A

- **A_{xi}** eigenvalue indices → spectral indices
- **Abraham–Klamt descriptors** → Linear Solvation Energy Relationships
- **Abraham's general equation** → Linear Solvation Energy Relationships
- **absolute hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **absorption parameter** → property filters (⊙ drug-like indices)
- **Acceptable Daily Intake** → biological activity indices (⊙ toxicological indices)
- **acceptor superdelocalizability** ≡ *electrophilic superdelocalizability* → quantum-chemical descriptors
- **ACCS** ≡ *Activity Class Characteristic Substructures* → substructure descriptors (⊙ structural keys)
- **ACC transforms** ≡ *Auto-Cross-Covariance transforms* → autocorrelation descriptors
- **ACD/log *P*** → lipophilicity descriptors
- ***ACGD* index** → charged partial surface area descriptors
- **acid dissociation constant** → physico-chemical properties (⊙ equilibrium constants)
- **activation energy index** → quantum-chemical descriptors (⊙ highest occupied molecular orbital energy)
- **activation hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **Activity Class Characteristic Substructures** → substructure descriptors (⊙ structural keys)
- **acyclic graph** ≡ *tree* → graph
- **acyclic polynomial** ≡ *matching polynomial* → Hosoya Z-index

ADAPT descriptors

ADAPT descriptors [Jurs, Chou *et al.*, 1979; Jurs, Hasan *et al.*, 1988], implemented in the homonymous software ADAPT (Automated Data Analysis and Pattern Recognition Toolkit), fall into three general categories: → *topological indices*, → *geometrical descriptors* (including → *principal moments of inertia*, → *volume descriptors*, and → *shadow indices*), and → *electronic descriptors* (including partial atomic charges and the → *dipole moment*); moreover, → *molecular weight*, → *count descriptors*, and a large number of → *substructure descriptors* are also generated. In addition, the → *charged partial surface area descriptors* constitute a fourth class of descriptors derived by combining electronic and geometrical information.

ADAPT software allows (a) molecular descriptor generation; (b) objective feature selection to discard descriptors that contain redundant or minimal information; and (c) multiple regression

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

ISBN: 978-3-527-31852-0

analysis by genetic algorithm or simulated annealing variable selection, or computational → *artificial neural networks*.

Several molecular properties have been modeled by ADAPT descriptors, such as *biological activities* [Henry, Jurs *et al.*, 1982; Jurs, Hasan *et al.*, 1983; Jurs, Stouch *et al.*, 1985; Walsh and Claxton, 1987; Wessel, Jurs *et al.*, 1998; Eldred, Weikel *et al.*, 1999; Eldred and Jurs, 1999; Patankar and Jurs, 2000, 2002; He, Jurs *et al.*, 2003, He *et al.*,2005; Benigni, 2005]; *boiling point* [Smeeks and Jurs, 1990;Stanton, Jurs *et al.*, 1991; Stanton, Egolf *et al.*, 1992; Egolf and Jurs, 1993a; Egolf, Wessel *et al.*, 1994; Wessel and Jurs, 1995a, 1995b; Goll and Jurs, 1999a; Stanton, 2000]; *chromatographic indices* [Anker, Jurs *et al.*, 1990; Sutter, Peterson *et al.*, 1997]; *aqueous solubilities* [Dunnivant, Elzerman *et al.*, 1992; Nelson and Jurs, 1994; Sutter and Jurs, 1996; Mitchell and Jurs, 1998a], *critical temperature and pressure* [Turner, Costello *et al.*, 1998]; *ion mobility constants* [Wessel and Jurs, 1994;Wessel, Sutter *et al.*, 1996]; *reaction rate constants* [Bakken and Jurs,1999a, 1999b]; and other various properties [Egolf and Jurs, 1992;Russell, Dixon *et al.*, 1992; Stanton and Jurs, 1992; Egolf and Jurs, 1993b; Engelhardt and Jurs, 1997; Mitchell and Jurs, 1997; Goll and Jurs, 1999b; Johnson and Jurs, 1999; Kauffman and Jurs, 2000, 2001a; Mattioni and Jurs, 2003].

- **additivity model** ≡ *Free–Wilson model* → Free–Wilson analysis
- **additive adjacency matrix** → adjacency matrix
- **additive chemical adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **additive–constitutive models** → group contribution methods
- **additive model of inductive effect** → electronic substituent constants (⊙ inductive electronic constants)
- **ADI** ≡ *Acceptable Daily Intake* → biological activity indices (⊙ toxicological indices)
- **adjacencies** → graph

■ **adjacency matrix (A)** (≡ *vertex adjacency matrix*)

The adjacency matrix **A** is one of the fundamental → *graph theoretical matrices*; it represents the whole set of connections between adjacent pairs of atoms [Trinajstić, 1992]. The entries a_{ij} of the matrix equal 1 if vertices v_i and v_j are adjacent (i.e., the atoms i and j are bonded) and zero otherwise:

$$[\mathbf{A}]_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E}(\mathcal{G}) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{E}(\mathcal{G})$ is the set of the graph edges.

This is the classical definition of the adjacency matrix, which refers to a → *simple graph*, where multiple bonds are not accounted for. The adjacency matrix is symmetric with dimension $A \times A$, where A is the number of atoms and it is usually derived from a → *H-depleted molecular graph*.

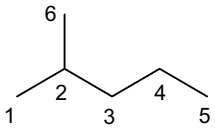
The ith row sum of the adjacency matrix is called → *vertex degree*, denoted by δ_i and defined as

$$\delta_i \equiv VS_i(\mathbf{A}) = \sum_{j=1}^{A} a_{ij}$$

where *VS* is the → *vertex sum operator*. The vertex degree represents the number of σ bonds of the ith atom.

Example A1

Adjacency matrix **A** and vertex degrees δ_i of 2-methylpentane.



A =

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

The **total adjacency index** A_V is a measure of the graph connectedness and is calculated as the sum of all the entries of the adjacency matrix of a molecular graph, which is twice the number B of graph edges [Harary, 1969a; Rouvray, 1983]:

$$A_V = \sum_{i=1}^{A}\sum_{j=1}^{A} a_{ij} = \sum_{i=1}^{A} \delta_i = 2 \cdot B$$

For example, the total adjacency index of 2-methylpentane is $A_V = 1 + 3 + 2 + 2 + 1 + 1 = 10$, which is twice the number of edges equal to five in the H-depleted molecular graph of this molecule. Therefore, the number of entries equal to 1 in the adjacency matrix is $2B$, while the number of entries equal to zero is $A^2 - 2B$; in particular, for acyclic graphs the total number of entries equal to 1 is $2(A-1)$ and the number of entries equal to zero is $A^2 - 2(A-1)$; for monocyclic graphs, the values are $2A$ and $A^2 - 2A$, respectively. The total adjacency index is sometimes calculated as the half sum of the adjacency matrix elements.

The global connectivity of a graph can also be characterized by the average of the total adjacency index as [Bonchev and Buck, 2007]

$$\bar{\delta} = \frac{A_V}{A} = \frac{2B}{A}$$

where A is the number of graph vertices. If calculated from the → *H-filled molecular graph*, this average index is one of the two → *Schäfli indices*, called connectivity.

The doubly normalized total adjacency index is called **density index** and is defined as

$$\bar{\bar{\delta}} = \frac{A_V}{A \cdot (A-1)} = \frac{2B}{A \cdot (A-1)} \quad \text{or} \quad \bar{\bar{\delta}} = \frac{A_V}{A^2} = \frac{2B}{A^2}$$

The adjacency matrix is one important source of molecular descriptors. Simple → *topological information indices* can be calculated on both the equality and the magnitude of adjacency matrix elements. → *Walk counts* and → *self-returning walk counts* that coincide with the spectral moments of the adjacency matrix are calculated by the increasing powers of the adjacency matrix [McKay, 1977; Jiang, Tang *et al.*, 1984; Hall, 1986; Kiang and Tang, 1986; Jiang and Zhang, 1989, 1990; Marković and Gutman, 1991; Jiang, Qian *et al.*, 1995; Marković and Stajkovic, 1997; Marković, 1999].

→ *Spectral indices*, → *determinant-based descriptors*, and → *characteristic polynomial-based descriptors* of the adjacency matrix are largely used in QSAR modeling.

The **clustering coefficient of a vertex**, denoted as C_i, is a local vertex invariant derived from the adjacency matrix by considering the first-neighbor interconnectivity [Bonchev and Buck, 2007]. It was proposed as a measure of the clustered structure of a graph around a vertex and is defined as

$$C_i = \frac{2 \cdot b_i}{\delta_i \cdot (\delta_i - 1)} \quad 0 \leq C_i \leq 1$$

where b_i is number of edges between the first neighbors of the ith vertex, measuring to what extent the first neighbors of the ith vertex are linked between themselves:

$$b_i = \frac{1}{2} \cdot \sum_{j=1}^{A} a_{ij} \cdot \sum_{m=1}^{A} a_{jm} \cdot a_{mi} \quad m \neq i$$

where A is the number of vertices, and a_{ij}, a_{jm}, and a_{mi} are the elements of the adjacency matrix. Then, the terms a_{ij} of the first summation are equal to 1 only for the vertices v_j, which are connected to the ith vertex, while terms $a_{jm} \cdot a_{mi}$ in the second summation are equal to 1 for those pairs of vertices v_j and v_m, which are contemporarily neighbors of the ith vertex and are bonded to each other, and are zero otherwise.

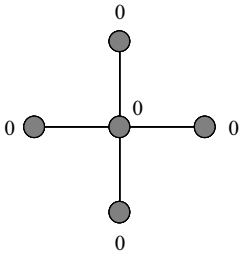
The **overall degree of clustering of a graph** is given by [Bonchev and Buck, 2007]

$$\bar{C} = \frac{1}{A} \cdot \sum_{i=1}^{A} C_i$$

It should be noted that clustering around a vertex is possible only in trimembered cycles; in all other structures, there are no edges between the first-neighbor vertices, for example, $b_i = 0$.

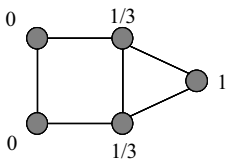
Example A2

Calculation of the density index $\bar{\bar{\delta}}$ and overall degree of clustering $\bar{C}$ for graphs A, B, and C. Each vertex is labeled with its clustering coefficient.



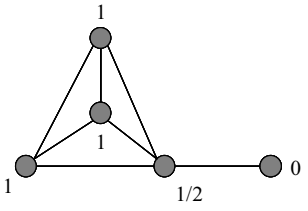
A

$A = 5 \quad B = 4$
$\bar{\bar{\delta}} = 0.4$
$\bar{C} = 0$



B

$A = 5 \quad B = 6$
$\bar{\bar{\delta}} = 0.6$
$\bar{C} = 0.333$



C

$A = 5 \quad B = 7$
$\bar{\bar{\delta}} = 0.7$
$\bar{C} = 0.7$

To account for multiple bonds, → *atom connectivity matrix*, → *adjacency matrix of a multigraph*, and → *adjacency matrix of a general graph* can be used instead of the adjacency matrix of a simple graph. To account for heteroatoms, different → *weighted adjacency matrices* were proposed, such as the → *augmented adjacency matrix* and → *chemical adjacency matrix*.

The **additive adjacency matrix** is derived from the adjacency matrix substituting row elements equal to 1, corresponding to pairs of adjacent vertices, with the vertex degrees of the connected vertices as

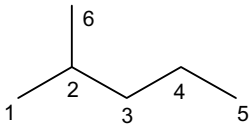
$$[{}^{\delta}\mathbf{A}]_{ij} = \begin{cases} \delta_j & \text{if } (i,j) \in \mathcal{E}(G) \\ 0 & \text{otherwise} \end{cases}$$

where δ_j is the vertex degree of the jth vertex connected to the ith vertex.

This matrix is a special case of → *distance degree matrices* obtained by the parameter combination $\alpha = 0$, $\beta = 0$, $\gamma = 1$. The row sum of the additive adjacency matrix is the → *extended connectivity* of first-order EC^1 defined by Morgan. This local invariant was used to calculate the → *eccentric adjacency index*. A modification of this matrix, which accounts for heteroatoms, is the → *additive chemical adjacency matrix*.

Example A3

Additive adjacency matrix ${}^{\delta}\mathbf{A}$ and extended connectivities EC^1 of 2-methylpentane.



${}^{\delta}\mathbf{A} =$

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

Other topological matrices are derived from the adjacency matrix, such as → *Laplacian matrix* and the powers of the adjacency matrix used to obtain walk counts and the corresponding molecular descriptors.

The **fragmental adjacency matrix** ${}^{m}\mathbf{A}_{\mathbf{F}}$ of mth order is a generalization of the adjacency matrix **A**, which encodes information about adjacencies of the K fragments of the same mth order (i.e., the same number m of edges) contained in the molecular graph instead of adjacencies between vertices [Guevara, 1999]. This matrix is a square symmetric ($K \times K$) matrix whose elements are different from zero only if two fragments i and j are adjacent, that is, they have $m - 1$ edges in common. The **fragmental degree** is defined in the same way as the vertex degree, that is, the row sum of the fragmental adjacency matrix, and, therefore, represents the number of fragments adjacent to the fragment considered. Then, by using the fragmental degree in place of the vertex degree, the **fragmental connectivity index** can be calculated in the same way as the → *Randić connectivity index*.

Moreover, the adjacency matrix can be transformed into a **decimal adjacency vector**, denoted by $\mathbf{a}^{\mathbf{10}}$, of A elements each being a local vertex invariant obtained by the following expression [Schultz and Schultz, 1991]:

$$a_i^{10} = (2 \cdot a_{i1})^{A-1} + (2 \cdot a_{i2})^{A-2} + \cdots + (2 \cdot a_{iA})^{0}$$

where a_{ij} is the jth column element of the ith row of the adjacency matrix **A**. In this way, the information contained in the adjacency matrix is compressed into an A-dimensional vector. For

example, a row of the adjacency matrix equal to [0 1 1 1 0] gives a value of 14, obtained as

$$a_i^{10} = (2 \cdot 0)^{5-1} + (2 \cdot 1)^{5-2} + (2 \cdot 1)^{5-3} + (2 \cdot 1)^{5-4} + (2 \cdot 0)^0 = 14$$

The elements of the decimal adjacency vector are integers that were used for → *canonical numbering* of molecular graphs [Randić, 1974].

From the decimal adjacency vector, three different indices were proposed as molecular descriptors:

(a) the sum of the elements of the vector $\mathbf{a^{10}}$, that is,

$$A1 = \sum_{i=1}^{A} a_i^{10}$$

(b) the sum of the linear combination of vertex degrees δ_i, each weighted by the corresponding decimal adjacency vector element a_i^{10}, that is,

$$A2 = \sum_{i=1}^{A} \delta_i \cdot a_i^{10}$$

(c) the sum of the elements of the *A*-dimensional vector **d** obtained by multiplying the topological → *distance matrix* **D** by the decimal adjacency vector, that is,

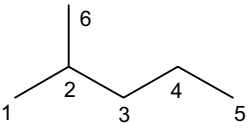
$$A3 = \sum_{i=1}^{A} [\mathbf{d}]_i$$

where the vector **d** is calculated as

$$\mathbf{d} = \mathbf{D} \cdot \mathbf{a}^{10}$$

Example A4

Decimal adjacency vector of 2-methylpentane and related molecular descriptors.



Atom	1	2	3	4	5	6		a_i^{10}		d_i
1	0	1	2	3	4	2		16		159
2	1	0	1	2	3	1		41		84
3	2	1	0	1	2	2	×	20	=	123
4	3	2	1	0	1	3		10		202
5	4	3	2	1	0	4		4		301
6	2	1	2	3	4	0		16		159

$a_1^{10} = (2 \times 1)^{6-2} = 16$
$a_2^{10} = (2 \times 1)^{6-1} + (2 \times 1)^{6-3} + (2 \times 1)^{6-6} = 41$
$a_3^{10} = (2 \times 1)^{6-2} + (2 \times 1)^{6-4} = 20$
$a_4^{10} = (2 \times 1)^{6-3} + (2 \times 1)^{6-5} = 10$
$a_5^{10} = (2 \times 1)^{6-4} = 4$
$a_6^{10} = (2 \times 1)^{6-2} = 16$

$A1 = 16 + 411 + 20 + 10 + 4 + 16 = 107$
$A2 = 1 \times 16 + 3 \times 41 + 2 \times 20 + 2 \times 10 + 1 \times 4 + 1 \times 16 = 219$
$A3 = 159 + 84 + 123 + 202 + 301 + 159 = 1028$

📖 [Lukovits, 2000; Schultz, 2000]

- **adjacency matrix of a general graph** → weighted matrices (⊙ weighted adjacency matrices)
- **adjacency matrix of a multigraph** → weighted matrices (⊙ weighted adjacency matrices)
- **adjacency plus distance matrix** → Schultz molecular topological index
- **adjacent eccentric distance sum index** → eccentricity-based Madan indices (⊙ Table E1)
- **adjusted R^2** → regression parameters
- **adjusted retention time** → chromatographic descriptors (⊙ retention time)
- **admittance matrix** ≡ *Laplacian matrix*
- **ADME properties** → drug design

■ **adsorbability index** (AI)

An empirical molecular descriptor derived from a → *group contribution method* based on molecular refractivity to predict activated carbon adsorption of 157 compounds [Abe, Tatsumoto *et al.*, 1986]. This index was also applied to predict the → *soil sorption partition coefficient* of the same 157 compounds [Okouchi and Saegusa, 1989; Okouchi, Saegusa *et al.*, 1992].

The adsorbability index is calculated by the expression

$$\mathrm{AI} = \sum\nolimits_i f_i \cdot \mathrm{N}_i + \sum\nolimits_j c_j$$

where the summations run over atomic and functional groups; f_i indicates the contribution to activated carbon adsorption of the ith atom- or group type and N_i the number of atoms or groups of type i; c_j represents a special correction factor accounting for functional group effects.

Atomic and group contributions and correction factors are reported in Table A1.

Table A1 Values of f and c factors proposed by Abe, *et al.* [Abe, Tatsumoto *et al.*, 1986].

Atom/group	f	Group	c
C	0.26	Aliphatic:	
H	0.12	−OH (alcohols)	−0.53
N	0.26	−O− (esters)	−0.36
O	0.17	−CHO (aldehydes)	−0.25
S	0.54	N (amines)	−0.58
Cl	0.59	−COOR (esters)	−0.28
Br	0.86	>C=O (ketones)	−0.30
NO_2	0.21	−COOH (fatty acids)	−0.03
−C=C−	0.19		
Iso	−0.12	α-Amino acids	−1.55
Tert	−0.32		
Cyclo	−0.28	All groups in aromatics	0

For example, for benzene
$\mathrm{AI} = 6 \times f_{\mathrm{C}} + 6 \times f_{\mathrm{H}} + 3 \times f_{\mathrm{C=C}} = 6 \times 0.26 + 6 \times 0.12 + 3 \times 0.19 = 2.85$;
for 1,1,2-tricholoroethane $\mathrm{AI} = 2 \times f_{\mathrm{C}} + 3 \times f_{\mathrm{H}} + 3 \times f_{\mathrm{Cl}} = 2 \times 0.26 + 3 \times 0.12 + 3 \times 0.59 = 2.65$

- ***AEI* indices** → spectral indices (⊙ A_{xi} eigenvalue indices)
- **AFC method** ≡ *KOWWIN* → lipophilicity descriptors

affinity fingerprints

Affinity fingerprints are → *vectorial descriptors* of molecules either comprising their binding affinities and docking scores or superpositioning pseudoenergies against a reference panel of uncorrelated proteins or small drug molecules [Briem and Lessel, 2000]. These molecular descriptors can be used both for high-throughput compound screening and the → *similarity/diversity* analysis and for the prediction of biological activities of compounds.

In contrast to most other molecular descriptors, affinity fingerprints are not directly derived from molecular structures.

***In vitro* affinity fingerprints** are based on binding affinities, experimentally determined, and can be used to estimate general cross-reactivity and, then, possible toxicity in the drug design process [Weinstein, Kohn *et al.*, 1992; Kauvar, Higgins *et al.*, 1995; Weinstein, Myers *et al.*, 1997; Dixon and Villar, 1998]. The underlying assumption is that compounds binding similarly to all the proteins in the reference panel are likely also to have similar affinity to their target receptor.

***Virtual* affinity fingerprints** (or ***in silico* affinity fingerprints**) are derived by computational methods and, thus, are vectorial descriptors where experimentally determined binding affinities of molecules are replaced by some calculated scores with respect to the reference panel [Briem and Lessel, 2000].

Some *virtual* affinity fingerprints are explained below.

DOCKSIM fingerprints are derived by computational docking of the molecules into binding pockets of protein structures solved by X-ray crystallography [Briem and Kuntz, 1996]. Therefore, they are 3D vectorial descriptors collecting the docking scores (**DOCK scores**) with respect to the protein-binding site of the reference panel; the scores are obtained by rigid docking of the molecules, and the reference panel contains eight uncorrelated and arbitrarily selected protein structures.

Flexsim-X fingerprints are vectors of docking scores as the DOCKSIM fingerprints, but the scores are obtained by flexible docking of molecules, and the reference panel contains around 40 protein structures, optimized by systematic and genetic algorithm (GA)-based procedures [Lessel and Briem, 2000].

Flexsim-S fingerprints are *virtual* affinity fingerprints where the docking scores are replaced by the superpositioning pseudoenergies, which measure the alignment quality of ligands onto a set of small reference molecules [Lemmen, Lengauer *et al.*, 1998]. Also for Flexsim-S fingerprints, size and composition of the reference panel should be properly optimized.

Flexsim-R fingerprints are *virtual* affinity fingerprints specifically designed for similarity assessments of small fragments, such as R-groups of combinatorial libraries [Weber, Teckentrup *et al.*, 2002].

Molecular hashkeys are another kind of *virtual* affinity fingerprints derived from surface-based comparisons of ligands with a reference panel comprising small, drug-like molecules instead of proteins [Ghuloum, Sage *et al.*, 1999]. A molecular hashkey is a vectorial descriptor of fixed dimension that captures information about the surface properties of a molecule. The elements of the hashkey of a molecule are values of its molecular surface similarity to a set of basis molecules in low-energy-fixed conformations. The molecule is flexibly aligned to each of the set of basis molecules to maximize molecular surface similarity.

- ***AH* weighting scheme** → weighting schemes
- **Aihara resonance energy** → delocalization degree indices
- **AI indices** → atom-type-based topological indices

■ **AIM theory** (≡ *Atoms-in-Molecules theory*)
Bader's Atoms-in-Molecules (AIM) quantum theory [Bader, 1990] provides a bridge between quantum chemistry and chemical concepts and the framework for reconstructing large molecules from a number of small electron density fragments. In the AIM theory, the electron density of a molecule is partitioned into distinct electron density basins, that is, regions occupied by the corresponding atoms, each containing an atomic nucleus. These electron density atomic fragments are essentially bounded by surfaces of zero net flux in the electron density.

An atomic property P can be then expressed as the integral of the corresponding property density ρ over an atomic region Ω as

$$P(\Omega) = \int_{\Omega} \rho_P d\tau$$

These atomic properties possess a high degree of transferability from the electronic environment in one molecule to another molecule with similar environments. Consequently, the properties of a whole molecule or a functional group can be obtained by adding the atomic properties as

$$P(\text{molecule}) = \sum_{\Omega} P(\Omega).$$

Based on the AIM theory are → *TAE descriptors* and the → *delocalization index DI*.

🕮 [Song, Breneman *et al.*, 2002; Lamarche and Platts, 2003; Chaudry and Popelier, 2004; Krygowski, Ejsmont *et al.*, 2004]

➢ **Akaike Information Criterion** → regression parameters
➢ **alert indices** → property filters

■ **algebraic operators**
Algebraic operators play a meaningful role in the framework of → *molecular descriptors*, since they represent the fundamental mathematical tool used to transform into single numerical quantities the information encoded in → *matrices of molecules*.

Let **M** be a generic matrix with n rows and p columns, denoted as

$$\mathbf{M} \equiv [m_{ij}] = \begin{vmatrix} m_{11} & m_{12} & \dots & \dots & m_{1p} \\ \vdots & & & & \vdots \\ m_{n1} & m_{n2} & \dots & \dots & m_{np} \end{vmatrix}$$

The matrix elements m_{ij} are commonly denoted as

$$m_{ij} \equiv [\mathbf{M}]_{ij} \equiv (i,j)$$

A column vector **v** is a special case of matrix having n rows and one column; the row vector $\mathbf{v}^{\mathbf{T}}$ is a special case of matrix having one row and p columns.

Some definitions of matrix algebra [Ledermann and Vajda, 1980; Golub and van Loan, 1983; Mardia, Kent *et al.*, 1988], algebraic operators and set theory are given below.

• **characteristic polynomial**
Let **M** be a square matrix ($n \times n$) and x a scalar variable, the characteristic polynomial *Ch* is defined as

$$Ch(\mathbf{M}, x) = \det(\mathbf{M} - x\mathbf{I}) = \sum_{i=0}^{n} a_i \cdot x^{n-i}$$

where **I** is the identity matrix, that is, a matrix having the diagonal elements equal to 1 and all the off-diagonal elements equal to zero, and a_i the polynomial coefficients. The characteristic polynomial is obtained by expanding the determinant and, then, collecting terms with equal powers of x.

The **eigenvalues** λ of the matrix **M** are the n roots of its characteristic polynomial, and the set of the eigenvalues is called **spectrum of a matrix**, denoted as $\mathbf{\Lambda(M)}$.

Determinant and trace of **M** are given by the following expressions:

$$\det(\mathbf{M}) = a_n = \prod_{i=1}^{n} \lambda_i \quad tr(\mathbf{M}) = a_1 = \sum_{i=1}^{n} \lambda_i$$

respectively, where a_n and a_1 are the characteristic polynomial coefficients corresponding to i equal to n and 1, respectively.

For each eigenvalue λ_i, there exists a nonzero vector $\mathbf{v}_i$ satisfying the following relationship:

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

The n-dimensional vectors $\mathbf{v}_i$ are called **eigenvectors** of **M**.

A large number of → *characteristic polynomial-based descriptors* and → *spectral indices* are defined in literature, to study both molecular graphs and model physico-chemical properties of molecules.

• **cardinality of a set**
The cardinality of a set S is the number of elements in S and is indicated as $|S|$.

• **column sum operator**
This operator, denoted as CS_j, performs the sum of the elements of the jth matrix column:

$$CS_j(\mathbf{M}) \equiv \sum_{i=1}^{n} m_{ij}$$

The **column sum vector**, denoted by **cs**, is a p-dimensional vector collecting the results obtained by applying the column sum operator to all the p columns of the matrix.

• **determinant**
The determinant of a $n \times n$ square matrix **M**, denoted by det(**M**), is a scalar quantity and is defined as

$$\det(\mathbf{M}) = \sum\nolimits_{\pi} s(\pi) \cdot m_{1,\,i_1} \cdot m_{2,\,i_2} \cdot \ldots \cdot m_{n,\,i_n}$$

where the summation ranges over all $n!$ permutations π of the symbols 1, 2, …, n. Each permutation π of degree n is given by

$$\pi = \begin{pmatrix} 1 & 2 & \ldots & n \\ i_1 & i_2 & \ldots & i_n \end{pmatrix}$$

where $i_1, i_2, \ldots, i_n$ are the symbols 1, 2, ..., n in some order. The sign function $s(\pi)$ is defined as

$$s(\pi) = \begin{cases} +1 & \text{if } \pi \text{ is even} \\ -1 & \text{if } \pi \text{ is odd} \end{cases}$$

Related to the definition of determinant are permanent, pfaffian, and hafnian.

The **permanent**, denoted by per(**M**), also referred to as the positive determinant, is defined by omitting the sign function $s(\pi)$ [Kasum, Trinajstić *et al.*, 1981; Schultz, Schultz *et al.*, 1992, 1995; Cash, 1995a, 1998; Jiang, Liang *et al.*, 2006] as

$$\mathrm{per}(\mathbf{M}) = \sum\nolimits_{\pi} m_{1,\, i_1} \cdot m_{2,\, i_2} \cdot \ldots \cdot m_{n,\, i_n}$$

where π runs over the $n!$ permutations.

From the permanent, the corresponding permanent polynomial was also defined [Kasum, Trinajstić *et al.*, 1981; Cash, 2000b].

The **immanant**, denoted by $d_\lambda(\mathbf{M})$, is defined as

$$d_\lambda(\mathbf{M}) = \sum\nolimits_{\pi} \chi_\lambda(\pi) m_{1,\, i_1} \cdot m_{2,\, i_2} \cdot \ldots \cdot m_{n,\, i_n}$$

where π runs over the $n!$ permutations. $\chi_\lambda(\pi)$ is an irreducible character of the symmetric group indexed by a partition λ of n.

The **pfaffian**, denoted by pfa(**M**), and the **hafnian**, denoted by haf(**M**), are analogous to the determinant except for the summation that goes over all the permutations π ($i_1, i_2, \ldots, i_n$) and must also satisfy the limitations

$$i_1 < i_2, i_3 < i_4, \ \ldots, i_{n-1} < i_n; \qquad i_1 < i_3 < i_5 < \ldots < i_{n-1}$$

The entries of the main diagonal are excluded from the calculation of the pfaffian and hafnian [Caianiello, 1953, 1956]. Hafnians and pfaffians differ in the sign function $s(\pi)$ that is included in the definition of pfaffian only.

The hafnian calculated considering only the entries above the main diagonal is called **short hafnian**, shaf(**M**), whereas the hafnian calculated considering both entries above and below the main diagonal can also be referred to as **long hafnian**, lhaf(**M**) [Schultz and Schultz, 1992; Schultz, Schultz *et al.*, 1995].

For example, for a matrix **M** of order 4, pfaffian, long hafnian, and short hafnian are the following:

$$\begin{aligned} \mathrm{pfa} &= m_{12} \cdot m_{34} - m_{13} \cdot m_{24} + m_{14} \cdot m_{23} \\ \mathrm{shaf} &= m_{12} \cdot m_{34} + m_{13} \cdot m_{24} + m_{14} \cdot m_{23} \\ \mathrm{lhaf} &= m_{12} \cdot m_{21} \cdot m_{34} \cdot m_{43} + m_{13} \cdot m_{31} \cdot m_{24} \cdot m_{24} + m_{14} \cdot m_{41} \cdot m_{23} \cdot m_{32} \end{aligned}$$

Some molecular descriptors, called → *determinant-based descriptors*, are calculated as the determinant of → *matrices of molecules*. Moreover, permanents, short and long hafnians, calculated on the topological → *distance matrix* **D**, were used as graph invariants by Schultz and called **per(D) index**, **shaf(D) index**, and **lhaf(D) index** [Schultz and Schultz, 1992; Schultz, Schultz *et al.*, 1992].

🕮 [Schultz and Schultz, 1993; Schultz, Schultz *et al.*, 1993, 1994, 1995, 1996; Chan, Lam *et al.*, 1997; Gutman, 1998; Cash, 2000a, 2002a, 2003]

• **diagonal matrix**
A diagonal matrix is a square matrix whose diagonal terms m_{ii} are the only nonzero elements. The **diagonal operator** $\mathcal{D}(\mathbf{M})$ is an operator that transforms a generic square matrix **M** into a diagonal matrix:

$$\mathcal{D}(\mathbf{M}) = \begin{vmatrix} m_{11} & \dots & 0 & \dots & 0 \\ 0 & \dots & m_{ii} & \dots & 0 \\ 0 & \dots & 0 & \dots & m_{nn} \end{vmatrix}$$

• **Hadamard matrix product**
The Hadamard product of two matrices **A** and **B** of the same dimension is denoted as $\otimes$ and defined as

$$[\mathbf{A} \otimes \mathbf{B}]_{ij} = [\mathbf{A}]_{ij} \times [\mathbf{B}]_{ij}$$

that is, the elements of the resulting matrix are obtained by the scalar product of the corresponding elements of **A** and **B** matrices.

• **identity matrix (I)**
The identity matrix is a square diagonal matrix defined as

$$\mathbf{I} = \begin{vmatrix} 1 & \dots & 0 & \dots & 0 \\ 0 & \dots & 1 & \dots & 0 \\ 0 & \dots & 0 & \dots & 1 \end{vmatrix}$$

• **polynomial**
A polynomial $\mathcal{P}(x)$ of the x variable is a linear combination of its powers, usually written as

$$\mathcal{P}(x) = a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n$$

where n is the order of the polynomial. The values of x, for which $\mathcal{P}(x)$ is zero, are the *roots* of the polynomial.

Several polynomials associated with graphs were defined, such as the → *characteristic polynomials*, → *counting polynomials*, → *matching polynomial*, chromatic polynomial, and Tutte polynomial [Noy, 2003].

• **product of matrices**
Let **A** (n, m) and **B** (m, p) be two matrices. The product of the two matrices is defined as

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{C}$$

where the resulting product matrix **C** has n rows and p columns. Each scalar element c_{ij} of the **C** matrix is obtained by the scalar product between the ith row of the **A** matrix and the jth column of the **B** matrix. The row vector is represented as $\mathbf{a}_i^{\mathrm{T}}$ and the column vector as $\mathbf{b}_j$; the resulting element c_{ij} is then calculated as

$$\mathbf{a}_i^{\mathbf{T}} \cdot \mathbf{b}_j \equiv c_{ij} = \sum_{k=1}^{m} a_{ik} \cdot b_{kj}$$

A basic condition for the product of two matrices is that the number of columns of the left matrix and the number of rows of the right matrix are equal (m).

The kth **power matrix** $\mathbf{A}^k$ is a special case of the matrix product:

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

The main properties of the product of two matrices are

$$\text{(a) } \mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}, \qquad \text{(b) } (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}), \qquad \text{(c) } (\mathbf{A} \cdot \mathbf{B})^{\mathbf{T}} = \mathbf{B}^{\mathbf{T}} \cdot \mathbf{A}^{\mathbf{T}}$$

• **row sum operator** (≡ *vertex sum operator*)
This operator, denoted as VS_i, performs the sum of the elements of the ith matrix row:

$$VS_i(\mathbf{M}) = \sum_{j=1}^{p} m_{ij}$$

where p is the number of columns of the **M** matrix.

The **row sum vector**, **rs**, is an n-dimensional vector collecting the results obtained by applying the row sum operator to all the n rows of the matrix.

This operator is used to derive → *Local Vertex Invariants* from → *graph theoretical matrices*. For symmetric matrices, the local vertex invariants obtained by applying this operator on the transposed matrix coincide with those obtained by applying the operator on the original matrix.

• **scalar product of vectors**
Let **a** and **b** be two column vectors with the same dimension n. The scalar product between the two vectors is defined as the sum of the products of the corresponding elements of the row vector $\mathbf{a}^{\mathbf{T}}$ and the column vector **b** or, vice versa, of the row vector $\mathbf{b}^{\mathbf{T}}$ and the column vector **a**:

$$\mathbf{a}^{\mathbf{T}} \cdot \mathbf{b} = \mathbf{b}^{\mathbf{T}} \cdot \mathbf{a} = \sum_{k=1}^{n} a_k \cdot b_k$$

• **sparse matrices**
These are matrices with relatively few nonzero elements. A **binary sparse matrix B** is a sparse matrix comprised of elements equal to zero or 1. The **geodesic matrix** is a binary sparse matrix ${}^m\mathbf{B}$ defined as [Harary, 1969a]

$$[{}^m\mathbf{B}]_{ij} = \begin{cases} 1 & \text{if } d_{ij} = m \\ 0 & \text{otherwise} \end{cases}$$

where m defines the order of the matrix and d_{ij} is the → *topological distance* between vertices v_i and v_j. The geodesic matrix is largely applied to calculate → *autocorrelation descriptors*, → *Estrada generalized topological indices*, → *higher order Wiener numbers*, → *interaction geodesic matrices*.

Let **M** be a matrix $A \times A$ representing a → *molecular graph* $\mathcal{G}$, where A is the number of vertices. To obtain a ***m*th order sparse matrix** ${}^{m}\mathbf{M}$ from any matrix **M**, the → *Hadamard matrix product* is performed as the following:

$$ {}^{m}\mathbf{M} = \mathbf{M} \otimes {}^{m}\mathbf{B} $$

where the superscript "*m*" of the sparse matrix ${}^{m}\mathbf{M}$ means that all of the **M** matrix elements are taken as zero but those corresponding to pairs of vertices v_i and v_j at topological distance m and ${}^{m}\mathbf{B}$ constitute the geodesic matrix defined above.

The → *adjacency matrix* **A** of a molecular graph $\mathcal{G}$ is an example of binary sparse matrix, only the off-diagonal entries *i–j* corresponding to pairs of adjacent vertices v_i and v_j, that is, vertices connected by a bond, being equal to one. Using the adjacency matrix as the multiplier in the Hadamard product, it follows

$$ {}^{1}\mathbf{M} \equiv \mathbf{M}_e = \mathbf{M} \otimes \mathbf{A} $$

where ${}^{1}\mathbf{M}$ is a **first-order sparse matrix**, also called edge-matrix, denoted as $\mathbf{M}_e$.

Opposite to sparse matrices are **dense matrices**, that is, matrices with several nonzero entries [Randić and DeAlba, 1997].

• **stochastic matrices** (≡ *probability matrices, transition matrices*)
These are square matrices **M** for which each row sum, *right stochastic matrices*, or each column sum, *left stochastic matrices*, is equal to 1, that is, the row elements or the column elements consist of nonnegative real numbers that can be interpreted as probabilities:

$$ VS_i(\mathbf{M}) = 1 \quad \text{or} \quad CS_j(\mathbf{M}) = 1 $$

where *VS* is the → *row sum operator* and *CS* the → *column sum operator*.

Stochastic matrices for which both row and column sums are equal to 1 are called *double stochastic matrices*. Stochastic matrices are defined in the framework of the → *MARCH-INSIDE descriptors*, → *TOMOCOMD descriptors*, and → *walk counts*.

• **sum of matrices**
Let **A** (*n*, *p*) and **B** (*n*, *p*) be two equal-sized matrices. The sum of the two matrices is defined as

$$ \mathbf{A} + \mathbf{B} = \mathbf{C} $$

Each scalar element c_{ij} of the matrix **C** is obtained by summing up the corresponding elements of the two matrices, that is,

$$ c_{ij} = a_{ij} + b_{ij} $$

A basic condition for the sum of two matrices is that the two matrices have the same dimension.

The main properties of the sum of two matrices are

$$ \text{(a) } \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \qquad \text{(b) } (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \qquad \text{(c) } \alpha(\mathbf{A} + \mathbf{B}) = \alpha\,\mathbf{A} + \alpha\,\mathrm{B} $$

where α is a scalar value.

• **total sum operator**
This operator S performs the sum of all of the elements of a matrix $\mathbf{M}$ of size $n \times p$:

$$S(\mathbf{M}) \equiv \sum_{i=1}^{n} \sum_{j=1}^{p} m_{ij} = \sum_{i=1}^{n} VS_i(\mathbf{M}) = \sum_{j=1}^{p} CS_j(\mathbf{M})$$

where VS_i and CS_j are the row sum operator and the column sum operator, respectively.

• **trace**
The trace of a square matrix $\mathbf{M}$ (i.e., $n = p$), denoted by $tr(\mathbf{M})$, is the sum of the diagonal elements:

$$tr(\mathbf{M}) \equiv \sum_{i=1}^{n} m_{ii}$$

• **transposition of a matrix**
The matrix $\mathbf{M}^{\mathrm{T}}$ is the transposed matrix of $\mathbf{M}$ if its elements are

$$[\mathbf{M}^{\mathrm{T}}]_{ij} = [\mathbf{M}]_{ji}$$

If the dimension of $\mathbf{M}$ is $n \times p$, the transposed matrix $\mathbf{M}^{\mathrm{T}}$ has dimension $p \times n$.

• **unit matrix (U)**
The unit matrix is a square matrix defined as

$$\mathbf{U} = \begin{vmatrix} 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & \dots & 1 \\ 1 & \dots & 1 & \dots & 1 \end{vmatrix}$$

➢ **algebraic semisum charge transfer index** → topological charge indices
➢ **algebraic structure count** → Kekulé number

■ alignment rules

In most → *grid-based QSAR techniques*, which use as the molecular descriptors energy values of → *molecular interaction fields* (steric, hydrophobic, coulombic, etc.), rules for alignment of all the molecules in the data set are required for comparability purposes. In effect, the energy value at each grid point **p** depends on the relative orientation of the compound with respect to the grid. As a consequence, the use of the grid points as molecular descriptors requires the mandatory step of aligning the molecules of the considered → *data set* in such a way that each of the thousands of grid points represents, for all the molecules, the same kind of information, and not spurious information due to lack of invariance in the rotation of the molecules in the grid.

Therefore, in applying grid-based QSAR techniques there are, in most cases, two closely related problems: the selection of a suitable molecular conformation for each compound and the relative alignment of the compounds, either among themselves or with respect to any → *receptor*, if its structure is known.

The ideal choice of conformers for QSAR would be the bioactive one. Wherever experimental structural data (e.g., X-ray data) on ligands bound to targets exist, the bioactive ligand conformation is available and should be used to derive an alignment rule.

When no structural data are available for the receptor, methods that explore conformational space may find the best relative match among the different ligands. During this process, low-energy conformations are selected to obtain the best match from all the different conformations. The solution is usually not unique because other conformations may bind to the unknown receptor, and multiple alignment rules, based on different starting hypotheses, should be considered when no structural information and no rigid compounds are available.

The success or failure of the grid-based methods to find acceptable → *quantitative structure–activity relationships* strongly depends on how the molecules are aligned in the grid on which the molecular interaction fields are sampled. In effect, problems may be mainly due to (a) an alignment that leads to a resulting common structure, that is, the pharmacophore, not reliable, and (b) the same grid points in different molecules represent chance variation in model geometry.

To avoid the drawbacks of the molecule alignment, several approaches based on different criteria were proposed; two basic alignment techniques are explained below.

- **point-by-point alignment**

For a set of congeneric compounds, the atoms of each compound are superimposed on their common backbone, aligning as much of each structure as possible.

For structurally diverse compounds, hypotheses on the → *pharmacophore* can provide an approach to overcome ambiguities in atom superimposition and identify a suitable alignment.

- **field-fitting alignment**

In this approach, the molecules are aligned by maximizing the degree of similarity between their molecular interaction fields. Different types of probes result in different fields as well as different molecular alignments. Therefore, the selection of suitable fields (and how to weight them) depends on external considerations.

Moreover, a difficulty in field-based alignment is that molecular regions not relevant, that is, parts of the molecule not involved in ligand–receptor interactions, may distort the alignment.

📖 [Kato, Itai *et al.*, 1987; Mayer, Naylor *et al.*, 1987; Kearsley and Smith, 1990; Manaut, Sanz *et al.*, 1991; Cramer III, DePriest *et al.*, 1993; Dean, 1993; Klebe, 1993, 1998; Waller and Marshall, 1993; Waller, Oprea *et al.*, 1993; Klebe, Mietzner *et al.*, 1994; Cramer III, Clark *et al.*, 1996; Petitjean, 1996; Greco, Novellino *et al.*, 1997; Handschuh, Wagener *et al.*, 1998; Langer and Hoffmann, 1998b; Norinder, 1998; Bernard, Kireev *et al.*, 1999; Robinson, Lyne *et al.*, 1999; Lemmen and Lengauer, 2000; Vedani, McMasters *et al.*, 2000; Jewell, Turner *et al.*, 2001; Makhija and Kulkarni, 2001b; Nissink, Verdonk *et al.*, 2001; Pitman, Huber *et al.*, 2001; Wildman and Crippen, 2001; Zhu, Hou *et al.*, 2001; Bringmann and Rummey, 2003; Bultinck, Kuppens *et al.*, 2003; Bultinck, Carbó-Dorca *et al.*, 2003; Hasegawa, Arakawa *et al.*, 2003; Marialke, Körner *et al.*, 2007]

➢ **Alikhanidi vertex degree** → vertex degree
➢ **all-path matrix** → path counts
➢ **all-path Wiener index** → path counts
➢ **all possible models** → variable selection
➢ **ALOGP** → lipophilicity descriptors (⊙ Ghose–Crippen hydrophobic atomic constants)

ALPHA descriptor

This is a vectorial molecular descriptor derived from the trajectories obtained by molecular dynamic simulation applying a technique of Gaussian smoothing [Tuppurainen, Viisas *et al.*, 2004]. For each trajectory coordinate *x*, the *ALPHA* descriptor is defined as

$$ALPHA(x) = \sum_{i=1}^{N} \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\left[(x-\alpha_i)^2/2 \cdot \sigma^2\right]}$$

where α and σ are the mean and standard deviation of the Gaussian function and the summation denotes over N-overlaid Gaussian functions; α values are first transformed to a bounded range (e.g., 0.5–3). Then, a Gaussian kernel of fixed standard deviation σ (a parameter to be optimized) is placed over each α value. Finally, the quantity *ALPHA*(*x*) is calculated at intervals of L (usually L is set at $\sigma/2$) resulting into a (pseudo) spectrum, which can be used as a molecular descriptor for QSAR modeling.

The dimensionality of the *ALPHA* descriptor is high (depending strongly on the value of σ) and, thus, the PLS method is suggested to compress the data.

- **Altenburg polynomial** → counting polynomials
- **altered Wiener indices** → Wiener index
- **amino acid descriptors** → biodescriptors
- **amino acid sequences** ≡ *peptide sequences*
- **Amoore shape indices** → shape descriptors

amphiphilic moments

The amphiphilicity of a compound is defined as the difference between the free energy of transfer of a compound from the aqueous phase to the air–water interface and the free energy of micelle formation and is quantified by means of surface tension measurements.

Amphiphilic moments are defined as vectors pointing from the center of the hydrophobic domain to the center of the hydrophilic domain of a molecule. It is defined as [Fischer, Gottschlich *et al.*, 1998; Fischer, Kansy *et al.*, 2001]:

$$\bar{A} = \sum_{i=1}^{A} d_i \cdot \alpha_i$$

where *d* is the distance of an identified charged residue from the farthest hydrophobic/hydrophilic residues. Each atom is weighted by its hydrophobic/hydrophilic property α on the basis of an atom contribution method [Meylan and Howard, 1995].

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 [Cruciani, Crivori *et al.*, 2000].

- **AMSP** ≡ *Autocorrelation of Molecular Surface Properties* → autocorrelation descriptors
- **Andrews' curves** → molecular descriptors (⊙ transformations of molecular descriptors)
- **Andrews descriptors** → count descriptors
- **angular distance** → similarity/diversity
- **angular separation** → similarity/diversity (⊙ Table S7)
- **a_N-index** → determinant-based descriptors (⊙ general a_N-index)

- **anisometry** → shape descriptors
- **anisotropy of the polarizability** → electric polarization descriptors
- **Ant Colony fitness function** → regression parameters
- **antibonding orbital information index** → information theoretic topological index
- **anticonnectivity indices** → variable descriptors

applicability domain

The concept of the applicability domain concerns the predictive use of QSAR/QSPR models and, then, is closely related to the concept of model validation (→ *validation techniques*). In other words, the applicability domain is a concept related to the quality of the QSAR/QSPR model predictions and prevention of the potential misuse of model's results. A key component of the prediction quality is indeed to define when a QSAR/QSPR model is suitable to predict a property/activity of a new compound [Tropsha, Gramatica *et al.*, 2003; Jaworska, Nikolova-Jeliazkova *et al.*, 2004; Dimitrov, Dimitrova *et al.*, 2005; Jaworska, Nikolova-Jeliazkova *et al.*, 2005; Netzeva, Worth *et al.*, 2005; Nikolova-Jeliazkova and Jaworska, 2005].

A model will yield reliable predictions when model assumptions are fulfilled and unreliable predictions when they are violated. In particular, for QSAR/QSPR models, based on statistical mining techniques, the → *training set* and the model prediction space are the basis for the estimation of space where predictions are reliable.

Two basic approaches were proposed for evaluating the applicability domain.

The first approach to applicability domain evaluation is the statistical analysis of the training set, trying to define the best conditions for interpolated prediction that is usually more reliable than extrapolation. Extrapolation is not a problem in principle, because extrapolated results from theoretically well-founded models can often be reliable. However, QSAR/QSPR models are usually based on empirical, and limited experimental evidence and/or are only locally valid; therefore, extrapolation usually results in high uncertainty and not reliable predictions.

Different approaches to estimate interpolation regions in a multivariate space were evaluated by Jaworska [Jaworska, Nikolova-Jeliazkova *et al.*, 2005], based on (a) ranges of the descriptor space; (b) distance-based methods, using Euclidean, Manhattan, and Mahalanobis distances, Hotelling T^2 method and leverage values; and (c) probability density distribution methods based on parametric and nonparametric approaches. Both ranges and distance-based methods were also evaluated in the principal component space by → *Principal Component Analysis*.

Another approach to applicability domain evaluation is based on the → *similarity/diversity* of the compound considered with respect to those belonging to the training set; a QSAR/QSPR prediction should be reliable if the compound is, in some way, similar to one or more compounds present in the training set [Nikolova and Jaworska, 2003]. High similarity is simply another way to use the interpolation ability of the model in place of the extrapolation.

A stepwise procedure was also proposed [Dimitrov, Dimitrova *et al.*, 2005] based on a four-stage procedure: (1) a study of the variations of molecular parameters that may affect the quality of the measured endpoint significantly (e.g., molecular weight, absorption, water solubility volatility, etc.); (2) an analysis of the structural domain based on a set of → *atom-centered fragment descriptors* that could be used to characterize the structural domain of the atoms present in the training set; (3) an analysis of the mechanistic domain focused on functional groups whose reactivity modulates the endpoint studied or structural fragments used in group contribution models; and (4) an analysis of the metabolic domain by simulators, although the metabolic aspects are usually not included in QSAR models.

🕮 [Martin, Kofron *et al.*, 2002; Eriksson, Jaworska *et al.*, 2003; Papa, Villa *et al.*, 2005; Tetko, Bruneau *et al.*, 2006; Zhang, Golbraikh *et al.*, 2006; Stanforth, Kolossov *et al.*, 2007]

- **arcs** → graph
- **arithmetic mean** → statistical indices (⊙ indices of central tendency)
- **arithmetic topological index** → vertex degree
- **aromatic bond count** → multiple bond descriptors
- **aromaticity** → delocalization degree indices
- **aromaticity indices** → delocalization degree indices
- **artificial neural networks** → chemometrics
- **aryl electronic constants** → electronic substituent constants
- ***ASIIg* index** → charge descriptors (⊙ charge-related indices)
- **asphericity** → shape descriptors
- **association coefficients** → similarity/diversity
- **asymptotic Q^2 rule** → regression parameters
- **ATAC** ≡ *Atom-Type AutoCorrelation* → autocorrelation descriptors
- **atom–atom polarizability** → electric polarization descriptors
- **atom-centered fragment descriptors** ≡ *Augmented Atoms* → substructure descriptors
- **atom connectivity matrices** → weighted matrices (⊙ weighted adjacency matrices)
- **atom count** ≡ *atom number*
- **atom detour eccentricity** → detour matrix
- **atom eccentricity** → distance matrix
- **atom electronegativity** → atomic properties
- **Atom Environment descriptors** → substructure descriptors (⊙ fingerprints)
- **atomic charges** → quantum-chemical descriptors
- **atomic charge-weighted negative surface area** → charged partial surface area descriptors
- **atomic charge-weighted positive surface area** → charged partial surface area descriptors

■ **atomic composition indices** (≡ *composition indices*)

Molecular → *0D descriptors* with high degeneracy, derived from the chemical formula of compounds and defined as → *information indices* of the elemental composition of the molecule. They can be considered → *molecular complexity indices* that take into account the molecular diversity in terms of different atom types.

→ *Average molecular weight* and → *relative atom-type count* are simple molecular descriptors that encode information on atomic composition. Other important descriptors of the atomic composition are based on the → *total information content* and the → *mean information content*, defined as

• **total information index on atomic composition** (I_{AC})

The total information content on atomic composition of the molecule is calculated from the complete molecular formula, hydrogen included, as

$$I_{AC} = A^h \cdot \log_2 A^h - \sum_g A_g \cdot \log_2 A_g$$

where A^h is the total number of atoms (hydrogen included) and A_g is the number of atoms of chemical element of type g [Dancoff and Quastler, 1953].

For example, benzene has 6 carbon and 6 hydrogen atoms; then, as $A^h = 12$ and $A_g = 6$ for both equivalence classes, $I_{AC} = 12$.

• **mean information index on atomic composition ($\bar{I}_{AC}$)**
The mean information content on atomic composition is the mean value of the total information content and is calculated as

$$\bar{I}_{AC} = -\sum_g \frac{A_g}{A^h} \cdot \log_2 \frac{A_g}{A^h} = -\sum_g p_g \cdot \log_2 p_g$$

where A^h is the total number of atoms (hydrogen included), A_g is the number of atoms of type g and p_g is the probability to randomly select a gth atom type [Dancoff and Quastler, 1953]. For example, for benzene $\bar{I}_{AC} = 1$.

- **atomic connectivity indices** ≡ *local connectivity indices* → connectivity indices
- **atomic dispersion coefficient** → hydration free energy density
- **Atomic Environment Autocorrelations** → autocorrelation descriptors
- **atomic ID numbers** → ID numbers
- **atomic information content** → atomic information indices

atomic information indices

Atomic descriptors related to the internal composition of atoms [Bonchev, 1983].

The **atomic information content** I_{at} is the → *total information content* of an atom viewed as a system whose structural elements, that is, protons p, neutrons n, and electrons el, are partitioned into nucleons, p + n, and electrons el:

$$I_{at} = (N_n + N_p + N_{el}) \cdot \log_2(N_n + N_p + N_{el}) - N_{el} \cdot \log_2 N_{el} - (N_n + N_p) \cdot \log_2(N_n + N_p)$$

where N_n, N_p, and N_{el} are the numbers of neutrons, protons, and electrons, respectively [Bonchev and Peev, 1973].

To account for the different isotopes of a given chemical element, the **information index on isotopic composition** was defined as

$$I_{IC} = \sum_k (I_{at})_k \cdot f_k$$

where the sum runs over all isotopes of the considered chemical element, $(I_{at})_k$ is the atomic information content of the kth isotope and f_k is its relative amount [Bonchev and Peev, 1973].

The **information index on proton–neutron composition** is an atomic descriptor defined as total information content of the atomic nucleus:

$$I^{n,p} = (N_n + N_p) \cdot \log_2(N_n + N_p) - N_n \cdot \log_2 N_n - N_p \cdot \log_2 N_p$$

where N_n and N_p are the numbers of neutrons and protons, respectively [Bonchev, Peev *et al.*, 1976].

The **nuclear information content** I_{NUCL} is a molecular descriptor calculated as the sum of information indices on the proton–neutron composition of all the nuclei of a molecule:

$$I_{NUCL} = \sum_{i=1}^{A} I_i^{n,p}$$

where A is the number of atoms and $I_i^{n,p}$ is the information index on proton–neutron composition of the nucleus of the ith atom. This index also accounts for molecular size by means of the number of atomic nuclei.

- **atomic molecular connectivity index** → connectivity indices
- **atomic moments of energy** → self-returning walk counts
- **atomic multigraph factor** → bond order indices (⊙ conventional bond order)
- **atomic path count** → path counts
- **atomic path count sum** → path counts
- **atomic path number** ≡ *atomic path count* → path counts
- **atomic path/walk indices** → shape descriptors (⊙ path/walk shape indices)
- **atomic polarization** → electric polarization descriptors

atomic properties

"Most atomic properties are a consequence of atomic structure, which in turn must be related to the inherent nature of the component electrons and nuclei. Therefore it is almost inevitable that such properties be related to one another, if only because of their common origin. It should not be surprising that a particular property, here the electronegativity, can be derived from or correlated with a wide variety of other properties, with reasonable agreement among the several results" [Sanderson, 1988].

Atomic properties P are physics and chemical observables characterizing each chemical element. They play a fundamental role in the definition of most of the molecular descriptors, being physico-chemical properties, as well as biological, toxicological, and environmental properties, deeply determined by the chemical elements constituting the molecule itself.

In Table A2, some important atomic properties are listed for the most common chemical elements.

Table A2 Atomic properties for some chemical elements.

Atom	Z	L	Z^v	R^{vdw}	R^{cov}	m	V^{vdw}	χ^{SA}	α	IP	EA
H	1	1	1	1.17	0.37	1.01	6.71	2.59	0.67	13.598	0.754
Li	3	2	1	1.82	1.34	6.94	25.25	0.89	24.3	5.392	0.618
Be	4	2	2	–	0.90	9.01	–	1.81	5.60	9.323	–
B	5	2	3	1.62	0.82	10.81	17.88	2.28	3.03	8.298	0.277
C	6	2	4	1.75	0.77	12.01	22.45	2.75	1.76	11.260	1.263
N	7	2	5	1.55	0.75	14.01	15.60	3.19	1.10	14.534	–
O	8	2	6	1.40	0.73	16.00	11.49	3.65	0.80	13.618	1.461
F	9	2	7	1.30	0.71	19.00	9.20	4.00	0.56	17.423	3.401
Na	11	3	1	2.27	1.54	22.99	49.00	0.56	23.6	5.139	0.548
Mg	12	3	2	1.73	1.30	24.31	21.69	1.32	10.6	7.646	–
Al	13	3	3	2.06	1.18	26.98	36.51	1.71	6.80	5.986	0.441
Si	14	3	4	1.97	1.11	28.09	31.98	2.14	5.38	8.152	1.385
P	15	3	5	1.85	1.06	30.97	26.52	2.52	3.63	10.487	0.747
S	16	3	6	1.80	1.02	32.07	24.43	2.96	2.90	10.360	2.077
Cl	17	3	7	1.75	0.99	35.45	22.45	3.48	2.18	12.968	3.613

(*Continued*)

Table A2 (*Continued*)

Atom	Z	L	Z^{v}	R^{vdw}	R^{cov}	m	V^{vdw}	χ^{SA}	α	IP	EA
K	19	4	1	2.75	1.96	39.10	87.11	0.45	43.4	4.341	0.501
Ca	20	4	2	–	1.74	40.08	–	0.95	22.8	6.113	0.018
Cr	24	4	6	2.20	1.27	52.00	44.60	1.66	11.60	6.767	0.666
Mn	25	4	7	2.18	1.39	54.94	43.40	2.20	9.40	7.434	–
Fe	26	4	8	2.14	1.25	55.85	41.05	2.20	8.40	7.902	0.151
Co	27	4	9	2.03	1.26	58.93	35.04	2.56	7.50	7.881	0.662
Ni	28	4	10	1.60	1.21	58.69	17.16	1.94	6.80	7.640	1.156
Cu	29	4	11	1.40	1.38	63.55	11.49	1.98	6.10	7.723	1.235
Zn	30	4	12	1.39	1.31	65.39	11.25	2.23	7.10	9.394	–
Ga	31	4	3	1.87	1.26	69.72	27.39	2.42	8.12	5.999	0.300
Ge	32	4	4	1.90	1.22	72.61	28.73	2.62	6.07	7.900	1.233
As	33	4	5	1.85	1.19	74.92	26.52	2.82	4.31	9.815	0.810
Se	34	4	6	1.90	1.16	78.96	28.73	3.01	3.77	9.752	2.021
Br	35	4	7	1.95	1.14	79.90	31.06	3.22	3.05	11.814	3.364
Rb	37	5	1	–	2.11	85.47	–	0.31	47.3	4.177	0.486
Sr	38	5	2	–	1.92	87.62	–	0.72	27.6	5.695	0.110
Mo	42	5	6	2.00	1.45	95.94	33.51	1.15	12.80	7.092	0.746
Ag	47	5	11	1.72	1.53	107.87	21.31	1.83	7.20	7.576	1.302
Cd	48	5	12	1.58	1.48	112.41	16.52	1.98	7.20	8.994	–
In	49	5	3	1.93	1.44	114.82	30.11	2.14	10.20	5.786	0.300
Sn	50	5	4	2.22	1.41	118.71	45.83	2.30	7.70	7.344	1.112
Sb	51	5	5	2.10	1.38	121.76	38.79	2.46	6.60	8.640	1.070
Te	52	5	6	2.06	1.35	127.60	36.62	2.62	5.50	9.010	1.971
I	53	5	7	2.10	1.33	126.90	38.79	2.78	5.35	10.451	3.059
Gd	64	6	10	2.59	1.79	157.25	72.78	2.00	23.50	6.150	0.500
Pt	78	6	10	1.75	1.28	195.08	22.45	2.28	6.50	9.000	2.128
Au	79	6	11	1.66	1.44	196.97	19.16	2.54	5.80	9.226	2.309
Hg	80	6	12	1.55	1.49	200.59	15.60	2.20	5.70	10.438	–
Tl	81	6	3	1.96	1.48	204.38	31.54	2.25	7.60	6.108	0.200
Pb	82	6	4	2.02	1.47	207.20	34.53	2.29	6.80	7.417	0.364
Bi	83	6	5	2.10	1.46	208.98	38.79	2.34	7.40	7.289	0.946

Z, atomic number; L, principal quantum number; Z^{v}, number of valence electrons; R^{vdw}, van der Waals atomic radius; R^{cov}, covalent radius; m, atomic mass; V^{vdw}, van der Waals volume; χ^{SA}, Sanderson electronegativity; α, atomic polarizability (10^{-24} cm^3); IP, ionization potential (eV); EA, electron affinity (eV).

For atomic mass, van der Waals volume, Sanderson electronegativity, and atom polarizability, the scaled values with respect to the carbon atom are listed in Table A3.

Table A3 Atomic mass (m), van der Waals volume (V^{vdw}), Sanderson electronegativity (χ^{SA}), and polarizability (α): original values and scaled values with respect to the carbon atom value.

	Atomic mass		Volume		Electronegativity		Polarizability	
ID	m	m/m_C	V^{vdw}	V^{vdw}/V_C^{vdw}	χ^{SA}	χ^{SA}/χ_C^{SA}	α	α/α_C
H	1.01	0.084	6.709	0.299	2.592	0.944	0.667	0.379
B	10.81	0.900	17.875	0.796	2.275	0.828	3.030	1.722

(*Continued*)

Table A3 (*Continued*)

	Atomic mass		Volume		Electronegativity		Polarizability	
ID	**m**	m/m_C	V^{vdw}	V^{vdw}/V_C^{vdw}	χ^{SA}	χ^{SA}/χ_C^{SA}	α	α/α_C
C	12.01	1.000	22.449	1.000	2.746	1.000	1.760	1.000
N	14.01	1.166	15.599	0.695	3.194	1.163	1.100	0.625
O	16.00	1.332	11.494	0.512	3.654	1.331	0.802	0.456
F	19.00	1.582	9.203	0.410	4.000	1.457	0.557	0.316
Al	26.98	2.246	36.511	1.626	1.714	0.624	6.800	3.864
Si	28.09	2.339	31.976	1.424	2.138	0.779	5.380	3.057
P	30.97	2.579	26.522	1.181	2.515	0.916	3.630	2.063
S	32.07	2.670	24.429	1.088	2.957	1.077	2.900	1.648
Cl	35.45	2.952	23.228	1.035	3.475	1.265	2.180	1.239
Fe	55.85	4.650	41.052	1.829	2.000	0.728	8.400	4.773
Co	58.93	4.907	35.041	1.561	2.000	0.728	7.500	4.261
Ni	58.69	4.887	17.157	0.764	2.000	0.728	6.800	3.864
Cu	63.55	5.291	11.494	0.512	2.033	0.740	6.100	3.466
Zn	65.39	5.445	38.351	1.708	2.223	0.810	7.100	4.034
Br	79.90	6.653	31.059	1.384	3.219	1.172	3.050	1.733
Sn	118.71	9.884	45.830	2.042	2.298	0.837	7.700	4.375
I	126.90	10.566	38.792	1.728	2.778	1.012	5.350	3.040

Other atomic electronegativity scales are reported elsewhere (→ *electronegativity*, Table E7).

- **atomic refractivity** → physico-chemical properties (⊙ molar refractivity)
- **atomic self-returning walk count** → self-returning walk counts
- **atomic sequence count** → sequence matrices

■ **atomic solvation parameter** ($\Delta\sigma$)

An empirical atomic descriptor $\Delta\sigma$ proposed to calculate solvation free energy of a group X in terms of atomic contributions by the following equation:

$$\Delta G_X = \sum\nolimits_i \Delta\sigma_i \cdot SA_i$$

where the sum runs over all the nonhydrogen atoms of the X group, *SA* is the → *solvent accessible surface area* of the *i*th atom, and $\Delta\sigma$ denotes the corresponding atomic contribution to solvation energy [Eisenberg and McLachlan, 1986]. The atomic contributions $\Delta\sigma_i \cdot SA_i$ are the free energy of transfer of each atom to the solution; note that the areas *SA* depend on molecule conformation. Proposed to study protein folding and binding, the estimated values for the atomic solvation parameters $\Delta\sigma$ (in cal $Å^{-2}$ mol^{-1}) are $\Delta\sigma_C = 16$, $\Delta\sigma_N = -6$, $\Delta\sigma_O = -6$, $\Delta\sigma_{N^+} = -50$, $\Delta\sigma_{O^-} = 24$, and $\Delta\sigma_S = 21$ for carbon, nitrogen, oxygen, nitrogen cation, oxygen anion, and sulfur, respectively.

- **atomic valency index** → quantum-chemical descriptors
- **atomic walk count** → walk counts
- **atomic walk count sum** → walk counts

- **atomic weight-weighted adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **atomic weight-weighted distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **Atom-in-Molecules theory** ≡ *AIM theory*
- **atom-in-structure invariant index** → charge descriptors (⊙ charge-related indices)
- **atomistic topological indices** → count descriptors
- **atom-level composite ETA index** → ETA indices
- **atom leverage-based center** → center of a molecule

atom number (A) (≡ *atom count*)

This is the simplest measure with regard to molecular size, defined as the total number of atoms in a molecule. It is a global, zero dimensional, descriptor with a high degeneracy. In several applications for the calculation of → *molecular descriptors*, the atom number A refers only to nonhydrogen atoms.

The **information index on size** is the → *total information content* on the atom number, defined as

$$I_{\mathrm{SIZE}} = A^h \log_2 A^h$$

where the atom number A^h also takes hydrogen atoms into account [Bertz, 1981]. This index can also be calculated without considering hydrogen atoms.

Other related molecular descriptors are → *atomic composition indices*, several → *information indices* and → *graph invariants*.

- **atom-pair matching function** → molecular shape analysis
- **atom pairs** → substructure descriptors
- **atom polarizability** → electric polarization descriptors
- **Atom-Type AutoCorrelation** → autocorrelation descriptors
- **atom-type autocorrelation matrix** → weighted matrices (⊙ weighted distance matrices)

atom type-based topological indices

Atom-type topological indices are used to describe a molecule by information related to different atom types in the molecule. An atom-type index is usually derived from some properties of all the atoms of the same type and their structural environment. → *Atom-type E-state indices* of Kier and Hall, → *perturbation connectivity indices*, → *atom-type path counts*, and → *atom-type autocorrelation descriptors* are examples of these molecular descriptors.

AI indices are atom-type topological indices derived from the → *Xu index*, whose formula is applied to single atom types [Ren, 2002a, 2002b, 2002c, 2003a, 2003c, 2003d]. For any ith atom in the molecular graph, first a local vertex invariant, denoted as AI_i, is calculated as

$$AI_i = 1 + \phi_i = 1 + \frac{\delta_i^{\mathrm{m}} \cdot \sigma_i^2}{\sum_{i=1}^{A} \delta_i^{\mathrm{m}} \cdot \sigma_i}$$

where ϕ is a perturbation term reflecting the effects of the structural environment of the ith atom on the topological index and σ the → *vertex distance sum*. The expression defined above is based

on the → *Ren vertex degree* δ^m derived from the Kier–Hall valence vertex degree and defined as

$$\delta_i^m = \delta_i + \left[\left(\frac{2}{L_i}\right)^2 \cdot \delta_i^v + 1\right]^{-1} = \delta_i + (I_i \cdot \delta_i)^{-1}$$

where δ_i is the → *vertex degree* of the ith atom, L_i its principal quantum number, δ_i^v its → *valence vertex degree*, and I_i denotes the → *intrinsic state*. This formula is applied only to heteroatoms or carbon atoms with multiple bonds and/or bonded to heteroatoms; otherwise, the Ren vertex degree coincides with the simple vertex degree δ_i.

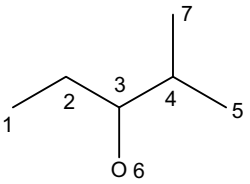
According to this definition of local vertex invariants, the *AI* index for the kth atom type is derived by adding the values of the local *AI* indices for all the atoms of type k.

$$AI(k) = \sum_{i=1}^{n_k} AI_i = n_k + \sum_{i=1}^{n_k} \phi_i = n_k + \frac{\sum_{i=1}^{n_k} \delta_i^m \cdot \sigma_i^2}{\sum_{i=1}^{A} \delta_i^m \cdot \sigma_i}$$

where n_k is the number of atoms of type k.

Example A5

Calculation of *AI* indices for 4-methyl-3-pentanol. σ is the vertex distance sum and δ^m the Ren vertex degree; **D** is the distance matrix.



D =

Atom	1	2	3	4	5	6	7	σ_i	δ_i^m
1	0	1	2	3	4	3	4	17	1.000
2	1	0	1	2	3	2	3	12	2.000
3	2	1	0	1	2	1	2	9	3.250
4	3	2	1	0	1	2	1	10	3.000
5	4	3	2	1	0	3	2	15	1.000
6	3	2	1	2	3	0	3	14	1.167
7	4	3	2	1	2	3	0	15	1.000

$$AI(-CH_3) = AI_1 + AI_5 + AI_7 = \left(1 + \frac{1 \times 17^2}{146.588}\right) + 2\left(1 + \frac{1 \times 15^2}{146.588}\right) = 8.041$$

$$AI(-CH_2-) = AI_2 = 1 + \frac{2 \times 12^2}{146.588} = 2.965$$

$$AI(-CH<) = AI_3 + AI_4 = \left(1 + \frac{3 \times 9^2}{146.588}\right) + \left(1 + \frac{3 \times 10^2}{146.588}\right) = 5.704$$

$$AI(-OH) = AI_6 = 1 + \frac{1.167 \times 14^2}{146.588} = 2.560$$

Based on the same approach as the *AI* indices but derived from the formula of the → *Lu index*, the **DAI indices** are atom-type topological indices that exploit bond length-weighted

interatomic distances calculated by adding the relative bond lengths of the edges along the shortest path [Lu, Guo *et al.*, 2006b, 2006c, 2006d; Lu, Wang *et al.*, 2006]. For any *i*th atom in the molecular graph, the local vertex invariant DAI_i is calculated as

$$DAI_i = 1 + \phi_i = 1 + A\frac{\sum_{j=1}^{A}[\mathbf{D}(r^*)]_{ij}}{\sum_{i=1}^{A}\sum_{j=1}^{A}[\mathbf{D}(r^*)]_{ij}}$$

where ϕ is the perturbation term relative to the atom environment, A is the number of atoms, and $[\mathbf{D}(r^*)]_{ij}$ are the elements of the → *bond length-weighted distance matrix*.

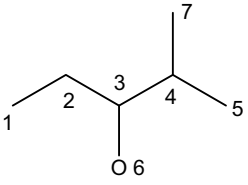
The *DAI* index for the *k*th atom type is calculated by adding contributions of all atoms of the considered type:

$$DAI(k) = \sum_{i=1}^{n_k} DAI_i = n_k + \sum_{i=1}^{n_k}\phi_i$$

where n_k is the number of atoms of type k.

Example A6

Calculation of *DAI* indices for 4-methyl-3-pentanol. VS_i indicates the matrix row sums; $\mathbf{D}(r^*)$ is the bond length-weighted distance matrix.



$\mathbf{D}(r^*)=$

Atom	1	2	3	4	5	6	7	VS_i
1	0	1	2	3	4	2.928	4	16.928
2	1	0	1	2	3	1.928	3	11.928
3	2	1	0	1	2	0.928	2	8.928
4	3	2	1	0	1	1.928	1	9.928
5	4	3	3	1	0	2.928	2	14.928
6	2.928	1.928	0.928	1.928	2.928	0	2.928	13.571
7	4	3	2	1	2	2.928	0	14.928

$$DAI(-CH_3) = DAI_1 + DAI_5 + DAI_7 = \left(1 + 7 \times \frac{16.928}{91.143}\right) + 2\left(1 + 7 \times \frac{14.928}{91.143}\right) = 6.593$$

$$DAI(-CH_2-) = DAI_2 = 1 + 7 \times \frac{11.928}{91.143} = 1.916$$

$$DAI(-CH<) = DAI_3 + DAI_4 = \left(1 + 7 \times \frac{8.928}{91.143}\right) + \left(1 + 7 \times \frac{9.928}{91.143}\right) = 3.448$$

$$DAI(-OH) = DAI_6 = 1 + 7 \times \frac{13.571}{91.143} = 2.042$$

- **atom-type count** → count descriptors
- **atom-type *E*-state counts** → electrotopological state indices
- **atom-type *E*-state indices** → electrotopological state indices
- **atom-type *HE*-state indices** → electrotopological state indices

- **atom-type interaction matrices** → weighted matrices (⊙ weighted distance matrices)
- ***ATS* descriptor** → autocorrelation descriptors (⊙ Moreau–Broto autocorrelation)
- **attractive steric effects** → minimal topological difference
- **augmented adjacency matrix** → weighted matrices (⊙ weighted adjacency matrices)
- **Augmented Atom keys** → substructure descriptors
- **Augmented Atoms** → substructure descriptors
- **augmented connectivity** → eccentricity-based Madan indices
- **augmented distance matrix** → weighted matrices (⊙ weighted distance matrices)
- **augmented eccentric connectivity index** → eccentricity-based Madan indices (⊙ Table E1)
- **augmented edge adjacency matrix** → edge adjacency matrix
- **augmented matrices** → matrices of molecules
- **augmented pair descriptors** → substructure descriptors
- **augmented valence** → vertex degree
- **augmented vertex degree** → weighted matrices (⊙ weighted adjacency matrices)
- **augmented vertex degree matrix** → weighted matrices (⊙ weighted distance matrices)
- **Austel branching index** → steric descriptors

autocorrelation descriptors

→ *Molecular descriptors* based on the autocorrelation function AC_k, defined as

$$AC_k = \int_a^b f(x) \cdot f(x+k) \cdot dx$$

where $f(x)$ is any function of the variable x and k is the lag representing an interval of x, and a and b define the total studied interval of the function. The function $f(x)$ is usually a time-dependent function such as a time-dependent electrical signal or a spatial-dependent function such as the population density in space. Then, autocorrelation measures the strength of a relationship between observations as a function of the time or space separation between them [Moreau and Turpin, 1996].

The autocorrelation function AC_k is the integration of the products of the function values calculated at x and $x + k$. This function expresses how numerical values of the function at intervals equal to the lag are correlated.

Autocorrelation functions AC_k can also be calculated for any ordered discrete sequence of n values $f(x_i)$ by summing the products of the ith value and the $(i + k)$th value as

$$AC_k = \frac{1}{(n-k) \cdot \sigma^2} \cdot \sum_{i=1}^{n-k} [(f(x_i) - \mu) \cdot (f(x_{i+k}) - \mu)]$$

where k is the lag, σ^2 is the variance of the function values, and μ is their mean. The lag assumes values between 1 and K, where the maximum value K can be $n-1$; however, in several applications, K is chosen equal to a small number ($K<8$). A lag value of zero corresponds to the sum of the square-centered values of the function.

Note that it is common practice in many disciplines to use the term *autocorrelation* even if the standardization by σ^2 is not applied; in this case, the correct term should be *autocovariance*.

Autocorrelation descriptors of chemical compounds are calculated by using various molecular properties that can be represented at the atomic level or molecular surface level or else.

A property of the autocorrelation function is that it does not change when the origin of the x variable is shifted. In effect, autocorrelation descriptors are considered → *TRI descriptors*, meaning that they have translational and rotational invariance.

Based on the same principles as the autocorrelation descriptors, but calculated contemporarily on two different properties $f(x)$ and $g(x)$, **cross-correlation descriptors** are calculated to measure the strength of relationships between the two considered properties. For any two ordered sequences comprised of a number of discrete values, the cross-correlation is calculated by summing the products of the ith value of the first sequence and the $(i + k)$th value of the second sequence as

$$CC_k = \frac{1}{(n-k) \cdot \sigma_{f(x)} \cdot \sigma_{g(x)}} \cdot \sum_{i=1}^{n-k} \left[(f(x_i) - \mu_{f(x)}) \cdot (g(x_{i+k}) - \mu_{g(x)}) \right]$$

where n is the lowest cardinality of the two sets. For the autocorrelation, cross-correlation is usually calculated without the standardization by the two standard deviations $\sigma_{f(x)}$ and $\sigma_{g(x)}$; in this case, the correct term should be *cross-covariance*.

The most common spatial autocorrelation molecular descriptors are obtained by taking the molecule atoms as the set of discrete points in space and an atomic property as the function evaluated at those points.

Common weighting schemes w used to describe atoms in the molecule are → *physico-chemical properties* such as atomic masses, → *van der Waals volumes*, → *atomic electronegativities*, → *atomic polarizabilities*, covalent radius, and so on. Alternatively, the weighting scheme for atoms can be based on → *local vertex invariants* such as the topological → *vertex degrees*, Kier–Hall → *intrinsic states* or → *E-state indices*, → *normalized distance complexity index*, and related indices. Most of these → *weighting schemes* are implemented in DRAGON software [DRAGON – Talete s. r.l., 2007; Mauri, Consonni *et al.*, 2006] allowing calculation of different types of autocorrelation descriptors. A comparison of QSARs based on autocorrelation descriptors derived from different weighting schemes is reported in [Kabankin and Gabrielyan, 2005].

For spatial autocorrelation molecular descriptors calculated on a molecular graph, the lag k coincides with the → *topological distance* between any pair of vertices.

Autocorrelation descriptors can also be calculated from 3D spatial molecular geometry. In this case, the distribution of a molecular property can be evaluated by a mathematical function $f(x, y, z)$, x, y, and z being the spatial coordinates, defined either for each point of molecular space or molecular surface (i.e., a continuous property such as electronic density or molecular interaction energy) or only for points occupied by atoms (i.e., atomic properties) [Wagener, Sadowski *et al.*, 1995].

The plot of an ordered sequence of autocorrelation descriptors from lag 0 to lag K is called **autocorrelogram** and is usually used to describe a chemical compound in → *similarity/diversity* analysis.

Maximum Auto-Cross-Correlation descriptors (or ***MACC* descriptors**) are autocorrelation and cross-correlation descriptors calculated by taking into account only the maximum product of molecular properties for each lag k:

$$MACC_k = \max_i (f(x_i) \cdot g(x_{i+k}))$$

This function was applied to derive → *maximal R indices* and, with the name of MACC-2 transform, to calculate the → *GRIND descriptors*.

Moreover, a special case of autocorrelation descriptors is the **Atom-Type AutoCorrelation** (*ATAC*), which is calculated by summing property values only of atoms of given types. The simplest atom-type autocorrelation is given by

$$ATAC_k(u,v) = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij}(u,v) \cdot \delta(d_{ij};k)$$

where u and v denote two atom types. $\delta_{ij}(u,v)$ is a Kronecker delta function assuming a value equal to 1 if the atoms i and j form a pair of types u and v or, equivalently, of types v and u; $\delta(d_{ij};k)$ is a Kronecker delta function equal to 1 if the interatomic distance d_{ij} is equal to the lag k, and zero otherwise.

This descriptor is defined for each pair of atom types and simply encodes the occurrence numbers of the given atom type pair at different distance values. It can be normalized by using two different procedures: the first one consists in dividing each $ATAC_k$ value by the total number of atom pairs at distance k independently of their types; the second one consists in dividing each $ATAC_k$ value by a constant, which can be equal to the total number of atoms in the molecule or, alternatively, to the total number of (u,v) atom type pairs in the molecule.

Atom types can be defined in different ways; they can be defined in terms of the simple chemical elements or may account also for atom connectivity, hybridization states, and pharmacophoric features. Atom-type autocorrelations can be viewed as a special case of the → *atom-type interaction matrices*, from which other kinds of descriptors can also be derived.

Atom-type autocorrelations have been used to derive some → *substructure descriptors* such as → *atom pairs*, → *CATS descriptors*, and related descriptors.

Examples of autocorrelation descriptors, which are derived from the molecular graph but exploit 3D spatial information, are → *GETAWAY descriptors*, → *PEST Autocorrelation Descriptors* and → *SWM signals*. Other autocorrelation descriptors were derived from → *molecular shape field*.

A collection of other auto- and cross-correlation descriptors is discussed in the following sections.

• **Moreau–Broto autocorrelation** (≡ *Autocorrelation of a Topological Structure, ATS*)
This is the most known spatial autocorrelation defined on a molecular graph $\mathcal{G}$ as [Moreau and Broto, 1980a, 1980b; Broto, Moreau *et al.*, 1984a]

$$ATS_k = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} w_i \cdot w_j \cdot \delta(d_{ij};k) = \frac{1}{2} \cdot (\mathbf{w}^{\mathbf{T}} \cdot {}^{k}\mathbf{B} \cdot \mathbf{w})$$

where w is any atomic property, A is the number of atoms in a molecule, k is the lag, and d_{ij} is the topological distance between ith and jth atoms; $\delta(d_{ij};k)$ is a Kronecker delta function equal to 1 if $d_{ij} = k$, zero otherwise. ${}^{k}\mathbf{B}$ is the kth order → *geodesic matrix*, whose elements are equal to 1 only for vertices v_i and v_j at topological distance k, and zero otherwise; $\mathbf{w}$ is the A-dimensional vector of atomic properties. The autocorrelation ATS_0 defined for the path of length zero is calculated as

$$ATS_0 = \sum_{i=1}^{A} w_i^2$$

that is, the sum of the squares of the atomic properties. Typical atomic properties are atomic masses, polarizabilities, charges, and electronegativities.

Moreau–Broto autocorrelations can be viewed as a special case of the → *interaction geodesic matrices*, from which other kinds of descriptors can also be derived.

It has to be noted that atomic properties w should be centered by subtracting the average property value in the molecule to obtain proper autocorrelation values. Hollas demonstrated that only if properties are centered, all autocorrelation descriptors are uncorrelated thus becoming more suitable for subsequent statistical analysis [Hollas, 2002].

For each atomic property w, the set of the autocorrelation terms defined for all existing topological distances in the graph is the ***ATS* descriptor** defined as

$$\{ATS_0, ATS_1, ATS_2, \ \ldots \ , ATS_D\}_w$$

where D is the → *topological diameter*, that is, the maximum distance in the graph. The plot of the *ATS* descriptor is the corresponding autocorrelogram.

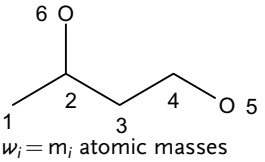
Average spatial autocorrelation descriptors are obtained by dividing each term by the corresponding number of contributions, thus avoiding any dependence on molecular size:

$$\overline{ATS}_k = \frac{1}{2\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} w_i \cdot w_j \cdot \delta(d_{ij}; k)$$

where Δ_k is the sum of the Kronecker delta, that is, the total number of vertex pairs at distance equal to k [Wagener, Sadowski *et al.*, 1995].

Example A7

Moreau–Broto autocorrelation descriptors calculated from the H-depleted molecular graph of 4-hydroxy-2-butanone.



$$ATS_0 = w_1^2 + w_2^2 + w_3^2 + w_4^2 + w_5^2 + w_6^2$$
$$ATS_1 = w_1 \cdot w_2 + w_2 \cdot w_3 + w_3 \cdot w_4 + w_4 \cdot w_5 + w_2 \cdot w_6$$
$$ATS_2 = w_1 \cdot w_3 + w_1 \cdot w_6 + w_2 \cdot w_4 + w_3 \cdot w_5 + w_3 \cdot w_6$$
$$ATS_3 = w_1 \cdot w_4 + w_2 \cdot w_5 + w_4 \cdot w_6$$

$w_i = m_i$ atomic masses

$$ATS_0 = 12^2 + 12^2 + 12^2 + 12^2 + 16^2 + 16^2 = 1088 \qquad \overline{ATS}_0 = 1088/6 = 181.3$$
$$ATS_1 = 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 816 \qquad \overline{ATS}_1 = 816/5 = 163.2$$
$$ATS_2 = 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 864 \qquad \overline{ATS}_2 = 864/5 = 172.8$$
$$ATS_3 = 12 \cdot 12 + 12 \cdot 16 + 12 \cdot 16 = 528 \qquad \overline{ATS}_3 = 528/3 = 176.0.$$

The *ATS* descriptor is a graph invariant describing how the property considered is distributed along the topological structure. Assuming an additive scheme, the *ATS* descriptor corresponds to a decomposition of the square molecular property Φ in different atomic contributions:

$$\Phi^2 = \left(\sum_{i=1}^{A} w_i \right)^2 = \sum_{i=1}^{A} w_i^2 + \sum\nolimits_{i \neq j} 2 \cdot w_i \cdot w_j = ATS_0 + 2 \cdot \sum_{k=1}^{D} ATS_k$$

where ATS_0 contains all atomic contributions to the square molecular property and ATS_k the interactions between each pair of atoms.

- **3D molecular autocorrelation**

Autocorrelation descriptors calculated for 3D spatial molecular geometry are based on interatomic distances collected in the → *geometry matrix* **G** instead of topological distances and the property function is still defined by the set of atomic properties.

The interatomic distance r is divided into elementary distance intervals of equal width (e.g., 0.5 Å). Each distance interval is defined by a lower and upper value of interatomic distance r_{ij}. All interatomic distances falling in the same interval are considered identical. For each distance interval, the autocorrelation function AC_k is obtained by summing all the products of the property values of atoms i and j whose interatomic distance r_{ij} falls within the considered interval $[r_u, r_v]_k$:

$$AC_k(r_u, r_v) = \sum_{i,j} w_i \cdot w_j \quad (r_u \leq r_{ij} \leq r_v)$$

📖 [Broto, Moreau *et al.*, 1984c; Broto and Devillers, 1990; Zakarya, Belkhadir *et al.*, 1993]

- **Moran coefficient** (I_k)

This is a general index of spatial autocorrelation that, if applied to a molecular graph, can be defined as

$$I_k = \frac{\frac{1}{\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} (w_i - \bar{w}) \cdot (w_j - \bar{w}) \cdot \delta(d_{ij}; k)}{\frac{1}{A} \cdot \sum_{i=1}^{A} (w_i - \bar{w})^2}$$

where w_i is any atomic property, $\bar{w}$ is its average value on the molecule, A is the number of atoms, k is the considered lag, d_{ij} is the topological distance between ith and jth atoms, and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δ_k is the number of vertex pairs at distance equal to k [Moran, 1950].

Moran coefficient usually takes value in the interval $[-1, +1]$. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values.

- **Geary coefficient** (c_k)

This is a general index of spatial autocorrelation that, if applied to a molecular graph, can be defined as

$$c_k = \frac{\frac{1}{2\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} (w_i - w_j)^2 \cdot \delta(d_{ij}; k)}{\frac{1}{(A-1)} \cdot \sum_{i=1}^{A} (w_i - \bar{w})^2}$$

where w_i is any atomic property, $\bar{w}$ is its average value on the molecule, A is the number of atoms, k is the lag considered, d_{ij} is the topological distance between ith and jth atoms, and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. Δ_k is the number of vertex pairs at distance equal to k [Geary, 1954].

Geary coefficient is a distance-type function varying from zero to infinite. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates in values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference "no correlation" is $c_k = 1$.

Table A4 Some autocorrelation descriptors for the data set of phenethylamines (Appendix C – Set 2).

Mol.	X	Y	ATS_1	ATS_2	ATS_3	ATS_4	I_1	I_2	I_3	I_4	c_1	c_2	c_3	c_4
1	H	H	2.952	3.313	3.473	3.554	−0.006	−0.056	−0.139	−0.319	0.504	0.804	1.364	2.193
2	H	F	3.032	3.422	3.567	3.598	−0.006	−0.055	−0.134	−0.322	0.512	0.782	1.299	2.198
3	H	Cl	3.096	3.508	3.641	3.635	−0.006	−0.077	−0.177	−0.368	0.531	0.811	1.306	2.114
4	H	Br	3.251	3.708	3.819	3.728	−0.005	−0.098	−0.203	−0.320	0.549	0.838	1.174	1.485
5	H	I	3.392	3.884	3.977	3.818	−0.004	−0.090	−0.174	−0.223	0.542	0.828	1.053	1.042
6	H	Me	3.003	3.383	3.533	3.582	−0.005	−0.045	−0.112	−0.290	0.503	0.768	1.280	2.176
7	F	H	3.032	3.422	3.567	3.640	−0.006	−0.055	−0.134	−0.299	0.512	0.782	1.299	2.034
8	Cl	H	3.096	3.508	3.641	3.710	−0.006	−0.077	−0.177	−0.366	0.531	0.811	1.306	2.008
9	Br	H	3.251	3.708	3.819	3.876	−0.005	−0.098	−0.203	−0.374	0.549	0.838	1.174	1.645
10	I	H	3.392	3.884	3.977	4.027	−0.004	−0.090	−0.174	−0.300	0.542	0.828	1.053	1.363
11	Me	H	3.003	3.383	3.533	3.609	−0.005	−0.045	−0.112	−0.261	0.503	0.768	1.280	2.009
12	Cl	F	3.165	3.598	3.828	3.748	−0.006	−0.073	−0.159	−0.365	0.539	0.793	1.208	2.025
13	Br	F	3.310	3.783	4.081	3.909	−0.005	−0.089	−0.194	−0.364	0.554	0.816	1.235	1.655
14	Me	F	3.079	3.485	3.663	3.651	−0.005	−0.045	−0.106	−0.266	0.511	0.752	1.168	2.025
15	Cl	Cl	3.221	3.671	3.966	3.780	−0.007	−0.095	−0.159	−0.403	0.561	0.825	1.201	1.969
16	Br	Cl	3.359	3.843	4.264	3.936	−0.006	−0.109	−0.141	−0.398	0.574	0.845	1.180	1.655
17	Me	Cl	3.140	3.566	3.763	3.686	−0.006	−0.063	−0.157	−0.302	0.529	0.778	1.210	1.942
18	Cl	Br	3.359	3.843	4.264	3.861	−0.006	−0.109	−0.141	−0.333	0.574	0.845	1.180	1.417
19	Br	Br	3.480	3.991	4.636	4.006	−0.005	−0.130	−0.029	−0.372	0.594	0.875	1.069	1.386
20	Me	Br	3.289	3.756	3.993	3.775	−0.005	−0.080	−0.214	−0.257	0.545	0.802	1.257	1.361
21	Me	Me	3.052	3.449	3.617	3.636	−0.005	−0.037	−0.083	−0.241	0.503	0.740	1.149	2.008
22	Br	Me	3.289	3.756	3.993	3.897	−0.005	−0.080	−0.214	−0.342	0.545	0.802	1.257	1.633

ATS, Moreau–Broto autocorrelations; *I*, Moran coefficient; *c*, Geary coefficient. Calculations are based on the carbon-scaled atomic mass as the weighting scheme for atoms (see Table A3).

• **Auto-Cross-Covariance transforms** (≡ *ACC transforms*)
These are autocovariances and cross-covariances calculated from sequential data with the aim of transforming them into → *uniform-length descriptors* suitable for QSAR modeling. ACC transforms were originally proposed to describe peptide sequences [Wold, Jonsson *et al.*, 1993; Sjöström, Rännar *et al.*, 1995; Andersson, Sjöström *et al.*, 1998; Nyström, Andersson *et al.*, 2000]. To calculate ACC transforms, each amino acid position in the peptide sequence is defined in terms of three orthogonal → z-*scores*, derived from a → *Principal Component Analysis* (PCA) of 29 physico-chemical properties of the 20 coded amino acids.

Then, for each peptide sequence, auto- and cross-covariances with lags $k = 1, 2, \ldots, K$, are calculated as

$$\mathrm{ACC}_k(j,j) = \sum_{i=1}^{n-k} \frac{z_i(j) \cdot z_{i+k}(j)}{n-k} \qquad \mathrm{ACC}_k(j,m) = \sum_{i=1}^{n-k} \frac{z_i(j) \cdot z_{i+k}(m)}{n-k}$$

where j and m indicate two different z-scores, n is the number of amino acids in the sequence, and index i refers to amino acid position in the sequence. z-score values, being derived from PCA, are used directly because they are already mean centered.

ACC transforms were also used to encode information contained into → *CoMFA fields* (steric and electrostatic fields) using as the lag the distance between grid points along each coordinate axis, along the diagonal, or along any intermediate direction. The cross-correlation terms were calculated by the products of the → *interaction energy values* for steric and electrostatic fields in grid points at distances equal to the lag. Different kinds of interactions, namely, positive–positive, negative–negative, and positive–negative, were kept separated, thus resulting in 10 ACC terms for each lag. The major drawback of these ACC transforms is that their values depend on molecule orientation along the axes [Clementi, Cruciani *et al.*, 1993b; van de Waterbeemd, Clementi *et al.*, 1993].

• **TMACC descriptors** (≡ *Topological MAximum Cross-Correlation descriptors*)
These are cross-correlation descriptors [Melville and Hirts, 2007] calculated by taking into account the topological distance d_{ij} between atoms i and j and four basic atomic properties: (1) Gasteiger–Marsili partial charges, accounting for electrostatic properties [Gasteiger and Marsili, 1980]; (2) Wildman–Crippen molar refractivity parameters, accounting for steric properties and polarizabilities [Wildman and Crippen, 1999]; (3) Wildman–Crippen log P parameters, accounting for hydrophobicity [Wildman and Crippen, 1999]; and (4) log S parameters, accounting for solubility and solvation phenomena [Hou, Xia *et al.*, 2004].

The general formula for the calculation of *TMACC* descriptors is

$$TMACC(\mathrm{P}, \mathrm{P}'; k) = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \mathrm{P}_i \cdot \mathrm{P}'_j \cdot \delta(d_{ij}; k)$$

where P and P′ are two atomic properties, A is the number of atoms in the molecule, k is the lag, d_{ij} is the topological distance between ith and jth atoms, Δ_k is the number of atom pairs located at topological distance k, and $\delta(d_{ij}; k)$ is the Kronecker delta equal to 1 if $d_{ij} = k$, zero otherwise. If only one property is considered, that is, $\mathrm{P} = \mathrm{P}'$, autocorrelations are obtained.

Moreover, because all the selected properties, except for molar refractivity, contain both positive and negative values, these are treated as different properties and cross-correlation terms are also calculated between positive and negative values of each property. Therefore, 7 autocorrelation terms and 12 cross-correlation terms constitute the final *TMACC* descriptor vector.

• **DZ$^{\mathrm{K}}$ descriptors**
These are a modification of the Moreau–Broto autocorrelation descriptors defined by using the topological distance in conjunction with the properties of the atoms [Zakarya, Nohair *et al.*, 2000]:

$$\mathrm{DZ}^{\mathrm{K}}_w = \sum_{k=1}^{K} \left[k \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} (w_i \cdot w_j)^{\alpha} \cdot \delta(d_{ij}; k) \right]$$

where w is the selected atomic property, K the maximum considered topological distance, and α an exponent taking values 1 or 0.5. In particular, for $\alpha = 1$, the following expression holds:

$$DZ_w^K = \sum_{k=1}^{K} k \cdot ATS_k(w)$$

where ATS_k is the Moreau–Broto autocorrelation relative to lag k.

The use of atomic properties such as atom connectivity, electronegativity, van der Waals volume, and molar refraction was suggested.

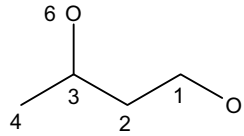
An extended form has been also proposed defined as

$${}^{e}DZ_w^K = \sum_{i=1}^{A} w_i + \sum_{k=1}^{K} \left[k \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} (w_i \cdot w_j)^{\alpha} \cdot \delta(d_{ij}; k) \right] = \sum_{i=1}^{A} w_i + DZ_w^K$$

where the sum of the atomic properties is added to the autocorrelation term.

Example A8

DZ^k autocorrelation descriptors from the H-depleted molecular graph of 4-hydroxy-2-butanone.



$DZ_m^1 = 1 \cdot ATS_1 = 816$
$DZ_m^2 = 1 \cdot ATS_1 + 2 \cdot ATS_2 = 816 + 2 \cdot 864 = 2544$
$DZ_m^3 = 1 \cdot ATS_1 + 2 \cdot ATS_2 + 3 \cdot ATS_3 = 816 + 2 \cdot 864 + 3 \cdot 528 = 4128$

$w_i = m_i$ atomic masses $\alpha = 1$

- **Molecular Electronegativity Edge Vector** (*VMEE*)

This is a modification of the Moreau–Broto autocorrelation defined by using reciprocal topological distances in conjunction with the Pauling atom electronegativities [Li, Fu *et al.*, 2001]. The autocorrelation value for each kth lag is calculated as

$$VMEE_k \equiv v_k = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{\chi_i^{PA} \cdot \chi_j^{PA}}{d_{ij}} \cdot \delta(d_{ij}; k), \quad k = 1, 2, 3, \ldots$$

where χ^{PA} is the atomic electronegativity and d_{ij} is the topological distance between ith and jth atoms. This autocorrelation vector was used in modeling biological activities of dipeptides.

- **3D topological distance-based descriptors** (S_k, X_k, I_k)

These are autocorrelation descriptors contemporarily based on topological and geometric distances, also called **3D TDB descriptors** [Klein, Kaiser *et al.*, 2004].

For each kth lag, steric descriptors, namely, **TDB-steric descriptors** S, are defined as

$$S_k = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} (R_i^{cov} \cdot r_{ij} \cdot R_j^{cov}) \cdot \delta(d_{ij}; k)$$

where Δ_k is the number of atom pairs located at a topological distance d_{ij} equal to k, r_{ij} is the geometric distance between ith and jth atoms, R^{cov} is the covalent radius of the atoms and δ is the Kronecker delta, which is equal to 1 when d_{ij} is equal to k and zero otherwise. In a similar way, electronic descriptors, namely, **TDB-electronic descriptors** X, are defined as

$$X_k = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} (\chi_i \cdot r_{ij} \cdot \chi_j) \cdot \delta(d_{ij}; k)$$

where χ is the sigma orbital electronegativity.

Together with steric and electronic descriptors, atom-type autocorrelation descriptors, namely, **TDB-atom type descriptors** I, are defined as

$$I_k(u,u) \equiv ATAC_k(u,u) = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} \delta_{ij}(u,u) \cdot \delta(d_{ij}; k)$$

where u denotes an atom type and $\delta_{ij}(u, u)$ is a Kronecker delta equal to 1 if both atoms i and j are of type u. These atom-type autocorrelations are calculated only for pairs of atoms of the same type. Moreover, unlike the previous two TDB descriptors (S_k and X_k), this autocorrelation descriptor does not account for 3D information.

- **Atomic Environment Autocorrelations** (AEA)

Aimed at characterizing the local environment of atoms, these descriptors are calculated by applying the autocorrelation function to encode spatial information relative to each single ith atom in a molecule as [Nohair, Zakarya *et al.*, 2002; Nohair and Zakarya, 2003]

$$AEA_{ik} = \sum_{j=1}^{A} (w_i \cdot w_j)^{\alpha} \cdot \delta(d_{ij}; k)$$

where w is any atomic property, α an adjustable parameter, and δ is the Kronecker δ function, which is equal to 1 when the topological distance d_{ij} between focused ith atom and any jth neighbor atom is equal to k. Atom connectivity, atomic van der Waals volume, and surface are among the suggested properties. Moreover, to take properties of the atoms along the i–j path of a topological distance d_{ij} equal to k also into account, a modified autocorrelation function was proposed as

$$AEA'_{ik} = \sum_{j=1}^{A} \left[w_i \cdot \left(\sum\nolimits_m w_m \right)_{ij} \cdot w_j \right]^{1/(k+1)} \cdot \delta(d_{ij}; k)$$

where the exponent is the reciprocal of the number of atoms along the shortest path connecting vertices i and j; w_i and w_j are properties of the two terminal vertices of the path, whereas w_m is the property of a vertex along the path.

- **Autocorrelation of Molecular Surface Properties** ($AMSP$)

This is a general approach for the description of property measures on the molecular surface by using uniform-length descriptors that consist of the same number of elements regardless of the

size of the molecule [Gasteiger, 2003a; Sadowski, Wagener *et al.*, 1995; Wagener, Sadowski *et al.*, 1995].

First, a number of points are randomly distributed on the molecular surface with a user-defined density and in an orderly manner to ensure a continuous surface. Then, the **Surface Autocorrelation Vector** (SAV) is derived by calculating for each lag k the sum of the products of the property values at two surface points located at a distance falling into the kth distance interval. This value is then normalized by the number Δ_k of the geometrical distances r_{ij} in the interval:

$$A(k) = \frac{1}{\Delta_k} \cdot \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} w_i \cdot w_j \cdot \delta(r_{ij}; k)$$

where N is the number of surface points and k represents a distance interval defined by a lower and upper bound.

It was demonstrated that to obtain the best surface autocorrelation vectors for QSAR modeling, the van der Waals surface is better than other molecular surfaces. Then, surface should have no fewer than five grid points per Å^2, and a distance interval not more than 1 Å should be used in the distance binning scheme (Figure A1).

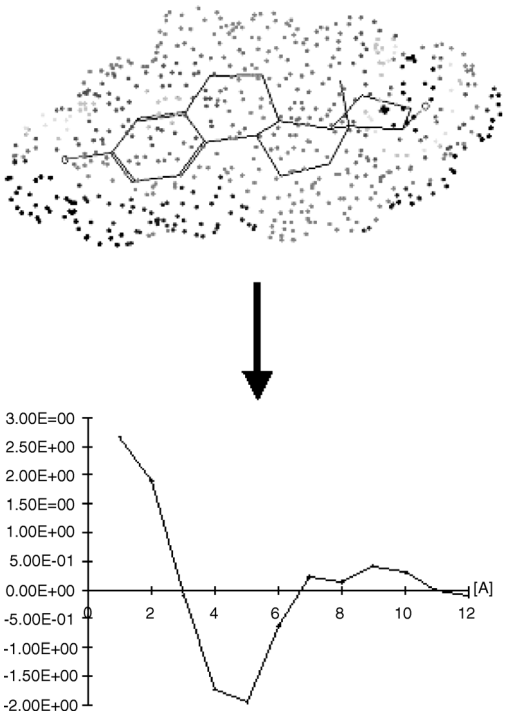


Figure A1 Surface autocorrelation vector of estradiol calculated by using MEP as the surface property.

📖 [Chastrette, Zakarya *et al.*, 1986; Devillers, Chambon *et al.*, 1986; Grassy and Lahana, 1993; Zakarya, Tiyal *et al.*, 1993; Clementi, Cruciani *et al.*, 1993b; van de Waterbeemd, Clementi *et al.*, 1993; Blin, Federici *et al.*, 1995; Sadowski, Wagener *et al.*, 1995; Bauknecht, Zell *et al.*, 1996; Patterson, Cramer III *et al.*, 1996; Huang, Song *et al.*, 1997; Anzali, Gasteiger *et al.*, 1998a; Legendre and Legendre, 1998; Devillers, 2000; Gancia, Bravi *et al.*, 2000; Gasteiger, 2003a; Moon, Song *et al.*, 2003; Cruciani, Baroni *et al.*, 2004]

- **Autocorrelation of a Topological Structure** ≡ *Moreau–Broto autocorrelation* → autocorrelation descriptors
- **Autocorrelation of Molecular Surface Properties** → autocorrelation descriptors
- **autocorrelogram** → autocorrelation descriptors
- **Auto-Cross-Covariance transforms** → autocorrelation descriptors
- **autoignition temperature** → physico-chemical properties (⊙ flash point)
- **autometricity class** → topological information indices (⊙ autometricity index)
- **autometricity index** → topological information indices
- **automorphism group** → graph
- **Avalon fingerprints** → substructure descriptors (⊙ fingerprints)
- **average atom charge density** → quantum-chemical descriptors
- **average atom eccentricity** → distance matrix
- **average binding energy** → scoring functions
- **average bond charge density** → quantum-chemical descriptors
- **average cyclicity index** → detour matrix
- **average distance between pairs of bases** → biodescriptors (⊙ DNA sequences)
- **average distance/distance degree** → molecular geometry
- **average distance sum connectivity** ≡ *Balaban distance connectivity index*
- **average electrophilic superdelocalizability** → quantum-chemical descriptors (⊙ electrophilic superdelocalizability)
- **average Fukui function** → quantum-chemical descriptors (⊙ Fukui functions)
- **average geometric distance degree** → molecular geometry
- **average graph distance degree** → distance matrix
- **average information content based on center** → centric indices
- **average local ionization energy** → quantum-chemical descriptors (⊙ electron density)
- **average molecular weight** → physico-chemical properties (⊙ molecular weight)
- **average nucleophilic superdelocalizability** → quantum-chemical descriptors (⊙ nucleophilic superdelocalizability)

■ **average quasivalence number** (AQVN)

It is a molecular descriptor calculated as average of the atomic numbers Z of the molecule atoms as [Veljković, Mouscadet *et al.*, 2007]

$$Z^{*} \equiv \mathrm{AQVN} = \frac{\sum_{i=1}^{A} Z_i}{A}$$

where A is the number of atoms. It is used in the definition of the **electron-ion interaction potential** (EIIP), proposed to estimate long-range properties of biological molecules [Veljković, 1980] and defined as

$$\mathrm{EIIP} = \frac{0.25 \cdot Z^* \cdot \sin(1.04 \cdot \pi \cdot Z^*)}{2\pi}$$

where Z^* is the average quasivalence number. Moreover, the ratio EIIP/AQVN was proposed as a → *drug-like index* for compounds.

- **average radical superdelocalizability** → quantum-chemical descriptors (⊙ radical superdelocalizability)
- **average row sum of the influence/distance matrix** → GETAWAY descriptors
- **average span** → size descriptors (⊙ span)
- **average vertex distance degree** → Balaban distance connectivity index
- **average writhe** → polymer descriptors
- ***A* weighting scheme** → weighting schemes
- ***AH* weighting scheme** → weighted matrices (⊙ weighted distance matrices)
- **AZV descriptors** → MPR approach
- **azzoo similarity coefficient** → similarity/diversity