\mathcal{G})$ or a weight w_i is assigned to each vertex $i \in \mathcal{V}(\mathcal{G})$.

- **line graph**
A line graph $L(\mathcal{G})$ is a graph obtained by representing the edges of the graph $\mathcal{G}$ by points and then by joining two such points with a line if the edges they represent are adjacent in the original graph; the following relation holds:

$$|\mathcal{V}(L(\mathcal{G}))| = |\mathcal{E}(\mathcal{G})|$$

Multiple edges are considered as independent vertices in the line graph.

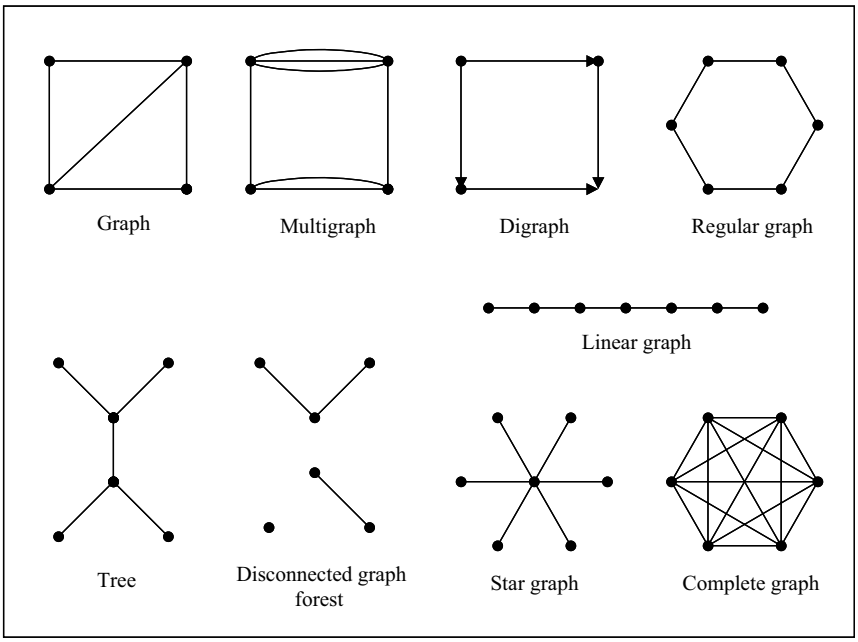


Figure G2 Examples of graphs.

- **chromatic graph**

Graph whose vertices or edges are symbolically differentiated by assigning 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 (→ *chromatic decomposition*).

- **isocodal graphs**

Graphs with identical atomic codes for all the vertices, that is, there exists a one-to-one correspondence between the atomic codes of all vertices. The most known atomic codes are → *walk count atomic code*, → *self-returning walk count atomic code*, and → *atomic path code*.

- **isospectral graphs** (≡ *cospectral graphs*)

Nonisomorphic graphs having the same → *characteristic polynomial*.

- **subspectral graphs**

Graphs whose eigenvalues of the characteristic polynomial are contained in the spectrum of another graph.

- **homeomorphic graphs**

Graphs obtained from the same graph by a sequence of line subdivisions.

- **chemical graph**

A chemically interpreted graph is called chemical graph, that is, graph representing a chemical system such as molecules, reactions, crystals, polymers, and orbitals. The common feature of chemical systems is the presence of sites (atoms, electrons, molecules, molecular fragments, etc.)

and connections (bonds, reaction steps, van der Waals forces, etc.) between them. In the graph representation of a chemical system, sites are replaced by vertices and connections by edges. The most common chemical graphs are molecular graphs and reaction graphs. The former correspond to specific chemical structures, whereas the latter to sets of chemical reactions. A topological representation of a molecule is given by a → *molecular graph*.

🕮 [Sachs, 1964; Graovac, Gutman *et al.*, 1972; Hosoya, 1972a; Balaban, 1976d, 1993e; Read and Corneil, 1977; Quintas and Slater, 1981; von Knop, Müller *et al.*, 1981; King, 1983; Randić, Woodworth *et al.*, 1983; Grossman, 1985; Balaban, Kennedy *et al.*, 1988; Balaban and Tomescu, 1988; Hansen and Jurs, 1988a; Gutman, 1991a; Liu and Klein, 1991; Polansky, 1991; Rouvray, 1991; Rücker and Rücker, 1991a, 1992; Bangov, 1992; Bonchev and Rouvray, 1992; Figueras, 1992; Gautzsch and Zinn, 1992a, 1992b, 1994; Ivanciuc and Balaban, 1992a, 1999c; Müller, Szymanski *et al.*, 1995; Lukovits, 1996a; Lepovic and Gutman, 1998; Balinska, Gargano *et al.*, 2001; Vukicević and Graovac, 2004b; Vukicević and Graovac, 2005]

- **graph automorphism** → graph
- **graph characteristic polynomial** → characteristic polynomial-based descriptors
- **graph center** ≡ center *of a graph*
- **graph coloring** → chromatic decomposition
- **graph distance code** → distance matrix
- **graph distance complexity** → topological information indices
- **graph distance count** → distance matrix
- **graph distance index** → distance matrix
- **graph eigenvalues** → characteristic polynomial-based descriptors
- **graph energy** → spectral indices

■ **graph entropy**

The graph entropy approach is based on the idea to catch the structural → *mean information content* in a graph by means of an information functional f [Dehmer, 2008a, 2008b; Dehmer and Emmert-Streib, 2008].

An **information functional** is a positive and monotonous function that captures structural information in a graph by defining the probability value for each graph vertex. The probability of the vertex v_i is defined as

$$p(v_i) = \frac{f(v_i)}{\sum_{j=1}^{A} f(v_j)}$$

where f represents an arbitrary information functional and A the total number of graph vertices. Because, by definition, it holds the equation:

$$p(v_1) + p(v_2) + p(v_3) + \ldots + p(v_A) = 1$$

the quantities $p(v_i)$ can be interpreted as the vertex probabilities.

Therefore, the graph entropy, denoted as I_f, is the structural mean information content defined as

$$I_f = -\sum_{i=1}^{A} \frac{f(v_i)}{\sum_{j=1}^{A} f(v_j)} \cdot \log_2 \frac{f(v_i)}{\sum_{j=1}^{A} f(v_j)}$$

The information functional is defined as

$$f(v_i) = \alpha^{[c_1 \cdot {}^1f_i + c_2 \cdot {}^2f_i + \cdots + c_D \cdot {}^Df_i]} \quad c_k > 0;\ \alpha > 0$$

where c_k are arbitrary real positive coefficients, D is the → *topological diameter*, and ${}^kf_i$ is the → *vertex distance count*, that is, the number of vertices at a → *topological distance k* from vertex v_i.

- **graphical bond order** → bond order indices
- **graphical bond order descriptors** → bond order indices (⊙ graphical bond order)
- **graphical matrices** → double invariants

■ **graph invariants** (≡ *graph-theoretical invariants*)
These are → *molecular descriptors* derived from a graph representation of the molecule and representing graph-theoretical properties that are preserved by isomorphism, that is, properties with identical values for → *isomorphic graphs*. A graph invariant may be a → *characteristic polynomial*, a sequence of numbers (→ *vectorial descriptors*) or a single numerical index obtained by the application of → *algebraic operators* to → *graph-theoretical matrices* and whose values are independent of vertex numbering or labelling [Kier and Hall, 1976a, 1986; Bonchev and Trinajstić, 1977; Bonchev, Mekenyan *et al.*, 1979; Balaban, Motoc *et al.*, 1983; Rouvray, 1983, 1995, 1989a; Basak, Magnuson *et al.*, 1987; Hansen and Jurs, 1988a; Basak, Niemi *et al.*, 1990c; Trinajstić, 1992; Randić, 1993a, 1998c, 2003b; Balaban, 1997a, 1998; Basak, Grunwald *et al.*, 1997; Diudea and Gutman, 1998; Balaban and Ivanciuc, 1999; Devillers and Balaban, 1999; Ivanciuc and Balaban, 1999c; Bonchev and Rouvray, 2000; Bonchev, 2003a; Ivanciuc, 2003c; Kerber, Laue *et al.*, 2004].

Single indices that are a numerical representation of the molecular structure derived from a → *molecular graph* are called **topological indices** (TIs) or **molecular topological indices** (MTIs). These are numerical quantifiers of molecular topology that are mathematically derived in a direct and unambiguous manner from the structural graph of a molecule, usually a → *H-depleted molecular graph*. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching, and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. In fact, topological indices were proposed to be divided into two categories: *topostructural* and *topochemical indices* [Basak, Gute *et al.*, 1997; Gute, Grunwald *et al.*, 1999]. **Topostructural indices** encode only information about the adjacency and distances between atoms in the molecular structure; **topochemical indices** quantify information about not only topology but also specific chemical properties of atoms such as their chemical identity and hybridization state (Figure G3).

→ *Topological information indices* are graph invariants, based on information theory and calculated as → *information content* of specified equivalence relationships on the molecular graph.

Topological indices are mainly based on distances between atoms calculated by the number of intervening bonds and are thus considered *through-bond* indices; they differ from → *topographic indices* and → *geometrical descriptors* that are, instead, considered *through-space* indices because they are based on interatomic → *geometric distances* [Diudea, Horvath *et al.*, 1995b; Balaban, 1997a].

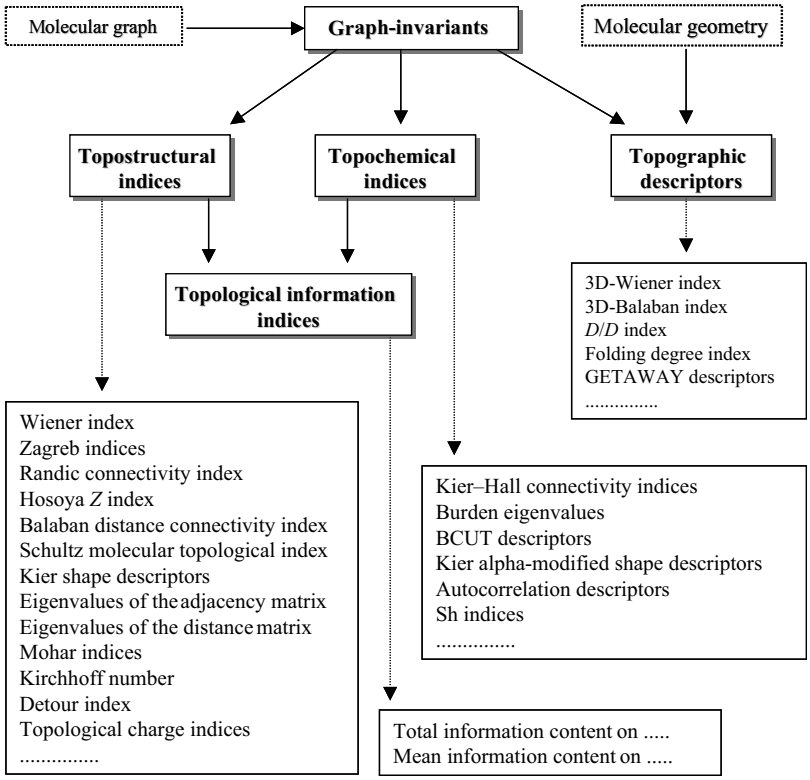


Figure G3 Different classes of graph invariants.

In general, TIs do not uniquely characterize molecular topology, different structures may have some of the same TIs. A consequence of topological indices' nonuniqueness is that they do not, in general, allow reconstructing molecule. Therefore, suitably defined ordered sequences of TIs can be used to characterize molecules with higher discrimination.

There are several ways to obtain topological descriptors. Simple topological indices consist in the counting of some specific graph elements; for examples, the → *Hosoya Z index*, → *path counts, walk counts*, → *self-returning walk counts*, → *Kier shape descriptors*, and → *path/walk shape indices*. However, the most common TIs are derived by applying some algebraic operators (e.g., the → *Wiener operator*) to → *graph-theoretical matrices*, such as → *adjacency matrix* **A**, → *distance matrix* **D**, → *detour matrix* **Δ**, → *Szeged matrices* **SZ**, → *Cluj matrices* **CJ**, → *layer matrices* **LM**, and → *walk matrices* **W**; among them there are the → *Wiener index*, → *Randić connectivity index* and related indices, → *Balaban distance connectivity index*, → *Schultz molecular topological index*, → *hyper-Wiener index*, → *quasi-Wiener index*, → *spectral indices*, → *determinant-based descriptors*, and → *Harary indices*. The most common functions to derive graph invariants from graph-theoretical matrices are listed in Table G2. Note that in functions $\mathcal{D}_1$ and $\mathcal{D}_2$, the most common parameter values are $\alpha = 1/2$ and $\lambda = 1$. Function $\mathcal{D}_3$ is used to generate descriptors derived from the matrix determinant and function $\mathcal{D}_4$ descriptors that

are linear combinations of the coefficients of the characteristic polynomial of a graph-theoretical matrix, such as the → *Hosoya-type indices*. Function $\mathcal{D}_5$ is based on the eigenvalues calculated from graph-theoretical matrices, and the related molecular descriptors are the so-called → *spectral indices*. Function $\mathcal{D}_6$ makes use of the matrix row sums VS_i (→ *row sum operator*) as the → *local vertex invariants* and, then, adds up the contributions from different graph fragments (e. g., edges), each weighted by the product of the local invariants of all the vertices contained in the fragment; → *connectivity-like indices*, such as the → *Randić connectivity index* and → *Balaban-like indices*, are calculated according to this function. Function $\mathcal{D}_7$ for $\alpha = 1/2$ and $\lambda = 2$ generates the → *hyper-Wiener-type indices*.

Table G2 Classical functions to derive molecular descriptors from graph-theoretical matrices.

ID	Function	ID	Function
1.	$\mathcal{D}_1(\mathbf{M};\alpha,\lambda) = \alpha \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{M}]_{ij}^{\lambda}$	5.	$\mathcal{D}_5(\mathbf{M}) = f(\mathbf{\Lambda}(\mathbf{M}))$
2.	$\mathcal{D}_2(\mathbf{M};\alpha;\lambda) = \alpha \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}[\mathbf{M}]_{ij}^{\lambda}$	6.	$\mathcal{D}_6(\mathbf{M};\alpha,\lambda) = \alpha \cdot \sum_{k=1}^{K}\left(\prod_{i=1}^{n_k} VS_i(\mathbf{M})\right)_k^{\lambda}$
3.	$\mathcal{D}_3(\mathbf{M};\alpha) = \alpha \cdot \det(\mathbf{M})$	7.	$\mathcal{D}_7(\mathbf{M};\alpha,\lambda) = \alpha \cdot \sum_{i=1}^{n}\sum_{j=1}^{n}([\mathbf{M}]_{ij}^{\lambda} + [\mathbf{M}]_{ij})$
4.	$\mathcal{D}_4(\mathbf{M};\alpha,\lambda) = \alpha \cdot \sum_{i=0}^{n} \lvert c(Ch(\mathbf{M};x))_i^{\lambda}\rvert$	8.	$\mathcal{D}_8(\mathbf{M};\alpha,\lambda) = \alpha \cdot \max_{ij}([\mathbf{M}]_{ij}^{\lambda})$

M is a graph-theoretical matrix, n the matrix dimension, $c(Ch(\mathbf{M};x))_i$ the ith coefficient of the characteristic polynomial of **M**, $\mathbf{\Lambda}(\mathbf{M})$ indicates the graph spectrum (i.e., the set of eigenvalues of **M**), and α and λ are real parameters. In function $\mathcal{D}_6$, $VS_i(\mathbf{M})$ is the ith matrix row sum, K the total number of selected graph fragments, and n_k the number of vertices in the kth fragment. a_{ij} indicates the elements of the adjacency matrix that are equal to 1 for pairs of adjacent vertices and zero otherwise.

Other topological indices can be obtained by using suitable functions applied to → *local vertex invariants*; the most common functions are atom and/or bond additive, resulting into descriptors, which correlate well physico-chemical properties, that are atom and/or bond additive themselves. → *Zagreb indices* and → *ID numbers* are derived according to this approach.

Some functions to derive molecular descriptors $\mathcal{D}$ from local vertex invariants, denoted by $\mathcal{L}$, are listed in Table G3. It should be noted that function $\mathcal{D}_4$, that is, the well-known Randić-type formula for $\alpha = 1$ and $\lambda = -1/2$, is restricted to pairs of adjacent vertices, a_{ij} being the elements of the → *adjacency matrix*, which are equal to 1 only for pairs of adjacent vertices and zero otherwise. Function $\mathcal{D}_6$ is an extension of function $\mathcal{D}_4$ to any type of graph fragments as in the → *Kier–Hall connectivity indices*. Function $\mathcal{D}_7$ gives → *autocorrelation descriptors*, while function $\mathcal{D}_8$ gives → *maximum auto-crosscorrelation descriptors*. Moreover, similar functions can be applied to → *local edge invariants* $\mathcal{L}_{ij}$ in place of local vertex invariants $\mathcal{L}_i$ so that other sets of molecular descriptors can be generated.

Table G3 Classical functions to derive molecular descriptors from local vertex invariants.

ID	Function	ID	Function
1.	$\mathcal{D}_1(\mathcal{L};\alpha,\lambda)=\alpha\cdot\sum_{i=1}^{A}\mathcal{L}_i^{\lambda}$	5.	$\mathcal{D}_5(\mathcal{L};\alpha,\lambda)=\alpha\cdot\sum_{i=1}^{A}\sum_{j=1}^{A}(\mathcal{L}_i\cdot\mathcal{L}_j)^{\lambda}\quad j\neq i$
2.	$\mathcal{D}_2(\mathcal{L};\alpha,\lambda)=\alpha\cdot\left(\prod_{i=1}^{A}\mathcal{L}_i\right)^{\lambda}$	6.	$\mathcal{D}_6(\mathcal{L};\alpha,\lambda)=\alpha\cdot\sum_{k=1}^{K}\left(\prod_{i=1}^{n_k}\mathcal{L}_i\right)_k^{\lambda}$
3.	$\mathcal{D}_3(\mathcal{L};\alpha)=\alpha\cdot\max_{i\in\mathcal{V}}(\mathcal{L}_i)$	7.	$\mathcal{D}_7(\mathcal{L};\alpha,\lambda,k)=\alpha\cdot\sum_{i=1}^{A}\sum_{j=1}^{A}(\mathcal{L}_i\cdot\mathcal{L}_j)^{\lambda}\cdot\delta(d_{ij};k)$
4.	$\mathcal{D}_4(\mathcal{L};\alpha,\lambda)=\alpha\cdot\sum_{i=1}^{A}\sum_{j=1}^{A}a_{ij}(\mathcal{L}_i\cdot\mathcal{L}_j)^{\lambda}$	8.	$\mathcal{D}_8(\mathcal{L};\alpha,\lambda,k)=\alpha\cdot\max_{i,j\in\mathcal{V}}[(\mathcal{L}_i\mathcal{L}_j)^{\lambda}\cdot\delta(d_{ij};k)]$

$\mathcal{L}_i$ and $\mathcal{L}_i$ are local invariants associated with the vertices v_i and v_j, respectively. A is the number of graph vertices, $\mathcal{V}$ denotes the set of graph vertices, and $\delta(d_{ij};k)$ is a Dirac delta function equal to 1 for pairs of vertices at topological distance d_{ij} equal to k and zero otherwise. In function $\mathcal{D}_4$, a_{ij} indicates the elements of the adjacency matrix, which are equal to 1 for pairs of adjacent vertices and zero otherwise. In function $\mathcal{D}_6$, the summation goes over fragments of a given type, K is the total number of selected graph fragments, and n_k is the number of vertices in the kth fragment.

Another way to derive topological indices is by generalizing the existing indices or graph-theoretical matrices. → *Kier–Hall connectivity indices*, → *higher order Wiener numbers*, → *generalized Wiener indices*, → *variable Zagreb indices*, → *generalized expanded Wiener numbers*, and → *generalized Hosoya indices* are all examples of the generalization of existing indices, while → *generalized distance matrix*, → *expanded distance matrices*, and → *graphical matrices* are examples of generalized matrices.

Several → *fragment topological indices* can be derived by any topological index calculated for molecular subgraphs.

Particular topological indices are derived from weighted molecular graphs where vertices and/or edges are weighted by quantities representing some 3D features of the molecule, like those obtained by → *computational chemistry*. The graph invariants obtained in this way encode both information on molecular topology and → *molecular geometry*. Examples of these topological descriptors are → *BCUT descriptors*, → *electronic-topological descriptors*, → *electron charge density connectivity index*, and several descriptors obtained from → *weighted matrices*.

→ *Triplet topological indices* were proposed based on a general matrix–vector multiplication approach and several → *combined descriptors* are combinations of existing descriptors.

Several graph invariants can also be derived by the **vector–matrix–vector multiplication approach** (or **VMV approach**) proposed by Estrada [Estrada, Rodriguez *et al.*, 1997; Estrada and Rodriguez, 1997; Estrada, 2001; Estrada and Gutierrez, 2001]. This approach allows to generate graph invariants $\mathcal{D}$ according to the following equation:

$$\mathcal{D}(\mathbf{M},\mathbf{v}_1,\mathbf{v}_2;\alpha,\lambda)=\alpha\cdot(\mathbf{v}_1^{\mathrm{T}}\cdot\mathbf{M}^{\lambda}\cdot\mathbf{v}_2)$$

where $\mathbf{v}_1$ and $\mathbf{v}_2$ are two column vectors collecting atomic properties or local vertex invariants, $\mathbf{M}$ is a graph-theoretical matrix, and α and λ are two real parameters. Examples of well-known molecular descriptors derived from the VMV approach are reported in Table G4; moreover, all the → *TOMOCOMD descriptors* are calculated by this approach.

Table G4 Examples of molecular descriptors derived from VMV approach.

M	v_1	v_2	λ	α	Descriptor
D	1	1	1	1/2	Wiener index, W
I	δ	δ	1	1	First Zagreb index, M_1
A	δ	δ	1	1	Second Zagreb index, M_2
D	1	1	−1	1/2	Harary index, H
A	$\delta^{-0.5}$	$\delta^{-0.5}$	1	1/2	Randić connectivity index, $^1\chi$
A	$\sigma^{-0.5}$	$\sigma^{-0.5}$	1	(1/2) $[B/(C+1)]$	Balaban distance connectivity index, J

D is the → *distance matrix*, **I** the → *identity matrix*, and **A** the → *adjacency matrix*. δ indicates the → *vertex degree* and σ the → *distance degree*; B is the number of graph edges and C the number of rings.

Another general procedure to generate graph invariants is that used to calculate the so-called **Molecular Descriptor Family** (MDF) [Jäntschi, 2004a, 2004b, 2005; Bolboacâ and Jäntschi, 2005b, 2005, 2006, 2007]. This procedure utilizes both topological and geometrical distances between atoms, 6 atomic properties as the weighting scheme for graph vertices, 24 formulas for interaction descriptors, 6 overlapping interaction models, 4 fragmentation criteria, and 19 fragmental property selector functions. To all 131 328 resulted values, 6 linearization operations are applied, and finally it results in a number of 787968 MDF values for a given molecule.

Graph invariants have been successfully applied in characterizing the structural similarity/dissimilarity of molecules and in QSAR/QSPR modeling.

Due to the large proliferation of graph invariants, the result of many authors following the procedures outlined above and other general schemes, some rules are needed to critically analyze such invariants, paying particular attention to their effective role in correlating physicochemical properties, biological and other experimental responses, and their chemical meaning. In this respect, a list of desirable attributes for topological indices was suggested by Randić [Randić, 1991b] (see Table M11).

Theoretical studies on variability (e.g., covariance) and correlation of graph invariants were presented by Hollas [Hollas, 2002, 2003, 2005a, 2005b, 2005c, 2006; Hollas, Gutman *et al.*, 2005].

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

➢ **graph kernel** ≡ *pseudo*center → center of a graph

■ Graph of Atomic Orbitals (GAO)

GAO is a molecular representation that accounts for the electron configuration of different atoms in the molecule. It is defined as a molecular graph where each vertex represents a group of atomic orbitals of the respective atom [Mercader, Castro *et al.*, 2000; Toropov and Toropova, 2000a, 2000b, 2001b; Toropova and Toropov, 2000].

Let $\mathcal{G}$ be the → *H-filled molecular graph* of a molecule and $V(\mathcal{G}) = \{v_1, v_2, \ldots, v_A\}$ be the set of vertices in $\mathcal{G}$, then the graph of atomic orbitals (GAO) is obtained from $\mathcal{G}$ by replacing each of its vertex v_i with a set of n_i distinct vertices, the value n_i depending on the type of atom (Table G5). Two vertices in the GAO are adjacent if and only if they correspond to two different and adjacent atoms. Consequently, two vertices in the GAO, representing different groups of orbitals of the same atom, are not adjacent (Figure G4).

Table G5 Groups of atomic orbital for the most frequently occurring atoms in organic molecules.

Atom	Groups of atomic orbitals	*n*
H	$1s^1$	1
C	$1s^1\ 2s^2\ 2p^2$	3
N	$1s^1\ 2s^2\ 2p^3$	3
O	$1s^1\ 2s^2\ 2p^4$	3
F	$1s^1\ 2s^2\ 2p^5$	3
S	$1s^1\ 2s^2\ 2p^6\ 3s^2\ 3p^4$	5
Cl	$1s^1\ 2s^2\ 2p^6\ 3s^2\ 3p^5$	5
Br	$1s^1\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 3d^{10}\ 4s^2\ 4p^5$	8
I	$1s^1\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 3d^{10}\ 4s^2\ 4p^6\ 4d^{10}\ 5s^2\ 5p^5$	11

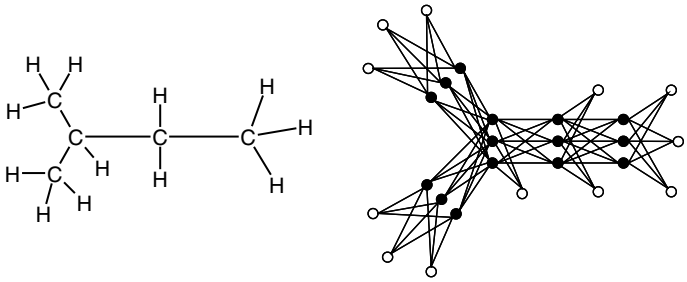


Figure G4 Graph of atomic orbitals (GAO) of 2-methylbutane.

GAO is an orbital-based graph-theoretical representation of molecules from which the common → *graph invariants*, such as connectivity, Zagreb, and Wiener indices, can be calculated, and thus it represents a source of orbital-based molecular descriptors, which can be generally called **GAO descriptors**. Note that → *orbital interaction graph of linked atoms* is another representation of molecules, which accounts for atom orbitals.

📖 [Toropov and Toropova, 2001b, 2003; Toropov, Toropova *et al.*, 2003, 2004]

- **graph potentials** → MPR approach
- **graph radius** → biodescriptors (⊙ DNA sequences)
- **graph-theoretical invariants** ≡ *graph invariants*
- **graph-theoretical matrices** → matrices of molecules
- **graph-theoretical shape coefficient** → shape descriptors (⊙ Petitjean shape indices)
- **graph theory** → graph
- **graph valence shells** ≡ *valence shell counts* → path counts
- **graph vertex complexity** → topological information indices
- **graph walk count** ≡ *molecular walk count* → walk counts
- **gravitational indices** → size descriptors
- **Green resonance energy** → delocalization degree indices
- **grid** → grid-based QSAR techniques
- **GRID/GOLPE method** → grid-based QSAR techniques (⊙ GRID method)

- **GRID method** → grid-based QSAR techniques
- **grid-based QSAR model** → grid-based QSAR techniques

grid-based QSAR techniques

These are QSAR techniques based on molecular descriptors calculated by embedding compounds into a fixed grid and encoding information about → *molecular interaction fields* (MIF) or physico-chemical properties related to ligand–receptor binding interactions [Esposito, Hopfinger *et al.*, 2003; Kubinyi, 2003a; Goodford, 2006].

Grid-based techniques could be used to model a variety of biological and physico-chemical properties; their most common application has, by far, been focused on ligand-target binding properties; moreover, grid-based descriptors can be compared to measure → *similarity/diversity* of molecules. Grid-based descriptors mainly characterize molecular shape and charge distribution in the 3D space responsible for nonbonding interactions involved in ligand–receptor binding. Therefore, they give the possibility of representing the molecular interaction regions of interest graphically, a big advantage in pharmacological studies.

The basic starting steps in grid-based techniques are the generation of three-dimensional structures of molecules, conformational search, and, in most of the cases, alignment of all the data set molecules according to some → *alignment rules*; alignment can be among the molecules themselves or with respect to a reference compound or a → *pharmacophore*.

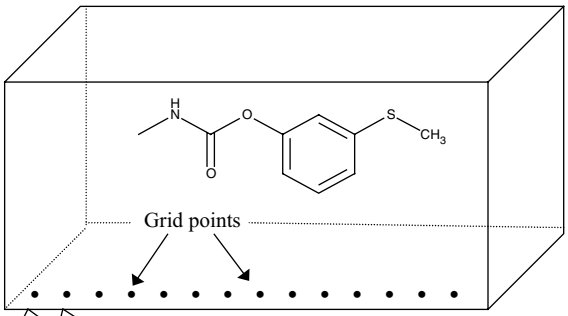
Once the molecules have been aligned, a rigid orthogonal grid of regularly spaced points representing an approximation of the binding site cavity space is established around each compound. A **grid** is a regular 3D array of $N_x \times N_y \times N_z$ points (N_G), that is, a lattice of grid points with N_x points along the *X*-axis, N_y points along the *Y*-axis, and N_z points along the *Z*-axis, each point **p** being characterized by the Cartesian coordinates (x, y, z) in the 3D space. The grid can be chosen to embed all the atoms of all compounds of the data set or else cover common particular regions of interest in the compounds. The density of the grid must be such as to sample the potential energies of the theoretically continuous scalar field reliably. A density sampling about 0.25–0.50 Å for sharp fields like molecular electrostatic potential seems to preserve field invariance [Todeschini, Moro *et al.*, 1997]. Steps of 2 Å are the most commonly used; however, a finer grid was suggested to obtain better predictive models [Liljefors, 1998].

Then, for each molecule embedded in the grid, specific values are calculated at every grid point; these, usually, are molecular **interaction energy values** or some function of them, taken as molecular descriptors within the framework of grid-based QSAR techniques. Usually, a reasonable selection of interaction energy values is performed based on energy cut-off values, selected molecular regions, or other specific criteria, depending on the method.

Finally, the classical **grid-based QSAR model** is estimated from interaction energy values. Linear models are the most common and take the following general form:

$$\hat{y}_i = b_0 + \sum_{j=1}^{F} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \sum_{z=1}^{N_z} b_{j,xyz} \cdot E_{ij,xyz} = \sum_{j=1}^{F} \sum_{k=1}^{N_G} b_{jk} \cdot E_{ijk}$$

where F is the number of fields used in the analysis, that is, the number of molecular interaction fields; N_x, N_y, and N_z are the number of grid points along the *X*-axis, *Y*-axis, and *Z*-axis, respectively; $N_G = N_x \times N_y \times N_z$ is the total number of grid points; and $E_{ij,\,xyz}$ is the potential interaction energy of the *i*th compound for the *j*th field in the grid coordinate (x, y, z). The k index of the last summation runs over the grid points in a vectorial representation (Figure G5).



	Steric field				Electronic field			
Compound	Point 1	Point 2	.....	Point N_G	Point 1	Point 2	.....	Point N_G
1								
2								
.....								
n								

Figure G5 Construction of the data set matrix in grid-based QSAR techniques.

Due to the large number of descriptors (commonly 15 000–20 000 for each field), multivariate regression analysis is usually performed by partial least squares regression (PLS), with or without → *variable selection*. Alternatively to grid-based QSAR models, → *similarity/diversity* between molecules can be measured by comparing their interaction fields.

The most popular grid-based QSAR techniques are *GRID* and *CoMFA*, based on molecular interaction fields derived from different probes; these methods together with a number of related techniques are discussed below; other related techniques are → *hydration free energy density* and → *Comparative Molecular Surface Analysis*.

A critical point of most of the grid-based techniques is the alignment of molecules, which determines to what extent the descriptors differ from one molecule to the next. Consequently, it substantially influences the results of the analysis. Hence, significant and reliable results can only be expected if the alignment was carried out properly and unambiguously. Often, the need for an alignment limits the application of these grid-based descriptors to homogeneous data sets, and even, then, the alignment is not always easily performed. To overcome this drawback, different research groups started to develop alignment-independent molecular descriptors. The first alignment-independent descriptors derived from scalar fields are *G-WHIM descriptors* [Todeschini, Moro *et al.*, 1997], based on the theoretical principles of the → *WHIM descriptors* but applied to → *molecular interaction fields*. *VolSurf* [Cruciani, Pastor *et al.*, 2000] and *GRIND* [Pastor, Cruciani *et al.*, 2000] *descriptors* are other grid-based descriptors independent of any previous alignment of the molecules.

• **GRID method**

This method is a computational procedure for detecting favorable binding sites on a molecule of known structure [Goodford, 1985, 1995]. A small molecule, such as water (the → *probe*) is used to generate → *interaction energy values* at all the grid points. The probe is rotated at the grid point until it makes the most favorable energetic interactions with the target. The final 3D array of energies constitutes the *GRID Map* for that particular probe. These energies are calculated by a sophisticated Empirical Force-Field method (*GRID Force Field*). Typically, grid spacing of 0.5 Å is used.

The main advantage of the GRID method is the great variety of available probes, represented by several functional groups such as water, methyl, ammonium, carboxylate, and benzene; in particular, among them there are probes that can both accept and donate a hydrogen bond (e.g., water), probes that cannot turn around (e.g., carbonyl probe), and a hydrophobic probe, named DRY.

The GRID method, unlike other grid-based techniques, explicitly takes the flexibility of the molecule into account in ligand–receptor interactions. The conformational flexibility of compounds is studied, allowing them to be attracted or repelled by the probe as the probe is moving around [Liljefors, 1998]. This algorithm works by dividing the target molecule into a flexible core and flexible side chain on an atom basis.

The **GRID/GOLPE method** is a QSAR methodology that combines interaction fields calculated by GRID with statistical analysis implemented in the program → *GOLPE* [Baroni, Clementi *et al.*, 1992; Baroni, Costantino *et al.*, 1993a, 1993b; GOLPE – Multivariate Infometric Analysis s.r.l., 2007]. Applications of GRID/GOLPE method are [Cruciani, Clementi *et al.*, 1993, 1994, 1998; Cruciani and Watson, 1994; Norinder, 1996b; Cruciani, Pastor *et al.*, 1997; Pastor, Cruciani *et al.*, 1997; Cinone, Höltje *et al.*, 2000; Fichera, Cruciani *et al.*, 2000; Lozano, Pastor *et al.*, 2000; Sippl, 2006].

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

• **Comparative Molecular Field Analysis** (≡ *CoMFA*)

This is the most popular QSAR approach among the grid-based QSAR techniques. CoMFA compares the molecular potential energy fields of a set of molecules and searches for differences and similarities that can be correlated with differences and similarities in the property values considered [Cramer III, Patterson *et al.*, 1988; Marshall and Cramer III, 1988; Cramer III, DePriest *et al.*, 1993; Folkers, Merz *et al.*, 1993a, 1993b; Kim, 1995a; Oprea and Waller, 1997; Martin, 1998; Norinder, 1998; Kubinyi, 2003a].

The first step of the CoMFA approach consists in the selection of a group of compounds having a common → *pharmacophore*, in the generation of three-dimensional structures of reasonable conformation and in their alignment.

The grid established around each compound is referred to the **CoMFA lattice**; the grid point distance is arbitrarily chosen (2 Å by default), bearing in mind that even small desirable distances lead to too great a number of grid points; the walls of the grid usually extend at least 4 Å beyond the union volume of the superimposed molecules. The rigidity of receptor walls derived from the use of a rigid grid is a basic assumption and approximation in the CoMFA method.

In the original CoMFA method only two fields of noncovalent ligand–receptor interactions were calculated: the steric field that is a → *Lennard-Jones 6–12 potential function* and the electrostatic field that is a → *Coulomb potential energy function*. Usually, the two fields are kept separate to facilitate the interpretation of the final results. As steric and electrostatic

fields account only for enthalpic contributions to free binding energy, other fields that account for solvation and entropic terms should be added. For example, hydrophobic interactions, which are entropic properties, are accounted for by the use of → *HINT* and → *molecular liphophilicity potential*. Since the Lennard-Jones potential is characterized by very steep increases in energy at short distances from the molecular surface, it was proposed to use the van der Waals volume intersections between probe and ligand molecule for steric field calculation [Sulea, Oprea *et al.*, 1997]; this molecular interaction field was called intersection volume field (INVOL).

Interaction energy values at the grid points are the **CoMFA descriptors** and are collected into a QSAR matrix whose rows represent the molecules and columns the grid points for each field considered.

Unreasonably large positive energy values, that is, grid points inside the molecules, are set constant at chosen cutoff value. They mainly derive from the large values of van der Waals repulsion caused by even a small overlap of ligand atoms and probe atoms. Moreover, grid points without variance, that is, within the volume shared by all molecules, or with small variance, that is, far away from the van der Waals surface of molecules, are discarded. Moreover, other parameters such as → *log P* or → *quantum-chemical descriptors* can be added as variables to the QSAR matrix after properly scaling. The combination of global parameters and CoMFA fields leads to a **mixed CoMFA approach** [Kubinyi, 1993b].

The **mixed CoMFA model** is defined as

$$\hat{Y}_i = b_0 + \sum_{j=1}^{F} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} \sum_{z=1}^{N_z} b_{j,xyz} \cdot E_{ij,xyz} + \sum_{j=1}^{J} b_j \cdot \Phi_{ij} = \sum_{j=1}^{F} \sum_{k=1}^{N_G} b_{jk} \cdot E_{ijk} + \sum_{j=1}^{J} b_j \cdot \Phi_{ij}$$

where Φ_{ij} is any global molecular property and J the total number of molecular properties considered.

Partial least squares (PLS) regression is usually performed to search a correlation between the thousands of CoMFA descriptors and biological responses. However, usually after → *variable selection*, the PLS model is transformed into and presented as a multiple regression equation to allow comparison with classical QSAR models.

Developments of the CoMFA approach have been also proposed based on a selection of molecule regions of interest for binding interactions [Cho and Tropsha, 1995; Cruciani, Clementi *et al.*, 1998; Tropsha and Cho, 1998].

A critical review of CoMFA applications is given in [Kim, Greco *et al.*, 1998] and a complete list of references 1993–1997 in [Kim, 1998].

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

• **Comparative Molecular Similarity Indices Analysis** (≡ *CoMSIA*)

CoMSIA is a method of measuring the similarity of molecules on the basis of their physicochemical properties. It implements the steric, electrostatic, hydrophobic, and hydrogen-bonding → *similarity indices* utilized in the molecular alignment program SEAL [Abraham *et al.*, 1994; Diudea, 1997d; Klebe, 1998; Klebe and Abraham, 1999].

Using the → *similarity score* A_F based on the weighted combination of steric, electrostatic, and hydrophobic properties, molecule alignment is performed starting from a random orientation of two molecules relative to each other; the best alignment is achieved with the maximum similarity score.

Moreover, → *molecular interaction fields* are calculated for each molecule in terms of similarity indices instead of the usual interaction potential functions, such as Lennard-Jones and Coulomb potential functions. Similarity fields are calculated representing the similarity between molecules and different probe atoms. In particular, the similarity values at the intersections of the regularly spaced grid (1.1 and 2.0 Å) relative to the *j*th physico-chemical property between the *i*th compound and a probe atom is calculated as

$$(A_F)_{ik,j} = \sum\nolimits_t w_{probe,j} \cdot w_{tj} \cdot e^{-\alpha \cdot r_{ik}^2}$$

where the summation goes over all atoms of the molecule, $w_{probe,j}$ and w_{tj} are, respectively, the actual value of the *j*th property of the probe and the *t*th atom of the target molecule, α is an attenuation factor, and r_{tk} is the geometric distance between the probe atom at the *k*th grid point and the *t*th atom of the molecule. Large values of α give rise to a strong distance-dependent attenuation of the similarity measures, that is, only local similarities are considered; otherwise, for small α values, global molecular features are of greater importance. The probe interaction with the molecule is, then, calculated for each grid point, including those inside the molecule atomic van der Waals radius, avoiding the need for cutoff as in CoMFA.

The studied properties are electrostatic, steric, hydrophobic, hydrogen-acceptor, and hydrogen-donor abilities; for electrostatic properties, the probe assumes charge $+1$, for steric properties radius 1 Å, for hydrophobicity and hydrogen-bonding abilities a value of $+1$.

These indices replace the distance functions used in the standard Lennard-Jones and Coulomb potential functions, which generate unrealistically extreme values as the surface of the considered molecule is approached.

CoMSIA results show regions of the compound that prefer or dislike the presence of a group with a specific physico-chemical property.

A molecular → *similarity matrix* can be obtained both from the similarity scores between pairs of molecule and any distance function applied to similarity fields [Klebe, Abraham *et al.*, 1994; Diudea, 1997d; Klebe, 1998; Kubinyi, Hamprecht *et al.*, 1998].

📖 [Hou, Li *et al.*, 2000; Anzini, Cappelli *et al.*, 2001; Ducrot, Andrianjara *et al.*, 2001; Makhija and Kulkarni, 2001b, 2002a; Zhu, Hou *et al.*, 2001; Buolamwini and Assefa, 2002; Doytchinova and Flower, 2002; Schleifer and Tot, 2002; Sreenivasa and Kulkarni, 2002; Wellsow, Machulla *et al.*, 2002; Xu, Zhang *et al.*, 2002; Assefa, Kamath *et al.*, 2003; Boström, Böhm *et al.*, 2003; Bringmann and Rummey, 2003; Liu, Yang *et al.*, 2003; Raichurkar and Kulkarni, 2003; Chen, Yao *et al.*, 2004; Kelkar, Pednekar *et al.*, 2004; Medina-Franco, Rodríguez-Morales *et al.*, 2004; Sutherland and Weaver, 2004; Jójárt, Martinek *et al.*, 2005; Zhao, Yu *et al.*, 2005]

• **G-WHIM descriptors** (≡ *Grid-Weighted Holistic Invariant Molecular descriptors*)
Based on a similar approach to that used to define → *WHIM descriptors*, G-WHIM descriptors are global molecular descriptors of → *molecular interaction fields* [Todeschini, Moro *et al.*, 1997; Todeschini and Gramatica, 1998].

Once the optimal choices for the grid have been made, the G-WHIM descriptors are used to condense the whole information contained in the scalar field constituted by the calculated → *interaction energy values* into a few global parameters, whose values are independent of the molecular orientation within the grid.

For each molecule, the G-WHIM descriptors are calculated by the following steps:

a. The molecule is freely and separately embedded in the center of the grid.

b. The interaction energy values are calculated at all the grid points by using the selected probe.

c. The interaction energy values are used as weighting scheme for the grid points. This is the main difference between G-WHIM and WHIM descriptors, for which the defined atomic properties are used to weight molecule atoms.

d. Finally, the G-WHIM descriptors are calculated in the same way as the WHIM descriptors, that is, by the calculation of a weighted covariance matrix, principal component analysis, and the calculation of statistical parameters on the projected points along each principal component (i.e., on the score values) (Figure G6).

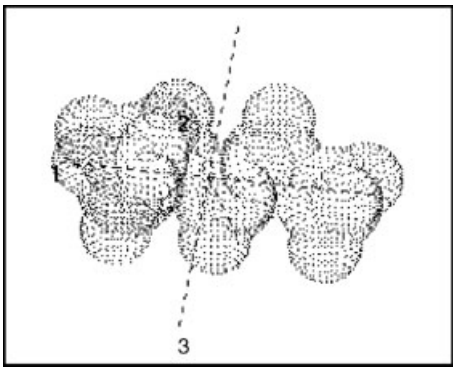


Figure G6 Principal axes of the grid interaction energy values of the 2-methylpentane.

It should be noted that only points with nonzero interaction energy are effective in the computation of the descriptors. Second, when the calculated interactions give both positive and negative values, interaction energy values cannot be used directly in this form as statistical weights, which must always be defined semipositive. In this case, the scalar field values are divided into two blocks: a grid-negative (positive) block containing the grid coordinates associated with negative (positive) interaction values, getting their absolute values and setting the positive (negative) values to zero.

This assumption leads to two sets of G-WHIM descriptors: one describing the positive part of the molecular field and the other describing the negative part.

Thus, for each region (positive (+) and negative (−)), the G-WHIM descriptors consist of 8 directional and 5 nondirectional molecular descriptors (26 for a complete description of each interaction field), calculated from each molecule:

$$\lambda_1(\pm) \quad \lambda_2(\pm) \quad \lambda_3(\pm) \quad \vartheta_1(\pm) \quad \vartheta_2(\pm) \quad \eta_1(\pm) \quad \eta_2(\pm) \quad \eta_3(\pm)$$

and

$$T(\pm), \quad A(\pm), \quad V(\pm), \quad K(\pm), \quad D(\pm)$$

The directional γ and the nondirectional G parameters, defined for → *WHIM descriptors* and containing information about the molecular symmetry, are not considered in the frame of the G-

WHIM approach as their meaning becomes doubtful, depending heavily upon the point sampling. However, information regarding molecular symmetry can be obtained by directly using the WHIM symmetry parameters.

The meaning of the G-WHIM descriptors is that previously defined for WHIM descriptors, but now the descriptors refer to the interaction molecular field instead of the molecule. For example, the eigenvalues λ_1, λ_2, and λ_3 relate to the interaction field size; the eigenvalue proportions ϑ_1 and ϑ_2 relate to the interaction field shape; the group of descriptors constituted by the inverse function of the kurtosis (k), that is, $\eta_m = 1/\kappa_m$ relate to the interaction field density along each axis. Moreover, global information about the interaction field is obtained as for → *global WHIM descriptors* (T, A, V, K, D), with the same meaning.

First [Todeschini, Moro *et al.*, 1997], the acentric factor was used as the shape descriptor, defined as $\omega = \vartheta_1 - \vartheta_3$, ranging from zero (spherical interaction field) to one (linear interaction field); this shape descriptor was later substituted by the global K shape descriptor.

It must be noted that the invariance to rotation of G-WHIM descriptors, that is, the independence of any molecular → *alignment rule*, is obtained if the grid points are dense enough. A too sparse distribution of grid points represents an inadequate sampling of the ideal scalar field and is not able to guarantee that the calculated scalar field is representative of the ideal scalar field in such a way as to preserve rotational invariance. If a molecule is placed into an infinite, isotropic, and even very dense grid, the scalar field $\mathcal{F}$ calculated at the grid points must contain the same information, independent of the molecule's orientation and depending only on the potential energy of the selected probe and the mathematical function representing the interaction. Thus $\mathcal{F}$ contains the whole information about the interaction properties of the molecule.

In practice, this ideal situation cannot be achieved, but it can be simulated by plunging the molecule into a finite grid: the aim is to represent the theoretical scalar field $\mathcal{F}$ by a finite sampling of this field.

G-WHIM descriptors can be calculated for any selected region of the field. To avoid irrelevant or unreliable chemical information due to long-range interactions, an energy cutoff criterion, which takes into account only interaction energy values relevant to the considered interaction (e. g., long-range or chemical interaction) is used. In this way, points far from the molecule and not contributing to the interaction are not included in the calculations, for example, the field values inside the van der Waals surface of the molecule are not considered. Moreover, specific chemical information is gained by using different energy cutoff values to select the regions, keeping in mind that the higher the cutoff value the smaller the considered region around the molecule (i. e., the total number of nonzero weighted grid points). For instance, surface points at a given cutoff value, that is, points on an isopotential energy surface, can be selected. G-WHIM descriptors calculated on the → *Connolly surface area* were called **MS-WHIM descriptors** [Bravi, Gancia *et al.*, 1997].

The ability to take the individual parameters provided by different cutoff values into account when defining different molecular interaction regions could possibly lead to a deeper chemical insight into molecular interactions and properties.

The G-WHIM approach integrates the information contained in → *WHIM descriptors* and overcomes any problem due to the alignment of the different molecules and the explosion of variables arising from traditional grid-based QSAR techniques, such as GRID and CoMFA.

🕮 [Ekins, Bravi *et al.*, 1999a, 1999b, 1999c; Zaliani and Gancia, 1999; Bravi and Wikel, 2000a, 2000b; Cosentino, Moro *et al.*, 2000; Gancia, Bravi *et al.*, 2000; Ekins, Durst *et al.*, 2001; Baumann, 2002b; Snyder, Sangar *et al.*, 2002]

• **Self-Organizing Molecular Field Analysis** (≡ *SOMFA*)
SOMFA is a grid-based approach that does not use a probe to determine interaction energies. Instead, each grid point is assigned the shape or → *molecular electrostatic potential* (MEP) value: (a) shape is represented by binary values equal to 1 for points inside the van der Waals envelope and zero otherwise; (b) electrostatic potential values at grid points are calculated from partial charges distributed across the atom centers [Robinson, Winn *et al.*, 1999].

Crucial to SOMFA is the concept of *mean centered activity* (A_i^c), which is the activity of a molecule obtained by subtracting the mean activity of the training set molecules. Therefore, the molecule activity has a scale such that the most active compounds have positive values and the least active ones have negative values. The mean centered activity is used as the weighting scheme for the grid points: the value of the shape or electrostatic potential at every grid point for a given molecule is multiplied by the mean centered activity. This procedure allows grid points to filter in such a way as to highlight the features that differentiate high-affinity and low-affinity compounds.

In general, a SOMFA master grid can be trained on any calculable molecular property that can be distributed in a grid by using all the training compounds. This grid is constructed by overlaying all the individual molecular grids; the value at the grid point (*x*,*y*,*z*) of the master grid is defined as

$$SOMFA(x,y,z) = \sum_{i=1}^{n} P_i(x,y,z) \cdot A_i^c$$

where *n* is the number of training compounds, P_i is the property of the *i*th compound at grid point (*x*, *y*, *z*), and A_i^c is its mean centered activity. Maximum and minimum grid values of the master grid can be displayed to highlight regions favorable or unfavorable to activity.

Finally, for a given property P, a SOMFA molecular descriptor can be calculated for each molecule as

$$SOMFA_i(P) = \sum_x \sum_y \sum_z P_i(x,y,z) \cdot SOMFA(x,y,z)$$

where $P_i(x, y, z)$ is the property value for the *i*th molecule at point (*x*, *y*, *z*) and *SOMFA*(*x*, *y*, *z*) is the value of the master grid at the same point.

A linear combination of the activities calculated from shape and MEP properties was also proposed as a model for the better prediction of activity:

$$\hat{A}_i = \alpha \cdot A_i^{SH} + (1-\alpha) \cdot A_i^{MEP}$$

where α is a parameter that can be optimized to maximize the predictive power of the model.

SOMFA, like other grid-based techniques, needs a reliable procedure for molecule alignment and suffers from the need for the bioactive conformations.

🕮 [Amat, Besalú *et al.*, 2001; Liu, Yin *et al.*, 2001b; Klein, Kaiblinger *et al.*, 2002; Klocker, Wailzer *et al.*, 2002a; Smith, Sorich *et al.*, 2003; Tuppurainen, Viisas *et al.*, 2004; Martinek, Ötvös *et al.*, 2005]

• **Voronoi Field Analysis** (≡ *VFA*)
Among the grid-based QSAR techniques, Voronoi Field Analysis was proposed with the aim of reducing the very large number of potential → *interaction energy values* assigned to the grid points of → *molecular interaction fields* [Chuman, Karasawa *et al.*, 1998]. Interaction energy values are assigned to each of the **Voronoi polyhedra** into which the superimposed molecular space is decomposed using an approach similar to that used in the → *Voronoi binding site models*.

The molecules in the data set are superimposed considering their conformational flexibility and the total volume of the superimposed molecular space is calculated, after expansion with a 4.0 Å of thick shell outside the surface. The total volume is divided into Voronoi polyhedra each including an atom as a reference point. The outer boundaries of the most outer subspaces are not bisecting planes between two reference points.

The reference points are assigned by selecting as the template the simplest molecule or the unsubstituted one, and the position in the template of the atoms including the hydrogens are automatically defined as reference points. As the second step, the largest compound in terms of number of atoms is selected and the atomic positions of this compound are compared with the previous ones: new reference points are defined if no reference point within 1.0 Å of each atom of this molecule is found. The remaining molecules are then selected in order of decreasing size and atomic positions are compared with reference points as in the previous step. As the final step, a Voronoi polyhedron is assigned to each reference point with its own molecular space; the *Voronoi polyhedron* is a region delimited by a set of planes, each of which bisects as well as is perpendicular to the line connecting the reference point with each of the neighboring reference points of the adjacent regions. In other words, each polyhedron is a set of points closer to the reference point than to any other.

After obtaining the decomposition of the superimposed molecular space into Voronoi polyhedra, a grid exactly containing the expanded molecular surface is defined, with grid points spaced at 0.3 Å, and potential energy values are calculated at each of the lattice points located inside the surface. The steric and electrostatic potential energy values calculated at each *k*th grid point are transformed into the corresponding Voronoi potential energy values VE_g by summing all the contributions of the grid points belonging to the *g*th Voronoy polyhedron VP_g:

$$VE_g^T = \sum\nolimits_k E_k^T \quad k \in VP_g$$

where superscript "*T*" denotes any kind of potential energy value (steric, electrostatic, etc.).

📖 [Aurenhammer, 1991]

• **GRIND descriptors** (≡ *GRid INdependent Descriptors*)
These are molecular descriptors derived from → *molecular interaction fields* (MIF) calculated by using different → *probes* and representing the geometrical relationships among MIF regions [ALMOND – Multivariate Infometric Analysis s.r.l., 2007; Pastor, Cruciani *et al.*, 2000; Cruciani, Pastor *et al.*, 2001b; Gratteri, Cruciani *et al.*, 2001].

The procedure for obtaining GRIND descriptors is threefold: (a) the molecular interaction field (MIF) is computed, (b) the MIF is filtered, considering only the regions in which the intensity of the field is maximum at relative distances, and (c) the GRIND descriptors are calculated on the basis of the maximum value of the property products obtained at different distances.

Molecular interaction fields are calculated according to the → *GRID method* and considering three different probes: the DRY probe for representing hydrophobic interactions, the O probe (carbonyl oxygen) that represents hydrogen-bonding donor groups, and the N1 probe (amide nitrogen) to represent hydrogen-bonding acceptor groups. By default, a grid-spacing of 0.5 Å is used with the grid extending 5 Å beyond a molecule.

In the second step, the most interesting regions are selected as those characterized by favorable interaction (negative) energies. Therefore, the method extracts from each MIF a number of grid points (about 100) that represent favorable probe–ligand interaction regions by using two optimality criteria: the intensity of the field at a grid point and the mutual distances between the chosen grid points.

In the third step, GRIND descriptors are calculated according to the distance between the grid points, basically using auto- (same probe) and cross-correlation (combinations of pairs of different probes) transforms (Table G6).

Table G6 Correlogram types used to generate GRIND descriptors.

Correlogram	Probe 1	Probe 2	Interaction
1	DRY	DRY	Hydrophobic
2	O	O	Hydrogen bond donor
3	N1	N1	Hydrogen bond acceptor
4	DRY	O	Hydrophobic and hydrogen bond donor
5	DRY	N1	Hydrophobic and hydrogen bond acceptor
6	O	Ni	Hydrogen bond donor and acceptor

Unlike the classical → *autocorrelation descriptors*, only the highest product of interaction energies per distance bin is stored as GRIND descriptor (**MACC-2 transform**). This difference is responsible for the reversibility of GRIND descriptors. Unlike most of the grid-based methods, GRIND descriptors are also independent of the molecule alignment.

🕮 [Afzelius, Masimirembwa *et al.*, 2002; Cruciani, Pastor *et al.*, 2002; Fontaine, Pastor *et al.*, 2003, 2004; Lapinsh, Prusis *et al.*, 2003; Crivori, Zamora *et al.*, 2004; Gratteri, Romanelli *et al.*, 2004; Aureli, Cruciani *et al.*, 2005; Sciabola, Alex *et al.*, 2005; Gedeck, Rohde *et al.*, 2006; Pastor, 2006; Urbano-Cuadrado, Carbó *et al.*, 2007]

• **VolSurf descriptors**

VolSurf descriptors, such as G-WHIM and GRIND, encode information present in → *molecular interaction fields* (MIF) calculated by the → *GRID* force field parametrization [Crivori, Cruciani *et al.*, 2000; Cruciani, Crivori *et al.*, 2000; Cruciani, Pastor *et al.*, 2000; Mannhold, Berellini *et al.*, 2006].

VolSurf descriptors were designed to compress relevant MIF information into a few alignment-independent descriptors encoding information about molecular size and shape, the overall distribution of hydrophobic and hydrophilic regions and the balance between them (Table G7).

Interaction fields obtained with different probes (H_2O, DRY, O) are analyzed and the volume and surface of the regions that encompass interaction energy values under certain cutoff limits,

together with some additional variables that express their geometrical spatial distribution, are calculated.

From each MIF, a unique framework (a volume and/or a surface) related to specific molecular properties is constructed. Similar to 2D images, each 3D molecular field map is made of a regular lattice of boxes called *voxels*, which represent attractive and repulsive forces between a probe and a molecule. Each voxel is defined by a volume, by a surface, and by an interaction energy value. By contouring the voxels at different energy levels, different images can be obtained. The images are then used to compute the volumes and the surfaces related to the contouring method selected. In the building phase of volumes, the voxels are grouped by a shape function that assigns the value of 1 to voxels, inside an energy range, and 0 to all other voxels. Subsequently, a simple summation over the selected voxels yields back the overall volume for the considered property. For example, when computing a molecular volume, under standard conditions, all voxels with an energy interaction greater than $+0.2$ kcal/mol are marked as 1 and then counted. As a defined volume is associated with each voxel, the total volume is obtained by multiplying the number of selected voxels by their volume. Conversely, when a hydrophilic volume is computed, only the voxels with interaction energy below -0.2 kcal/mol are marked as 1 and then counted.

Table G7 List of VolSurf descriptors [Zamora, Oprea *et al*., 2001; Mannhold, Berellini *et al*., 2006].

Symbol	Probe	Meaning
V	H_2O	Water-excluded volume (in $Å^3$) at $+0.2$ kcal/mol energy
S	H_2O	Accessible surface of the water interaction field at $+0.2$ kcal/mol energy
R	H_2O	Rugosity, defined as the ratio of volume (V) to surface (S)
G	H_2O	Globularity, defined as the ratio of surface (S) to the surface area of a sphere with the same volume (V)
W1–W8	H_2O	Volume of the hydrophilic interactions at eight different energy levels: -0.2, -0.5, -1.0, -2.0, -3.0, -4.0, -5.0, -6.0 kcal/mol
IW1–IW8	H_2O	Integy moments at the same energy levels as W1–W8
CW1–CW8	H_2O	Capacity factors, defined as the ratio of the hydrophilic surface to the total molecular surface (S), at the same energy levels as W1–W8
D1–D8	DRY	Volume of the hydrophobic interactions at eight energy levels: -0.2, -0.4, -0.6, -0.8, -1.0, -1.2, -1.4, -1.6 kcal/mol
ID1–ID8	DRY	Integy moments at the same energy levels as D1–D8
A	DRY	Strength of the amphiphilic moment
POL	DRY	Polarizability
BV1–BV3	DRY/H_2O	Best volumes calculated at the three maximum hydrophobic/hydrophilic regions
Emin1–Emin3	DRY/H_2O	Energy values of the three lowest energy minima
D12, D13, D23	DRY/H_2O	Distances between the three energy minima
HL	DRY/H_2O	Hydrophilic–lipophilic balance, defined as the ratio of the volume of hydrophobic regions at -4 kcal/mol to the volume of hydrophobic regions at -0.8 kcal/mol
CPP	DRY/H_2O	Critical packing parameter, defined as the ratio of surface of the hydrophobic regions at -0.6 kcal/mol to the surface of the hydrophilic regions at -3 kcal/mol

(*Continued*)

Table G7 (*Continued*)

Symbol	Probe	Meaning
Wp1–Wp8	O	Volume of the interactions with the probe O at eight different energy levels
HB1–HB8	O	Hydrogen-bond donor capacity of the target, defined as the difference between the volume of the hydrophilic interactions (W1–W8) and volume of the O probe interactions (Wp1–Wp8)
E	–	Elongation
EEFR	–	Elongation/elongation-fixed ratio
MW	–	Molecular weight
log *P*	DRY/H_2O	octanol–water partition coefficient
D	H_2O	Diffusivity
α	–	polarizability
HBP	Polar	Hydrogen bonding parameter

VolSurf descriptors related to molecular size and shape are: the *molecular volume* represents the water-excluded volume (in $Å^3$), calculated as the volume enclosed by the water-accessible surface computed at a repulsive value of $+0.20$ kcal/mol; the *molecular surface* represents the accessible surface (in $Å^2$) traced out by a water probe interacting at $+0.20$ kcal/mol when it rolls over the target molecule; the **rugosity** is a measure of a molecular wrinkled surface defined as the ratio of volume/surface, the smaller the ratio the larger the rugosity; the molecular **globularity** is defined as the ratio of the molecular surface over the surface area of a sphere of the same volume V. Globularity is 1.0 for perfect spherical molecules. It assumes values greater than 1.0 for real spheroidal molecules. Globularity is also related to molecular flexibility. **Elongation** represents the maximum extension a molecule could reach if properly stretched. **Elongation/Elongation-Fixed Ratio**, denoted as EEFR, represents the portion of the extension given by the rigid part of the molecule; within each molecule a fixed part is considered as the rigid core, and EEFR is defined as the ratio of the elongation to the elongation of the rigid core.

Another set of VolSurf descriptors consists of descriptors of hydrophilic regions. These are hydrophilic descriptors defined as the volume of the molecular envelope, which is accessible to and attractively interacts with water molecules. The volume of this envelope varies with the level of interaction energies. In general, hydrophilic descriptors computed from molecular fields of -0.2 to -1.0 kcal/mol account for polarizability and dispersion forces, whereas descriptors computed from molecular fields of -1.0 to -6.0 kcal/mol account for polar and H-bond donor–acceptor regions. Moreover, **best hydrophilic volumes** are six molecular descriptors that refer to the best three hydrophilic interactions generated by a water molecule; these are the first, second, and third volumes calculated in separate regions of maximum hydrophilicity. The best volumes are measured at -1.0 and -3.0 kcal/mol. **Capacity factors** are defined as the ratio of the hydrophilic surface over the total molecular surface. Capacity factors are calculated at eight different energy levels, the same levels used to compute the hydrophilic descriptors.

VolSurf descriptors of hydrophobic regions are: *hydrophobic descriptors* defined in terms of the volume of the molecular envelope generating attractive hydrophobic interactions and **best hydrophobic volumes** that represent the best three hydrophobic interactions generated by the DRY probe and measured at -0.6 and -1.0 kcal/mol. VolSurf computes hydrophobic

descriptors at eight different energy levels adapted to the usual energy range of hydrophobic interactions (i.e., from -0.2 to -1.6 kcal/mol).

Integy moments (*INTEraction enerGY moments*), like dipole moments, express the unbalance between the barycenter of a molecule and the center of its hydrophilic or hydrophobic regions. When referring to hydrophilic regions, integy moments are vectors pointing from the center of mass to the center of the hydrophilic regions; when the integy moment is high, there is a clear concentration of hydrated regions in only one part of the molecular surface. Small moments indicate that the polar moieties are either close to the center of mass or they balance at opposite ends of the molecule, so that their resulting barycenter is close to the center of the molecule. When referring to hydrophobic regions, integy moments measure the unbalance between the center of mass of a molecule and the center of the hydrophobic regions. All the integy moments can be visualized in the real 3D molecular space.

Local interaction energy minima are VolSurf descriptors representing the energy of interaction (in kcal/mol) of the best three local energy minima between the water probe and the target molecule. Alternatively, the minima can refer to the three deepest local minima in the → *molecular electrostatic potential* (MEP). They are generated both for probes H_2O and DRY. The *energy minima distances* are VolSurf descriptors that represent the distances between all combinations between the best three local energy minima of a molecular interaction field. They are generated both for probes H_2O and DRY. **Hydrophilic–lipophilic balance** is defined as the ratio of the volume of hydrophilic regions measured at -4.0 kcal/mol over the volume of hydrophobic regions measured at -0.8 kcal/mol. The balance describes which effect dominates in the molecule or if they are roughly equally balanced. The → *amphiphilic moment* is defined as a vector pointing from the center of the hydrophobic domain to the center of the hydrophilic domain. The vector length is proportional to the strength of the amphiphilic moment, and it may determine the ability of a compound to permeate a membrane. In contrast to the hydrophilic–lipophilic balance, the **critical packing parameter** refers just to molecular shape, being defined as

$$\mathrm{CPP} = \frac{volume(hydrophobic\ regions)}{surface(hydrophilic\ regions) \cdot length(hydrophobic\ regions)}$$

Lipophilic and hydrophilic calculations are performed at -0.6 and -3.0 kcal/mol, respectively. Critical packing is a good parameter to predict molecular packing such as in micelle formation and may be relevant in solubility studies in which melting point plays an important role.

Other VolSurf descriptors are: the → *molecular weight*; the → *log P*, computed via a linear model derived by fitting VolSurf descriptors to experimental data; the → *hydrogen bonding parameter* used to describe the H-bonding capacity of a molecular target, as obtained with a polar probe (e.g., the water probe); the **diffusivity**, which controls the dispersion of chemical in water fluid and is calculated according to a modified Stokes–Einstein equation; the → *polarizability*, defined as an estimate of the average molecular polarizability and calculated from the structure of a compound according to Cruciani [Cruciani, Crivori *et al.*, 2000] and the additive method of Miller [Miller, 1990b].

🕮 [Alifrangis, Christensen *et al.*, 2000; Testa and Bojarski, 2000; Ekins, Durst *et al.*, 2001; Filipponi, Cruciani *et al.*, 2001; Zamora, Oprea *et al.*, 2001, 2003; Cruciani, Pastor *et al.*, 2002; Oprea, 2002b; Oprea, Zamora *et al.*, 2002; Cruciani, Meniconi *et al.*, 2003; Fontaine, Pastor *et al.*, 2003; Menezes, Lopes *et al.*, 2003; Cianchetta, Mannhold *et al.*, 2004; Crivori, Zamora

et al., 2004; Jacobs, 2004; Kovatcheva, Golbraikh *et al.*, 2004; Schefzik, Kibbey *et al.*, 2004; Stærk, Skole *et al.*, 2004; Aureli, Cruciani *et al.*, 2005; Caron and Ermondi, 2005; de Cerqueira Lima, Golbraikh *et al.*, 2006; Lamanna, Catalano *et al.*, 2007]

- **4D-QSAR analysis**

4D-QSAR analysis is a grid-based technique that explicitely accounts for ligand conformational flexibility and explores different alignments of compounds [Hopfinger, Wang *et al.*, 1997; Albuquerque, Hopfinger *et al.*, 1998]. With respect to 3D-QSAR analysis, the fourth dimension of 4D-QSAR analsysis is the ensemble sampling.

Unlike the other grid-based techniques, molecular descriptors are not derived from → *molecular interaction fields* but from the partition of molecules into different parts that are expected to have different types of interactions with receptor sites.

To generate 4D-QSAR analysis descriptors, the following procedure is used. First, 3D molecular structures are constructed and the conformers of the minimum energy state are used as the initial structures of conformational search. The reference cell grid is constructed to arrange the largest compound in the data set and usually has a grid spacing of 1.0 Å. In the second step, atoms of each molecule are classified into eight types of → *Interaction Pharmacophore Elements* (IPEs) that are the generic type (i.e., any type of atom), nonpolar atom, positively charged atom, negatively charged atom, hydrogen-bond acceptor, hydrogen-bond donor, aromatic atom, and nonhydrogen atom (see Table 4-1 in → *4D-Molecular Similarity Analysis*).

The third step is to estimate the → *Conformational Ensemble Profile* (CEP) for each compound by molecular dynamic simulation; this profile encodes those conformations selected on the basis of the Boltzmann distribution. Then, different alignments are selected to compare the molecules of the training set. In the following step, each conformation of a molecule is placed in the reference grid space on the basis of the alignment scheme being explored and the thermodynamic probability of each grid cell occupied by each IPE type is computed.

The normalized occupancy of each grid cell by each IPE type over the CEP of each molecule, for a given alignment, constitutes a unique set of molecular descriptors referred to as **Grid Cell Occupancy Descriptors** (GCODs). These descriptors were used directly to estimate QSAR models and indirectly in 4D-Molecular Similarity Analysis to generate a set of → *spectral indices*.

🕮 [Ravi, Hopfinger *et al.*, 2001; Santos-Filho and Hopfinger, 2001, 2002; Esposito, Hopfinger *et al.*, 2003; Hong and Hopfinger, 2003; Romeiro, Albuquerque *et al.*, 2005]

- ➢ **Grid Cell Occupancy Descriptors** → grid-based QSAR techniques (⊙ 4D-QSAR analysis)
- ➢ **grid region selection methods** → variable selection
- ➢ **GRID electrostatic energy function** → molecular interaction fields (⊙ electrostatic interaction fields)
- ➢ **Grid-Weighted Holistic Invariant Molecular descriptors** ≡ *G-WHIM descriptors* → grid-based QSAR techniques
- ➢ **GRIND descriptors** → grid-based QSAR techniques
- ➢ **Grob inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- ➢ **group charge transfer** ≡ *charge transfer constant* → electronic substituent constants

■ **group contribution methods** (GCM)
Group contribution methods search for relationships between structural properties and a physico-chemical or biological response based on the following general models:

$$y_i = f(\mathrm{G}_1, \mathrm{G}_2, \ldots, \mathrm{G}_m; n_1, n_2, \ldots, n_m)$$

where the experimental property y_i for the ith compound is a function of m group contributions G_j and their occurrences n_j [Reinhard and Drefahl, 1999]. The **group contributions**, also known as **fragmental constants**, are numerical quantities associated with substructures of the molecule, such as single atoms, atom pairs, atom-centered substructures, molecular fragments, functional groups, and so on. The specification of the structural groups depends on the particular GCM scheme adopted. → *Cluster expansion of chemical graphs* is an example of a group contribution method based on all the connected subgraphs of the molecular graph.

Generally, the application of GCM to a molecule requires the following steps:

1 Identification of all groups in the molecule applicable to the particular GCM scheme. An automated search for substructures of interest for a given property is performed by the *CASE approach*.
2 Calculation of fragmental constants measuring contributions to the molecular property of the considered fragments by employing the function associated with the particular GCM.
3 Evaluation of some correction factors that should account for interactions among molecular groups.

Linear GCM models are defined as the following:

$$y_i = k_0 + \sum_{j=1}^{m} \mathrm{G}_j \cdot \mathrm{I}_{ij} \quad \text{or} \quad y_i = k_0 + \sum_{j=1}^{m} \mathrm{G}_j \cdot n_{ij}$$

where k_0 is a model-specified constant, j runs over the m group contributions defined within the GCM scheme, G_j is the contribution of the jth group. I_{ij} and n_{ij} are → *substructure descriptors*, and in particular, I_{ij} is a binary variable taking a value equal to 1 if the jth group is present in the ith molecule, zero otherwise, while n_{ij} is the number of times the jth group occurs in the ith molecule.

Nonlinear GCM models are usually defined as

$$y_i = k_0 + \sum_{j=1}^{m} \mathrm{G}_j \cdot n_{ij} - \left(\sum_{j=1}^{m} \mathrm{G}_j \cdot n_{ij} \right)^2$$

Moreover, mixed GCM models are defined by adding, usually, one or more molecular descriptors to the group contributions:

$$y_i = k_0 + \sum_{j=1}^{m} \mathrm{G}_j \cdot n_{ij} + \sum_{j'=1}^{p} \mathcal{D}_{ij'}$$

where the second summation runs over the p molecular descriptors defined in the GCM scheme and $\mathcal{D}_{ij'}$ is the j'th molecular descriptor value of the ith molecule.

The group contributions G are usually estimated by multivariate regression analysis, but they can also be experimental, theoretical, or user-defined quantities. For example, in the latter case,

the molecular weight can be viewed as a simple linear atom contribution model, where the group contributions are atomic masses. In the first case, large training sets are used to obtain reliable estimates of the group contributions. Usually a battery of group contributions (a field of scalar parameters) is defined taking into account several structural characteristics of the molecules, also sometimes adding extra terms (correction factors) referring to special substructures. If correction factors are considered, the GCM models are usually called **additive–constitutive models**.

Group contribution models were proposed for several molecular property estimations [Zhao, Abraham *et al.*, 2003b], such as boiling and melting points [Wang, Milne *et al.*, 1994; Krzyzaniak, Myrdal *et al.*, 1995; Le and Weers, 1995], → *molar refractivity* [Huggins, 1956; Ghose and Crippen, 1987], pK_a [Perrin, Dempsey *et al.*, 1981; Hilal, Karichoff *et al.*, 1995], critical temperatures, solubilities [Hine and Mookerjee, 1975; Klopman, Wang *et al.*, 1992; Myrdal, Ward *et al.*, 1993; Thomsen, Rasmussen *et al.*, 1999], soil sorption coefficients [Tao, Piao *et al.*, 1999], and several thermodynamic properties [Thinh and Trong, 1976; Yoneda, 1979; Reid, Prausnitz *et al.*, 1988; Suzuki, 2001; Béliveau, Tardif *et al.*, 2003; Béliveau, Lipscomb *et al.*, 2005]. The Rekker method [Nys and Rekker, 1973; Rekker, 1977a] is an example of group contribution method applied to the estimation of → *log P*. Another well-known group contribution model is that proposed by Atkinson for the evaluation of reaction rate constants with hydroxyl radicals of organic compounds [Atkinson, 1987, 1988].

🕮 [Smolenskii, 1964; Essam, Kennedy *et al.*, 1977; Ghose and Crippen, 1986; Klein, 1986; Elbro, Fredeslund *et al.*, 1991; Gao, Govind *et al.*, 1992; Drefahl and Reinhard, 1993; Bhattacharjee, 1994; Klopman, Li *et al.*, 1994; Yalkowsky, Dannenfelser *et al.*, 1994; Yalkowsky, Myrdal *et al.*, 1994; Meylan and Howard, 1995; Klein, Schmalz *et al.*, 1999; Platts, Butina *et al.*, 1999; Viswanadhan, Ghose *et al.*, 1999; Wildman and Crippen, 1999; Platts, Abraham *et al.*, 2000]

- **group contributions** → group contribution methods
- **group electronegativity** → atomic electronegativity
- **group molar refractivity** → physico-chemical properties (⊙ molar refractivity)
- **GTI** ≡ *Estrada Generalized Topological Indices* → variable descriptors
- **GUS index** → environmental indices (⊙ leaching indices)
- **Gutman index** ≡ *first Zagreb index* → Zagreb indices
- **Gutman molecular topological index** → Schultz molecular topological index
- **Gutmann's acceptor number** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **Gutmann's donor number** → Linear Solvation Energy Relationships (⊙ hydrogen-bond parameters)
- **GVW drug-like indices** → property filters (⊙ drug-like indices)
- **G-WHIM descriptors** → grid-based QSAR techniques