# P

- **PAD descriptors** ≡ *PEST Autocorrelation Descriptors* → TAE descriptor methodology
- **Padmakar–Ivan index** ≡ *PI index* → Szeged matrix
- **pair correlation cutoff selection** → variable reduction
- **Palm steric constant** → steric descriptors (⊙ Taft steric constant)
- **parachor** → physico-chemical properties
- **Para-Delocalization Index** → delocalization degree indices (⊙ Delocalization Index)
- **partial atomic charge** → quantum-chemical descriptors
- **partial charge weighted topological electronic index** → charge descriptors
- **partial equalization of orbital electronegativities** → electronegativity
- **partial local invariant** → iterated line graph sequence
- **partial negative surface area** → charged partial surface area descriptors
- **partial-order ranking methods** → chemometrics (⊙ ranking methods)
- **partial positive surface area** → charged partial surface area descriptors
- **partial Wiener indices** → Wiener index
- **partition-based methods** ≡ *cell-based methods*
- **partition coefficients** → physico-chemical properties
- **Pasaréti index** ≡ *all-path Wiener index* → path counts

■ **PASS** (≡ *Prediction of Activity Spectra of Substances*)
The computer system PASS was built to predict several hundreds of biological activities (main and side pharmacological activities, → *mode of action*, mutagenicity, carcinogenicity, teratogenicity, and embryotoxicity) [Filimonov and Poroikov, 1996, 2001; Poroikov, Filimonov *et al.*, 2000, 2003; Anzali, Barnickel *et al.*, 2001]. Most of the biological active compounds reveal a wide spectrum of different effects. Some of them are useful in the treatment of defined diseases, while others cause various side and toxic effects. The whole complex of activities caused by the compounds is called "biological activity spectrum of the substance." This spectrum is defined as the intrinsic property of a compound depending only on its molecular structure and physico-chemical characteristics.

PASS was trained on more than 30 000 compounds that reveal more than 500 kinds of different biological activities. The molecular descriptors used by PASS are → *MNA descriptors*.

- **path** → graph
- **path-Cluj matrices** → Cluj matrices
- **path-cluster subgraph** → molecular graph

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

ISBN: 978-3-527-31852-0

- **path connectivity** → weighted matrices (⊙ weighted distance matrices)
- **path count** ≡ *molecular path count* → path counts

## path counts (≡ *path numbers*)

Path counts are atomic and molecular descriptors obtained from a → *H-depleted molecular graph G*, based on the counting of graph → *paths*. Analogous to the → *atomic walk count*, the **atomic path count** (or **atomic path number**) ${}^{m}P_i$ is a → *local vertex invariant* encoding the atomic environment, defined as the number of paths of length *m* starting from the *i*th vertex to any other vertex in the graph. The length *m* of the path, that is, the number of edges along the path, is called **path order** [Randić, Brissey *et al.*, 1979; Randić and Wilkins, 1979b; Randić, 1979].

The **vertex path code** (or **Randić atomic path code**) of the *i*th vertex is the ordered sequence of atomic path counts, with respect to the path length:

$$\{{}^{1}P_i, {}^{2}P_i, \ldots, {}^{L}P_i\}$$

where $L = {}^{\Delta}\eta_i$ is the → *atom detour eccentricity* of the *i*th vertex, that is, the length of the longest path starting from the vertex $v_i$; it can be derived from the → *detour matrix* as the maximum value entry in the *i*th row. The atomic path count of first order ${}^{1}P_i$ is the → *vertex degree* $\delta_i$, while the atomic path count of zero order ${}^{0}P_i$ is always equal to 1. Vertex path codes for all nonhydrogen atoms in the molecule can be collected into a rectangular matrix that has been called → *path-sequence matrix* **SP**. The sum of all the elements in the vertex path code is the total number of paths of any length starting from the considered vertex and is called **atomic path count sum** $P_i$:

$$P_i = \sum_{m=1}^{L} {}^{m}P_i$$

The **molecular path count**, also called **path count**, **molecular path number** or **topological bond index** with the symbol $K_m$, is the total number of paths of length *m* in the graph and is denoted by ${}^{m}P$ ($m = 0,1,\ldots, L$), where *L* is the length of the longest path in the graph. ${}^{0}P$ coincides with the number *A* of graph vertices, ${}^{1}P$ with the number *B* of graph edges, ${}^{2}P$ with the → *connection number* $N_2$, that is, the number of two contiguous edges.

The molecular path count of order *m* is calculated by adding the corresponding atomic path counts of all vertices, then dividing by 2 since each path has been counted twice:

$${}^{m}P = \frac{1}{2} \cdot \sum_{i=1}^{A} {}^{m}P_i \quad m \neq 0$$

The path count ${}^{0}P$ is simply equal to *A*.

The **molecular path code** of the graph is the ordered sequence of molecular path counts:

$$\{{}^{0}P, {}^{1}P, {}^{2}P, \ldots, {}^{L}P\}$$

Molecular path codes are → *vectorial descriptors*, used, for example, to search for similarities among molecules, by choosing a suitable value for the maximum length *L* with respect to the set of studied molecules to obtain → *uniform-length descriptors*.

It is noteworthy that, for acyclic graphs, the molecular path code coincides with the → *graph distance code*.

Summing up all the elements of the molecular path code gives the **total path count** $P$ (also called **total path number**):

$$P = \sum_{m=0}^{L} {}^{m}P = A + \frac{1}{2} \cdot \sum_{m=1}^{L} \sum_{i=1}^{A} {}^{m}P_i = A + \frac{1}{2} \cdot \sum_{i=1}^{A} P_i$$

This descriptor is considered a quantitative measure of → *molecular complexity*.

For acyclic graphs, the total path count is simply calculated from the number $A$ of graph vertices as

$$P = \frac{A^2 + A}{2}$$

For simple structures, the path counts can be derived directly from the molecular graphs; otherwise specific algorithms are needed. For example, Randić's algorithm results [Randić, Brissey *et al.*, 1979] in path counts for nonequivalent vertices from the → *adjacency matrix*.

**Table P1** Outline of a generic path-sequence matrix with row and column sums.

| | Path length, *m* | | | | | |
|---|---|---|---|---|---|---|
| **Atom ID** | **0** | **1** | **2** | ... | ***L*** | **Atomic path count sums** |
| **1** | 1 | ${}^{1}P_1$ | ${}^{2}P_1$ | ... | ${}^{L}P_1$ | $P_1$ |
| **2** | 1 | ${}^{1}P_2$ | ${}^{2}P_2$ | ... | ${}^{L}P_2$ | $P_2$ |
| ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |
| ***A*** | 1 | ${}^{1}P_A$ | ${}^{2}P_A$ | ... | ${}^{L}P_A$ | $P_A$ |
| **Molecular path counts** | $A$ | ${}^{1}P$ | ${}^{2}P$ | ... | ${}^{L}P$ | $P$ |

The → *P matrix* is a graph representation of molecules based on the total path count.

Five **path count-based indices** were proposed by Balaban, defined as [Balaban, Beteringhe *et al.*, 2007]

$$\mathrm{Q} = \sum_{m=1}^{L} \frac{{}^{m}P^2}{(C+1)} \qquad \mathrm{S} = \sum_{m=1}^{L} \frac{{}^{m}P^{1/2}}{(C+1)} \qquad \mathrm{D} = \sum_{m=1}^{L} \frac{{}^{m}P^{1/2}}{m \cdot (C+1)}$$

$$\mathrm{A} = \sum_{m=1}^{L} \frac{{}^{m}P}{m \cdot (C+1)} \qquad \mathrm{P} = \sum_{m=1}^{L} \frac{{}^{m}P^{1/2}}{m^{1/2} \cdot (C+1)}$$

where $C$ is the → *cyclomatic number* and the summations run over the increasing path lengths.

The index Q increases with the molecule size and branching, whereas index S increases with size but decreases with branching; index D increases with cyclicity and decreases with branching. For acyclic graphs, index A is the → *Harary index* (denoted as $H$ by Trinajstić and $RDSUM$ by Balaban). Finally, index P increases with size and decreases with branching; for hydrocarbons, this index shows the minimum number of degeneracies with respect to the other path count-based indices.

**Atom-type path counts** ${}^{m}P_{\mathrm{X}}$ are defined as the number of paths originating from all the atoms of a given type. For example, the number of paths of length 3 originating from oxygen atoms in a molecule was used to predict boiling points of alcohols [Randić and Basak, 2001a].

To take into account multiple bonds and heteroatoms, **weighted path counts** can be calculated, either by introducing the weighting factors after the paths have been enumerated or by computing the weighted paths directly [Randić and Basak, 1999]. The sums of path weights obtained by applying different → *weighting schemes* to the graph edges are known as the → *ID numbers*; the most common weighting schemes are based on → *bond order indices*. Moreover, the **WTPT index** was proposed as the sum of the weights of all the paths starting from heteroatoms in the molecule [Bakken and Jurs, 1999b]; it is closely related to path counts of heteroatoms used in → *start-end vectors*.

→ *Variable path counts* are obtained by weighting the graph edges involving heteroatoms with one or more variable parameters [Amić, Basak *et al.*, 2002].

**Valence shell counts** or **graph valence shells**, denoted by ${}^{m}S$, are weighted path counts calculated by adding valence shells at the same separation $m$ for all atoms in a molecule [Randić, 2001b]. The concept of valence shell is similar to the concept of atomic path count; the difference is that instead of counting for each atom the number of neighbors at increasing length, one adds the → *vertex degree* of neighbors at increasing separation. The **valence shell** of order $m$ for the $i$th vertex is then defined as

$$ {}^{m}S_i = \sum_{j=1}^{A} \sum\nolimits_{p_{ij}} \delta_j \cdot \delta(|p_{ij}|; m) $$

where $\delta_j$ is the vertex degree of the $j$th atom, $|p_{ij}|$ is the length of a path connecting vertices $v_i$ and $v_j$, and $\delta(|p_{ij}|; m)$ the Dirac delta function equal to 1 when the length of the path $p_{ij}$ is equal to $m$, and zero otherwise. The first summation goes over all vertices in the graph, while the second one over all the paths connecting two vertices $v_i$ and $v_j$. A shell of order zero represents the vertex degree of a vertex, while a shell of order one represents the → *extended connectivity* of the vertex. The valence shell of a vertex can be viewed as the count of weighted paths starting from the vertex, where the weights are determined by the vertex degree of the other terminal vertex of the path. For acyclic graphs, the valence shells for the $i$th vertex reduce to the elements $lb_{im}$ in the $i$th row of the → *branching layer matrix* and are defined as the following [Lukovits, 2001a]:

$$ {}^{m}S_i \equiv lb_{im} = \sum_{j=1}^{A} \delta_j \cdot \delta(d_{ij}; m) $$

where $d_{ij}$ is the topological distance between vertices $v_i$ and $v_j$.

Then, the molecular valence shell count of $m$th order is calculated as

$$ {}^{m}S = \frac{1}{2} \cdot \sum_{i=1}^{A} {}^{m}S_i \quad m \neq 0 $$

Based on the length of the paths in the molecular graph, other local vertex invariants and molecular descriptors have been proposed.

The **path degree** or **vertex path sum**, is a local invariant, denoted by $\xi_i$ and defined as the sum of the lengths $m$ of all paths starting from vertex $v_i$:

$$ \xi_i = \sum_{m=1}^{L} {}^{m}P_i \cdot m $$

where $L = {}^{\Delta}\eta_i$ is the atom detour eccentricity of the $i$th vertex, that is, the length of the longest path starting from $v_i$ and ${}^{m}P_i$ is the number of paths of length $m$ from $v_i$. For acyclic graphs, the path degree $\xi_i$ coincides with the → *vertex distance degree* $\sigma_i$. Moreover, the path degrees are used as the weighting scheme for vertices to generate the → *path degree layer matrix* **LPD**.

By summing up path degrees over all vertices in the graph, the **all-path Wiener index** $W^{AP}$ (or **Pasaréti index**) is derived. This is a molecular descriptor proposed as a variant of the → *Wiener index* but with more discriminating power among cycle-containing structures, defined as [Lukovits, 1998a; Lukovits and Linert, 1998]

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^{A} \xi_i = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \sum\nolimits_{p_{ij}} |p_{ij}|$$

where the two outer summations on the right side run over all pairs of vertices in the graph and the inner summation runs over all paths $p_{ij}$ between the vertices $v_i$ and $v_j$; $|p_{ij}|$ denotes the length of the considered path. Its maximum value is equal to $A^2 \times (A-1) \times 2^{(A-4)}$ for a → *complete graph* with $A$ vertices.

It has to be noted that the all-path Wiener index coincides with a previously proposed global index obtained as the half-sum of any row of the path degree layer matrix **LPD**.

The all-path Wiener index can be calculated more easily from the **all-path matrix** $\mathbf{\Omega}^{AP}$ that is a square symmetric $A \times A$ matrix, $A$ being the number of graph vertices, whose $i$–$j$ entry is the sum of the lengths of all the paths $p_{ij}$ connecting vertices $v_i$ and $v_j$:

$$[\mathbf{\Omega}^{AP}]_{ij} = \begin{cases} \sum |p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $|p_{ij}|$ denotes the length of a path between $v_i$ and $v_j$. Diagonal elements are equal to zero by definition. → *Distance matrix* **D**, → *detour matrix* $\mathbf{\Delta}$ and → *detour distance – topological distance combined matrix* $\mathbf{\Delta} \wedge \mathbf{D}$ are closely related to the all-path matrix as they are based on the length of the shortest, longest, and longest plus shortest paths between any two vertices in the graph, respectively. It must be noted that for acyclic graphs all these matrices coincide, there being a unique path between two vertices.

The row sums of the all-path matrix are the path degrees $\xi_i$:

$$\xi_i \equiv VS_i(\mathbf{\Omega}^{AP}) = \sum_{j=1}^{A} [\mathbf{\Omega}^{AP}]_{ij}$$

where $VS_i$ indicates the → *vertex sum operator*.

The all-path Wiener index is then derived from the all-path matrix as the following:

$$W^{AP} \equiv Wi(\mathbf{\Omega}^{AP}) = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} [\mathbf{\Omega}^{AP}]_{ij}$$

where $Wi$ is the → *Wiener operator*.

Because the all-path Wiener index increases exponentially with the number $A$ of graph vertices, it was proposed [Lukovits, 1998a] to divide it by the average number $k$ of paths between vertices and the resulting quantity was called **Vérhalom index**:

$$\overline{W}^{AP} = W^{AP}/k$$

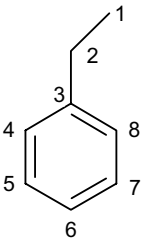
where $k$ is obtained by the ratio of the total number of paths (of order greater than zero) over the number of vertex pairs $A \times (A-1)/2$, where $A$ is the number of vertices in the graph:

$$k = \frac{2 \cdot P}{A \cdot (A-1)}$$

For simple cycles, $k = 2$.

**Example P1**

Path-sequence matrix **SP**, all-path matrix $\mathbf{\Omega}^{AP}$, path degrees $\xi_i$, and related indices for ethylbenzene.




**SP** =

| Atom | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | $P_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 12 |
| 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 12 |
| 3 | 1 | 3 | 3 | 2 | 2 | 2 | 0 | 0 | 12 |
| 4 | 1 | 2 | 3 | 3 | 2 | 2 | 1 | 1 | 14 |
| 5 | 1 | 2 | 2 | 3 | 3 | 3 | 1 | 0 | 14 |
| 6 | 1 | 2 | 2 | 2 | 4 | 4 | 0 | 0 | 14 |
| 7 | 1 | 2 | 2 | 3 | 3 | 3 | 1 | 0 | 14 |
| 8 | 1 | 2 | 3 | 3 | 2 | 2 | 1 | 1 | 14 |
| ${}^{m}P$ | 8 | 8 | 9 | 10 | 10 | 10 | 4 | 2 | 106 |

$$P = \sum_{m=0}^{7} {}^{m}P = 8 + \frac{1}{2} \cdot \sum_{i=1}^{8} P_i = 53$$

$$Q = \sum_{m=1}^{7} \frac{{}^{m}P^2}{(1+1)} = 232.5$$

$$S = \sum_{m=1}^{7} \frac{{}^{m}P^{1/2}}{(1+1)} = 9.365$$

$$D = \sum_{m=1}^{7} \frac{{}^{m}P^{1/2}}{m \cdot (1+1)} = 3.670$$

$$A = \sum_{m=1}^{7} \frac{{}^{m}P}{m \cdot (1+1)} = 10.643$$

$$P = \sum_{m=1}^{7} \frac{{}^{m}P^{1/2}}{m^{1/2} \cdot (1+1)} = 5.561$$

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^{8} \xi_i = 184$$

$$W^{AP} = \frac{1}{2} \cdot \sum_{i=1}^{8} \sum_{j=1}^{8} [\mathbf{\Omega}^{AP}]_{ij} = 184$$

$$\overline{W}^{AP} = \frac{W^{AP}}{k} = \frac{184}{53/28} = 97.2$$

$$\xi_1 = 1 \times 0 + 1 \times 1 + 1 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 2 \times 6 + 2 \times 7 = 53$$

$$\xi_2 = 1 \times 0 + 2 \times 1 + 2 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 2 \times 6 + 0 \times 7 = 42$$

$$\xi_3 = 1 \times 0 + 3 \times 1 + 3 \times 2 + 2 \times 3 + 2 \times 4 + 2 \times 5 + 0 \times 6 + 0 \times 7 = 33$$

$$\cdots\cdots$$

$$\xi_8 = 1 \times 0 + 2 \times 1 + 3 \times 2 + 3 \times 3 + 2 \times 4 + 2 \times 5 + 1 \times 6 + 1 \times 7 = 48$$

$\mathbf{\Omega}^{AP}$ =

| Atom | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\xi_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 10 | 10 | 10 | 10 | 10 | 53 |
| 2 | 1 | 0 | 1 | 8 | 8 | 8 | 8 | 8 | 42 |
| 3 | 2 | 1 | 0 | 6 | 6 | 6 | 6 | 6 | 33 |
| 4 | 10 | 8 | 6 | 0 | 6 | 6 | 6 | 6 | 48 |
| 5 | 10 | 8 | 6 | 6 | 0 | 6 | 6 | 6 | 48 |
| 6 | 10 | 8 | 6 | 6 | 6 | 0 | 6 | 6 | 48 |
| 7 | 10 | 8 | 6 | 6 | 6 | 6 | 0 | 6 | 48 |
| 8 | 10 | 8 | 6 | 6 | 6 | 6 | 6 | 0 | 48 |

**Table P2** Molecular path counts for C8 data set (Appendix C – Set 1).

| C8 | $^1P$ | $^2P$ | $^3P$ | $^4P$ | $^5P$ | $^6P$ | $^7P$ | C8 | $^1P$ | $^2P$ | $^3P$ | $^4P$ | $^5P$ | $^6P$ | $^7P$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *n*-Octane | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 33MM | 7 | 9 | 7 | 4 | 1 | 0 | 0 |
| 2M | 7 | 7 | 5 | 4 | 3 | 2 | 0 | 34MM | 7 | 8 | 8 | 4 | 1 | 0 | 0 |
| 3M | 7 | 7 | 6 | 4 | 3 | 1 | 0 | 2M3E | 7 | 8 | 8 | 5 | 0 | 0 | 0 |
| 4M | 7 | 7 | 6 | 5 | 2 | 1 | 0 | 3M3E | 7 | 9 | 9 | 3 | 0 | 0 | 0 |
| 3E | 7 | 7 | 7 | 5 | 2 | 0 | 0 | 223MMM | 7 | 10 | 8 | 3 | 0 | 0 | 0 |
| 22MM | 7 | 9 | 5 | 4 | 3 | 0 | 0 | 224MMM | 7 | 10 | 5 | 6 | 0 | 0 | 0 |
| 23MM | 7 | 8 | 7 | 4 | 2 | 0 | 0 | 233MMM | 7 | 10 | 9 | 2 | 0 | 0 | 0 |
| 24MM | 7 | 8 | 6 | 5 | 2 | 0 | 0 | 234MMM | 7 | 9 | 8 | 4 | 0 | 0 | 0 |
| 25MM | 7 | 8 | 5 | 4 | 4 | 0 | 0 | 2233MMMM | 7 | 12 | 9 | 0 | 0 | 0 | 0 |

📖 [Randić and Wilkins, 1979a, 1979c; Randić, 1980a, 1990a, 1991c, 1992c, 1996b, 1997a; Randić, Brissey *et al.*, 1980; Quintas and Slater, 1981; Wilkins, Randić *et al.*, 1981; Randić, Kraus *et al.*, 1983; Randić, 1984a; Randić, Jerman-Blazic *et al.*, 1987; Kunz, 1989; Randić and Jurs, 1989; Clerc and Terkovics, 1990; Hall, Kier *et al.*, 1993; Hall, Dailey *et al.*, 1993; Kier, Hall *et al.*, 1993; Pisanski and Žerovnik, 1994; Plavšić, Šoškić *et al.*, 1996b; Amić, Lučić *et al.*, 2001; Lukovits, Nikolić *et al.*, 2002]

- **path count-based indices** → path counts
- **path degree** → path counts
- **path degree layer matrix** → layer matrices
- **path-distance map matrix** → biodescriptors (⊙ proteomics maps)
- **path-distance-sum-connectivity matrix** → weighted matrices (⊙ weighted distance matrices)
- **path eccentricity** ≡ *atom detour eccentricity* → detour matrix
- **PathFinder fingerprints** → shape descriptors
- **path graph** ≡ *linear graph* → graph
- **path graphical bond order** → bond order indices (⊙ graphical bond order)
- **path-layer matrix** ≡ *path-sequence matrix* → sequence matrices
- **path length** → graph
- **path matrix** → double invariants
- **path matrix** ≡ *P-matrix* → bond order indices (⊙ graphical bond order)
- **path numbers** ≡ *path counts*
- **path order** → path counts
- **path-sequence matrix** → sequence matrices
- **path subgraph** → molecular graph
- **path-Szeged matrices** → Szeged matrices
- **path/walk shape indices** → shape descriptors
- **path-Wiener matrix** → Wiener matrix
- **path-χ matrix** → weighted matrices (⊙ weighted distance matrices)
- **pendent matrix** → superpendentic index
- **Pauling bond number** → delocalization degree indices (⊙ Krygowski bond energy)
- **PCA** ≡ *Principal Component Analysis*

- **PC-based drug-like index** → scoring functions
- **PDQ descriptors** ≡ *Pharmacophore-Derived Query descriptors* → substructure descriptors (⊙ pharmacophore-based descriptors)
- **PDR-FP fingerprints** → cell-based methods
- **PDT fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pearson's correlation coefficient** → statistical indices (⊙ correlation measures)
- **Pearson's first index** → statistical indices (⊙ moment statistical functions)
- **Pearson coefficient** → classification parameters
- **Pearson coefficient** → similarity/diversity
- **PEI** ≡ *Polarizability Effect Index* → electric polarization descriptors
- **PEOE** ≡ *Partial Equalization of Orbital Electronegativities* → electronegativity
- **per(D) index** → algebraic operators (⊙ determinant)

## periphery codes

These are binary molecular codes proposed to characterize the periphery shape of molecules embedded on a 2D hexagonal lattice [Balaban and Harary, 1968; Balaban, 1976b]. They are suitable for the shape characterization of planar benzenoids and annulenes. "Inside" and "Outside" regions of closed curves are indicated by binary labels 1 and 0, respectively, associated with the graph vertices [Randić and Razinger, 1995a, 1995b, 1997]. In other words, digit 1 is associated with movement toward Inside and digit 0 with movement Outside of each ring; a clockwise direction is adopted and the starting point on the periphery is the vertex satisfying the convention of lexicographic minimum. Other different canonical rules can be chosen to define periphery codes [Jerman-Blazic Dzonova and Trinajstić, 1982; Müller, Szymanski *et al.*, 1990a].

Periphery codes can be used to evaluate → *similarity/diversity* based on molecular shape among several compounds [Randić and Razinger, 1995b]. Moreover, periphery codes can also be used to distinguish between *cis*- and *trans*-isomers [Oth and Gilles, 1968; Balaban, 1969, 1997a] and recognize whether a atom molecule is chiral or not [Randić, 1998a]. In particular, for 2D-embedded molecules, the → *Randić chirality index* was proposed by calculating a particular periphery code from left to right and from right to left: if different results are obtained, then the molecule is chiral.

📖 [Balaban, 1971, 1988a; Randić and Mezey, 1996]

- **permanent** → algebraic operators (⊙ determinant)
- **permittivity** ≡ *dielectric constant* → physico-chemical properties
- **persistence** → environmental indices
- **persistence length** → size descriptors
- **perturbation connectivity indices** → connectivity indices
- **perturbation delta value** → vertex degree
- **perturbation geodesic matrices** → weighted matrices (⊙ weighted distance matrices)
- **perturbation graph matrices** → weighted matrices (⊙ weighted distance matrices)
- **Perturbation of an Environment Limited Concentric and Ordered** → DARC/PELCO analysis
- **PEST Autocorrelation Descriptors** → TAE descriptor methodology
- **PEST descriptors** → TAE descriptor methodology

- **Petitjean shape indices** → shape descriptors
- **pfaffian** → algebraic operators (⊙ determinant)
- **pH** → physico-chemical properties
- **pharmacological indices** → biological activity indices
- **pharmacophore** → drug design
- **pharmacophore-based descriptors** → substructure descriptors
- **Pharmacophore Definition Triplets fingerprints** ≡ *PDT fingerprints* → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pharmacophore-Derived Query descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pharmacophore Point Filter** → scoring functions
- **pharmacophore signature** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **PharmPrint descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **phase capacity ratio** ≡ *capacity factor* → chromatographic descriptors

## ■ physico-chemical properties

They constitute the most important class of experimental measurements and play a fundamental role as → *molecular descriptors* both for their availability as well as for their interpretability [Exner, 1966; Lyman, Reehl *et al.*, 1982; Reid, Prausnitz *et al.*, 1988; Horvath, 1992; Abraham, 1993c; Baum, 1997; Lide, 1999; Reinhard and Drefahl, 1999]. Physico-chemical properties are used both as the molecular properties to be correlated with molecular structure in QSPR modeling and as the molecular descriptors when searching for relationships with biological activities. Physico-chemical properties are constitutive parts of → *volume descriptors*, → *electric polarization descriptors*, → *spectra descriptors*, → *chromatographic descriptors*, and so on. Combinations of physico-chemical properties are largely used in the definition of → *environmental indices*. Other important physico-chemical properties are the so-called → *technological properties* useful to characterize materials.

Definitions of some important physico-chemical properties are given below.

• **boiling point** (BP)

Boiling point is the temperature at which the liquid and gas phases of a pure substance are in equilibrium at a specified pressure, that is, the temperature at which the substance changes its state from a liquid to a gas at a given pressure. The **normal boiling point** is the boiling point at normal atmospheric pressure (101.325 kPa). The SI units are Kelvin degrees K, nevertheless the Celsius degrees °C are still very much in use (°C = K − 273.15).

In terms of intermolecular interactions, the boiling point represents the temperature at which molecules possess enough thermal energy to overcome the various intermolecular attractions binding the molecules into the liquid (e.g. hydrogen bonds, dipole–dipole attraction, instantaneous-dipole induced-dipole attractions). Therefore the boiling point is also an index of the strength of intermolecular attractive forces.

The boiling point of a pure compound increases with the increase in the molecule size and molecular branching, with the presence of hydrogen-bonds and dipole–dipole interactions.

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

• **critical constants**

The critical pressure $P_c$, critical volume $V_c$, and critical temperature $T_c$ are the values of the pressure $P$, volume $V_m$, and thermodynamic temperature $T$ at which the densities of coexisting liquid and gaseous phases become identical.

The **critical temperature**, $T_c$, of a substance is the temperature above which distinct liquid and gas phases do not exist, that is, the temperature above which a gas cannot be liquefied by an increase of pressure. As the critical temperature is approached, the properties of the gas and liquid phases become the same resulting in only one phase: the supercritical fluid.

The **critical pressure**, $P_c$, is the vapor pressure at the critical temperature and critical volume.

The **critical volume**, $V_c$, is the volume of a fixed mass of a fluid at critical temperature and pressure.

🕮 [Needham, Wei *et al.*, 1988; Grigoras, 1990; Katritzky, Mu *et al.*, 1998; Turner, Costello *et al.*, 1998; Espinosa, Yaffe *et al.*, 2001; Wakeham, Cholakov *et al.*, 2002; Yao, Wang *et al.*, 2002]

• **density** ($\rho$)

The density of a substance is the mass m per unit volume V. For the common case of a homogeneous substance, it is expressed as

$$\rho = \frac{m}{V}$$

where m is the mass of the substance and V its volume. The SI units are kg $m^{-3}$.

In general, density can be changed by changing either the pressure or the temperature. Increasing the pressure will always increase the density of a material. Increasing the temperature generally decreases the density, but there are notable exceptions to this generalization (e.g., water).

• **dielectric constant** ($\varepsilon$)

The dielectric constant $\varepsilon$, also called **permittivity** and sometimes denoted by $\kappa$, is a measure of the ability of a substance to attenuate the transmission of an electrostatic force from one charged body to another [Karelson, 2001]. The lower the value, the greater the attenuation.

Based on the dielectric constant, the **Kirkwood function** is defined as [Kirkwood and Westheimer, 1938; Reichardt, 1990]

$$K_f = \frac{\varepsilon - 1}{2 \cdot \varepsilon + 1}$$

This function is used to study solvent effects and for classification of solvents. Moreover, the dielectric constant enters the definition of the → *molar refractivity*.

🕮 [Schweitzer and Morris, 1999; Sulea and Purisima, 1999; Sild and Karelson, 2002]

• **dielectric susceptibility** ($\chi^e$)

The dielectric susceptibility $\chi^e$ of a dielectric material is a measure of how easily it polarizes in response to an electric field. This, in turn, determines the electric permittivity of the material

and thus influences many other phenomena in that medium, from the capacitance of capacitors to the speed of light.

It is defined as the constant of proportionality relating the electric field **E** to the induced dielectric polarization density ***P*** such that

$$\boldsymbol{P} = \varepsilon_0 \cdot \chi^e \cdot \mathbf{E}$$

where $\varepsilon_0$ is the electric permittivity in vacuum.

The electric displacement **D** is related to the polarization density ***P*** by

$$\mathbf{D} = \varepsilon_0 \cdot \mathbf{E} + \boldsymbol{P} = \varepsilon_0 \cdot (1 + \chi^e) \cdot \mathbf{E} = \varepsilon_0 \cdot \varepsilon \cdot \mathbf{E}$$

where $\varepsilon$ is the → *dielectric constant* of the medium; the dielectric constant is related to the electric susceptibility as follows:

$$\varepsilon = 1 + \chi^e$$

• **enthalpies** (H)

The **enthalpy** or *heat content* (denoted as H) is a thermodynamic quantity describing the thermodynamic potential of a system, which can be used to calculate the "useful" work obtainable from a closed thermodynamic system under constant pressure.

The **standard reaction enthalpy** ($\Delta H^0$) is the variation of the enthalpy of a chemical reaction relatively to one mole of a specified reagent when both reagents and products are in their standard state (the most stable form of the element at 100 kPa of pressure and the specified temperature, usually 298 K or 25 °C).

Different enthalpies can be defined, depending on the involved thermodynamic process (Table P3) and their values are usually quoted in kJ/mol or kcal/mol or cal/g.

**Table P3** Usual symbols for standard reaction enthalpies.

| Symbol | Reaction | Symbol | Reaction |
|---|---|---|---|
| $\Delta H^0_f$ | Formation | $\Delta H^0_{fus}$ | Fusion |
| $\Delta H^0_c$ | Combustion | $\Delta H^0_{trans}$ | Transition phases |
| $\Delta H^0_{vap}$ | Vaporization | $\Delta H^0_{mix}$ | Mixing of fluids |
| $\Delta H^0_{sub}$ | Sublimation | $\Delta H^0_{ads}$ | Adsorption |

Negative values of standard reaction enthalpies indicate exothermic reactions, whereas positive values indicate endothermic reactions. Together with the → *molar volume*, vaporization enthalpy is used in determining the → *Hildebrand solubility parameter*.

📖 [Exner, 1973; Randić, 1991a; Li and You, 1993a; Pogliani, 1997b; Estrada, Torres *et al.*, 1998; Mercader, Castro *et al.*, 2000; Mercader, Castro *et al.*, 2001; Yao, Zhang *et al.*, 2001; Chickos, Nichols *et al.*, 2002; Puri, Chickos *et al.*, 2002a, 2002b, 2003; Toropov, Toropova *et al.*, 2004; Cao and Gao, 2005; Zhokova, Palyulin *et al.*, 2007]

- **equilibrium constants** ($K$)

The equilibrium constant is dimensionless quantity characterizing a chemical equilibrium in a chemical reaction. It is a useful tool in determining the concentration of various reactants and products in a system where chemical equilibrium occurs.

For example, for the reaction in solution,

$$aA + bB \rightleftarrows cC + dD$$

where $A$ and $B$ are reactant chemical species, $C$ and $D$ are product species, and $a$, $b$, $c$, and $d$ are the stoichiometric coefficients of the respective reactants and products, the equilibrium constant is given by

$$K = \frac{[C]^c \cdot [D]^d}{[A]^a \cdot [B]^b}$$

where $[A]$, $[B]$, $[C]$, and $[D]$ are the concentrations of the species involved in the reaction. A more precise definition is in terms of activity rather than concentration.

Equilibrium constants are often represented by the quantity p$K$ that is the negative logarithm (base 10) of an equilibrium constant $K$: $\mathrm{p}K = -\log_{10}K$.

A **dissociation constant** is a constant whose numerical value depends on the equilibrium between the dissociated and undissociated forms of a molecule. Higher the dissociation constant, greater the dissociation. Examples of dissociation constants are *substrate–enzyme dissociation constant* and the **acid dissociation constant** $\mathrm{p}K_\mathrm{a}$. This latter is defined as

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} + \log_{10}\left(\frac{\mathrm{AH}}{\mathrm{A}^-}\right)$$

where pH is the concentration of $\mathrm{H}^+$ species, AH is the conjugated acid and $\mathrm{A}^-$ the conjugated base ($\mathrm{p}K_\mathrm{a} < 2$ means strong acid; $\mathrm{p}K_\mathrm{a} > 2$ and $\mathrm{p}K_\mathrm{a} < 7$ mean weak acid; $\mathrm{p}K_\mathrm{a} > 7$ and $\mathrm{p}K_\mathrm{a} < 10$ mean weak base; $\mathrm{p}K_\mathrm{a} > 10$ means strong base).

In chemistry and biochemistry, a dissociation constant is a specific type of equilibrium constant that measures the propensity of a larger object to separate (dissociate) reversibly into smaller components, as when a complex falls apart into its component molecules, or when a salt splits up into its component ions. The dissociation constant is often also denoted as $K_\mathrm{d}$ and is the inverse of the *affinity constant*. In the special case of salts, the dissociation constant can also be called *ionization constant*.

The dissociation constant is commonly used in QSAR studies to describe the affinity between a ligand (such as a drug) and a protein, that is, how tightly a ligand binds to a particular protein. Ligand–protein affinities are influenced by noncovalent intermolecular interactions between the two molecules such as hydrogen-bonding, electrostatic interactions, hydrophobic, and Van der Waals forces.

Fundamental thermodynamic equations relate the equilibrium constant to Gibbs (G) free energy, enthalpy (H), and entropy (S):

$$\Delta\mathrm{G}^0 = \Delta\mathrm{H}^0 - T \cdot \Delta\mathrm{S}^0 = -RT \cdot \ln K$$

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

• **flash point** (FP)

The flash point is the temperature at which the vapor above a volatile liquid forms a combustible mixture with air. At the flash point, the application of a naked flame gives a momentary flash rather than continuous combustion, for which the temperature is too low.

At this temperature, the vapor may cease to burn when the source of ignition is removed. However, as the temperature rises still further, the combustible substance reacts with oxygen in the air in an exothermic oxidation process.

Closely related to the flash point, the **autoignition temperature** is defined as the lowest temperature at which a substance in air will ignite in the absence of a spark or flame. Autoignition occurs when the rate of heat evolved is greater than the rate at which heat is lost to the surroundings.

Flash point and autoignition temperature are → *technological properties* of compounds and important safety parameters [Katritzky, Maran *et al.*, 2000], often used as one descriptive characteristic of liquid fuel, but also used to describe liquids that are not used intentionally as fuels.

QSPR studies on flash points and autoignition temperatures are [Egolf and Jurs, 1992; Murugan, Grendze *et al.*, 1994; Katritzky, Lobanov *et al.*, 1996; Tetteh, Metcalfe *et al.*, 1996; Mitchell and Jurs, 1997; Tetteh, Suzuki *et al.*, 1999; Katritzky, Petrukhin *et al.*, 2001a; Stefanis, Constantinou *et al.*, 2004].

A → *group contribution method* was also proposed for the calculation of the flash point of chemicals [Albahri and George, 2003].

• **fugacity**

Fugacity is the tendency of a substance to move from one environmental compartment to another, that is, to prefer one phase (liquid, solid, gas) over another. At a fixed temperature and pressure, a chemical will have a different fugacity for each phase: the phase with the lowest fugacity will be the most favorable.

Originally, the term was applied to the tendency of a gas to expand or escape and related to its pressure in the system being studied.

• **Henry's law constant** (H)

The Henry's law gives the relationship between the partial pressure $P$ of a solute above the solution and its concentration $c$ in the solution; it is defined as

$$e^{P} = e^{\mathrm{H} \cdot \mathrm{c}}$$

or, using the natural logarithm, as

$$P = \mathrm{H} \cdot \mathrm{c}$$

where H is the Henry's law constant; its units are L· atm/mol, atm/(mol fraction), or Pa·m$^3$/mol.

The Henry's law constant varies with the solvent and the temperature.

🕮 [Nirmalakhandan and Speece, 1989b; Dunnivant, Elzerman *et al.*, 1992; ; Russell, Dixon *et al.*, 1992; Suzuki, Ohtaguchi *et al.*, 1992a; English and Carroll, 2001; Mariussen, Andersson *et al.*, 2001; Delgado and Alderete, 2002; Zhong, Yang *et al.*, 2002; Dearden and Schüürmann, 2003; Taskinen and Yliruusi, 2003; Wang, Tang *et al.*, 2003; Yaffe, Cohen *et al.*, 2003]

• **magnetic susceptibility** ($\chi^m$)
It is the degree of magnetization of a material in response to an applied magnetic field. To distinguish magnetic susceptibility from → *dielectric susceptibility*, it is often denoted by $\chi^m$ and it relates the magnetization **M** of a material with the intensity of the applied magnetic field **H**:

$$\mathbf{M} = \chi^m \cdot \mathbf{H}$$

The magnetic induction **B** is related to **H** by the relationship

$$\mathbf{B} = \mu_0 \cdot (\mathbf{H} + \mathbf{M}) = \mu_0 \cdot (1 + \chi^m) \cdot \mathbf{H} = \mu \cdot \mathbf{H}$$

where $\mu_0$ is the magnetic permeability in the vacuum and $\mu$ the **magnetic permittivity** of the material.

If $\chi^m$ is positive, that is, $(1 + \chi) > 1$, the material is called paramagnetic and the magnetic field is strengthened by the presence of the material. Alternatively, if $\chi^m$ is negative, that is, $(1 + \chi) < 1$, the material is diamagnetic and the magnetic field is weakened by the presence of the material.

🕮 [Dauben, Wilson *et al.*, 1968; Schmalz, Klein *et al.*, 1992; Estrada, 1998a]

• **melting point** (MP)
It is the temperature at which the solid and liquid states of a pure substance can exist in equilibrium; the melting point of a crystalline solid is the temperature at which it changes state from solid to liquid.

As heat is applied to a solid, its temperature increases until it reaches the melting point. At this temperature, additional heat converts the solid into a liquid without a change in temperature.

When considered as the temperature of the reverse change from liquid to solid, it is referred to as the **freezing point**. For most substances, melting and freezing points are equal.

Molecular size and symmetry usually increase the melting point; however, unlike the boiling point, the melting point is relatively insensitive to pressure. Melting points are often used to characterize organic compounds and to ascertain the purity. The melting point of a pure substance is always higher than the melting point of that substance when a small amount of an impurity is present. Moreover, together with the → *octanol–water partition coefficient*, melting point is used in the → *general solubility equation* to predict solubility of compounds.

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

• **molar refractivity** (MR)
The molar refractivity is the volume of the substance taken up by each mole of that substance. In SI units, MR is expressed as $m^3$/mol. MR is a molecular descriptor of a liquid, which contains both information about molecular volume and polarizability, usually defined by the Lorenz–Lorentz equation [Lorentz, 1880a, 1880b] (also known as the Clausius–Mosotti equation):

$$\mathrm{MR} = \frac{n_D^2 - 1}{n_D^2 + 2} \cdot \frac{\mathrm{MW}}{\rho} = \frac{\varepsilon - 1}{\varepsilon + 2} \cdot \bar{V}$$

where MW is the → *molecular weight*, $\rho$ the liquid → *density*, and $\bar{V}$ the → *molar volume*, and $n_D$ the → *refractive index* of the liquid referred to the sodium D line, and its square coincides with the → *dielectric constant* $\varepsilon$.

Molar refractivity is also proportional to → *polarizability* $\alpha$, by the following [Hansch and Leo, 1995]:

$$MR = \frac{4}{3} \cdot \pi \cdot N_A \cdot \alpha$$

where $N_A$ the Avogadro number (or Loschmidt constant), equal to $6.022\,141\,79 \times 10^{23}\ mol^{-1}$, that is, the number of molecules in a mole of substance.

Molar refractivity can be used to design a set of bioactive molecules so that covariance between MR and hydrophobicity is minimized; MR can serve as a measure of binding force between the polar portions of an enzyme and its substrate.

Alternative definitions of molar refractivity were proposed by Gladstone and Dale ($MR_{GD}$) [Gladstone and Dale, 1858] and Vogel ($MR_V$) [Vogel, 1948] as

$$MR_{GD} = (n-1) \cdot \frac{MW}{\rho} \qquad MR_V = n \cdot MW$$

where n is the refractive index.

Molar refractivity estimates by substituting the molar volume by → *Mc Gowan's characteristic volume* $V_X$ were proposed by Abraham *et al.* [Abraham, Whiting *et al.*, 1990b] as

$$MR_A = 10 \cdot f(n) \cdot V_X$$

where $f(n)$ is the → *refractive index function*. Moreover, to remove cohesive dispersion interactions, it was proposed to subtract the molar refractivity of the *n*-alkane with the same characteristic volume $V_X$:

$$R_2 = MR_A - MR_A^* = MR_A - (2.83195 \cdot V_X - 0.52553)$$

where $MR_A$ is the molar refractivity of the considered compound and $MR_A{}^*$ the molar refractivity of the *n*-alkane with the same characteristic volume $V_X$. The parameter $R_2$ can be considered a polarizability descriptor and is called **excess molar refractivity**. By definition, $R_2 = 0$ for all *n*-alkanes, and the same holds for branched alkanes.

When molar refractivity is determined using the sodium D-line, it coincides with the → *electron polarization*. Therefore, it can be considered contemporarily as being an → *electronic descriptor* as well as a → *steric descriptor* of compounds.

As molar refractivity is essentially an additive property, **group molar refractivity** is calculated as the difference between the molar refractivity of an X-substituted compound and the reference compound:

$$MR_X = MR_{X+REF} - MR_{REF}$$

This parameter is often used as a substituent steric constant in → *Hansch analysis*. To put the molar refractivities of the substituents on approximately the same scale as the → *hydrophobic substituent constants* $\pi$, the substituent MR values are often scaled down by a factor 0.1.

The difference between the molar refractivity of a substituent $MR_X$ and hydrogen $MR_H$ was used to estimate the difference in the interaction energy of a hydrogen-substituted parent

compound and an X-substituted analogue compound:

$$\Delta E_{INT} = \frac{-1673.6}{r_{XB}^6} \cdot (MR_X - MR_H) \text{ kJ/mol}$$

where $r_{XB}$ is the distance in angstroms between the group and the binding site [Pauling and Pressman, 1945].

Values for the atomic molar refractivity were also estimated by → *group contribution methods* [Ghose and Crippen, 1987].

The **molar refractivity partition index**, denoted as ${}^{P}MR_{\chi}$, is a → *Randić-like index* derived from the → *H-depleted molecular graph* of a compound as [Padrón, Carrasco *et al.*, 2002]

$${}^{P}MR_{\chi} = \sum\nolimits_b \left[\gamma_i^{MR} \cdot \gamma_j^{MR}\right]_b^{-1/2} \quad i \neq j$$

where the summation goes over all the bonds and $\gamma_i$ is the **atomic refractivity** of the *i*th atom plus the atomic refractivity of the hydrogens bonded to the *i*th atom; *i* and *j* indicate the two atoms forming the bond *b*.

**Table P4** Molar refractivity values for different atom types.

| Atom-type | Atomic refractivity | Atom type | Atomic refractivity |
|---|---|---|---|
| $Csp^3$ | 2.8128 | N(Ar) | 2.7662 |
| $Csp^2$ | 3.8278 | $NO_2$ | 3.5054 |
| Csp | 3.8974 | Ar−N=X | 3.8095 |
| C(Ar) | 3.5090 | F | 1.0632 |
| C=X | 3.0887 | Cl | 5.6105 |
| H | 0.9155 | Br | 8.6782 |
| −O− | 1.6351 | I | 13.8741 |
| =O | 1.7956 | $Ssp^3$ | 7.3190 |
| O=N | 2.1407 | $Ssp^2$ | 9.1680 |
| $Nsp^3$ | 3.0100 | R−SO−R | 6.0762 |
| $Nsp^2$, Nsp | 3.2009 | $R-SO_2-R$ | 5.3321 |

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

• **molecular weight** (MW)

Among the → *size descriptors*, molecular weight is the most simple and used molecular → *0D-descriptor*, calculated as the sum of the atomic weights of all the atoms in a molecule. It is related to molecular size and is atom-type sensitive. It is defined as

$$MW = \sum_{i=1}^{A} m_i$$

where m is the atomic mass and *i* runs over the *A* atoms of the molecule. The **average molecular weight** defined as

$$\overline{\mathrm{MW}} = \frac{1}{A} \cdot \sum_{i=1}^{A} \mathrm{m}_i = \frac{\mathrm{MW}}{A}$$

is also used as molecular descriptor and is related to → *atomic composition indices*.

**Square root molecular weight** (MW2), defined as $\mathrm{MW2} = \mathrm{MW}^{1/2}$, and **cubic root molecular weight** (MW3), defined as $\mathrm{MW3} = \mathrm{MW}^{1/3}$ and corresponding to a linear dimension of size, are also used as descriptors of molecule size.

• **parachor** (PA)

The parachor is defined by the Sudgen equation as [Sudgen, 1924]

$$\mathrm{PA} = \gamma^{1/4} \cdot \frac{\mathrm{MW}}{\rho_\mathrm{L} - \rho_\mathrm{V}} \approx \gamma^{1/4} \cdot \frac{\mathrm{MW}}{\rho_\mathrm{L}} = \gamma^{1/4} \cdot \bar{\mathrm{V}}$$

where MW is the → *molecular weight*, $\gamma$ the liquid → *surface tension*, and $\rho_\mathrm{L}$ and $\rho_\mathrm{V}$ the → *density* at a given temperature of liquid and vapor, respectively. The second relationship holds when the vapor density is negligible with respect to the liquid density, $\bar{\mathrm{V}}$ being the → *molar volume*. This expression is considered to be an additive quantity, that is, can be approximately expressed as a sum of empirical increments $\mathrm{PA}_i$ corresponding to the single atoms or groups in the molecule. As an additive quantity, the parachor has been used in solving various structural problems.

The parachor is related to physico-chemical properties depending on the molecule volume, that is, → *boiling point*. It is essentially constant over wide ranges of temperature.

📖 [Vogel, 1948; Quayle, 1953; Ahmad, Fyfe *et al.*, 1975; Briggs, 1981; Zhao, Abraham *et al.*, 2003a; Tiwari and Pande, 2006]

• **partition coefficients**

A partition coefficient or distribution coefficient is a measure of the equilibrium between two different means, such as two different phases or two different immiscible liquids [Dearden, 1985]. It is usually denoted by $K$ or $P$ and defined as the ratio of the concentrations of a compound in a two-compartment system under equilibrium conditions:

$$K \equiv P = \frac{[\mathrm{C}]_1}{[\mathrm{C}]_2}$$

where $[\mathrm{C}]_1$ and $[\mathrm{C}]_2$ are the concentrations of the solute in the two systems. The partition coefficients are usually transformed in a logarithmic form as

$$\log P = \log \frac{[\mathrm{C}]_1}{[\mathrm{C}]_2} = \log[\mathrm{C}]_1 - \log[\mathrm{C}]_2$$

Partition coefficients are dimensionless measures of the relative affinity of a molecule with respect to the two phases and depend on absorption, transport, and partitioning phenomena.

In most of the cases, the two phases are an organic phase and an aqueous phase, that is, the partitioning of a compound between a lipidic and an aqueous phase.

The best known of these partition coefficients is the one based on the solvents 1-octanol and water. The **octanol–water partition coefficient** $K_{\mathrm{ow}}$, very often expressed in its logarithmic form

log $K_{ow}$ (also denoted as log $P_{ow}$ or, often, simply as log $P$) is a measure of the hydrophobicity and hydrophilicity of a substance measured as partition between 1-octanol (the lipidic phase) and water (the polar phase):

$$K_{ow} \equiv P = \frac{[C]_{1\text{-}octanol}}{[C]_{water}}$$

To avoid possible associations of the solute in the organic phase, partition coefficients should be measured at low concentrations or extrapolated to infinite dilution of the solute.

In the context of drug-like substances, hydrophobicity is related to absorption, bioavailability, hydrophobic drug–receptor interactions, metabolism and toxicity. Closely related to log $P$ is the **octanol–water distribution coefficient** (log $D_{pH}$), accounting for partition of pH-dependent mixture of ionizable species. Ionization of any compound makes it more water soluble and then less lipophilic. The log $D$ can be calculated from log $P$ and → *acid dissociation constant* $pK_a$ by the following expression [Cronin, Aptula *et al.*, 2002b; Livingstone, 2003]:

$$\log D_{pH} = \log P - \log(1 + 10^{(pH - pK_a)\cdot I_{ab}})$$

where $I_{ab}$ is equal to 1 for acids and to –1 for bases.

Due to its importance in QSAR studies, several approaches were proposed for modeling → *lipophilicity* of chemical compounds.

Other common partition coefficients are soil sorption partition coefficient, gas–solvent partition coefficient and micelle–water partition coefficient, together with → *leaching indices*, which are partition indices thought of for environmental studies.

**Soil sorption partition coefficient** (or **soil–water partition coefficient**), denoted as $K_{oc}$ or **log $K_{oc}$**, accounts for sorption from water into soil. Because this often depends primarily on the soil's organic carbon content, measured values are usually normalized for the organic carbon (OC) content of soil, in which case the soil sorption equilibrium constant is expressed as

$$K_{oc} = \frac{[C_{soil}]/[C^0_{soil}]}{[C_w]/[C^0_w]}$$

where $[C_{soil}]$ is the concentration of solute per gram of carbon in a standard soil and $[C_w]$ is the concentration of solute per volume of aqueous solution. The standard state concentrations $[C^0_{soil}]$ and $[C^0_w]$ are typically chosen as 1 μg of solute/g of organic carbon for soil and 1 μg of solute/ml for aqueous solution.

Several models for estimating soil sorption coefficients take advantage of the correlation between $K_{oc}$ and other experimental partition coefficients, specially $K_{ow}$. For example, $K_{oc}$ values have been estimated from experimental octanol–water partition coefficients by

$$\log K_{oc} = m \cdot \log K_{ow} + b$$

where $m$ and $b$ are slope and intercept, respectively, of the developed linear regression models. Published values of $m$ and $b$ range from 0.5 to 1.1 and from −0.2 to 1.3, respectively, depending on the range of data employed in the individual regression [Winget, Cramer *et al.*, 2000]. Moreover, the → *adsorbability index* was proposed as a $K_{oc}$ descriptor.

**Gas–solvent partition coefficient** is known as the **Ostwald solubility coefficient** $L$ and is usually written in the logarithmic form as [Katritzky, Mu *et al.*, 1996a; Katritzky,

Oliferenko *et al.*, 2003a]

$$\log L = \log\left(\frac{[C_l]}{[C_g]}\right)$$

where $[C_l]$ and $[C_g]$ are the concentrations of the substance in the liquid solvent and in the gas, respectively. It is used in the → *Linear Solvation Energy Relationships*.

**Micelle–water partition coefficient**, denoted as $K_{mw}$ or in its logarithmic form as **log $K_{mw}$** or log $P_{mw}$, is the partition of a solute between micellar and aqueous phases [Tanaka, Nakamura *et al.*, 1994; Abraham, Chadha *et al.*, 1995a].

Micellization is typical of surfactants that are organic molecules having a chemical structure combining both a polar (amphiphobic) and a nonpolar (amphiphilic) group into a single molecule. When dissolved in a solvent at low concentration, they have the ability to adsorb at interfaces, thereby alter significantly physical properties of the interfaces. In particular, micellization is observed in surfactant solutions when the concentration exceeds the *critical micelle concentration* (cmc), whereas the physico-chemical properties of the aqueous solution change abruptly [Li, Zhang *et al.*, 2004; Jalali-Heravi and Konouz, 2005].

Micelle–water partition coefficients are extracted by micelle chromatography (high performance liquid chromatography, HPLC) using micelle aqueous solution as mobile phase. For determination of $K_{mw}$, → *retention times* are measured using a usual HPLC system at various concentrations of micelle in the aqueous mobile phase and then estimated from the following equation:

$$K_{mw} = k' \cdot \phi = k' \cdot \frac{V_{mc}}{V_{aq}}$$

where $k'$ is the → *retention factor* and $\phi$ is the phase ratio, defined as the volume of the micellar pseudostationary phase over that of the bulk aqueous phase ($V_{mc}/V_{aq}$), which is related to two intrinsic properties of the surfactant [Liu, Yao *et al.*, 2006; Katritzky, Pacureanu *et al.*, 2007].

As it was for the soil sorption partition coefficient, models for estimating micelle–water partition coefficients take advantage of the correlation with $K_{ow}$. For example, $K_{mw}$ values have been estimated from experimental octanol–water partition coefficients by

$$\log K_{mw} = m \cdot \log K_{ow} + b$$

where $m$ and $b$ are slope and intercept, respectively, of the developed linear regression model [Ishihama, Oda *et al.*, 1996; Trone, Leonard *et al.*, 2000].

📖 [Tanaka and Fujiwara, 1996; Winget, Cramer *et al.*, 2000; Chen, Harner *et al.*, 2003; Fichert, Yazdanian *et al.*, 2003; Basak, Mills *et al.*, 2004; Kahn, Fara *et al.*, 2005]

- **pH**

It is the common measure of the acid-base character of a solution, defined as

$$pH = -\log_{10}[H^+]$$

where $[H^+]$ is the concentration of hydrogen ions in moles per liter. The most precise definition is in terms of activity rather than concentration.

A solution of pH below 7 is acid, pH of 7 is neutral, pH over 7 is alkaline.

• **refractive index** (n)
The refractive index (or **index of refraction**) of a medium is defined as a ratio of the velocity of light in vacuum over the velocity of light in the substance of interest (a medium), or, in other words, is a measure for how much the speed of light (or other waves such as sound waves) is reduced inside the medium. For example, typical glass has a refractive index of 1.5, which means that light travels at $1/1.5 = 0.67$ times the speed in air or vacuum.

Used as an indicator of the purity of organic compounds, it is related to several electric and magnetic properties such as polarizability as well as to molar refractivity, critical temperature, surface tension, density, and boiling point. Usually, the refractive index is measured at the sodium D-line and indicated as $n_D^2$. Moreover, the **refractive index function** *f*(n) defined as

$$f(n) = \frac{n^2 - 1}{n^2 + 2}$$

was proposed as a molecular descriptor, accounting for composite solute interactions [Fuchs, Abraham *et al.*, 1982]. The refractive index of polymers is also among the important → *technological properties* of polymers.

📖 [Vogel, Cresswell *et al.*, 1951; Huggins, 1956; Katritzky, Sild *et al.*, 1998a, 1998b; Holder, Ye *et al.*, 2006a, 2006b; Cao and Gao, 2007]

• **solubility** (*S*)
Solubility is the maximum amount of solute that dissolves in a given quantity of solvent at a specific temperature, that is solubility refers to the ability for a given substance, the solute, to dissolve in a solvent. The resulting solution is called a saturated solution. Certain substances are soluble in all proportions with a given solvent, such as, for example, ethanol in water. This property is more correctly described as miscible.

Generally, for a solid in a liquid, solubility increases with temperature; for a gas, solubility decreases. Common measures of solubility include the mass of solute per unit mass of solution (mass fraction), mole fraction of solute, molality, molarity, and others.

Aqueous solubility is among the most important characteristics in ADME studies and plays a relevant role as physico-chemical descriptor in QSAR studies.

Solute–solvent interactions were largely studied and modeled by → *Linear Solvation Energy Relationships* and the → *Hildebrand solubility parameter.*

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

• **surface tension** (γ)
It is the attraction of molecules to each other on a liquid's surface, or, more specifically, the attractive intermolecular forces that liquid molecules below the surface exert on molecules at the surface. It is defined as

$$\gamma = [PA \cdot (\rho_L - \rho_V)]^4$$

where PA is the → *parachor*, and $\rho_L$ and $\rho_V$ are the liquid and vapor densities, respectively.

Surface tension creates a strong boundary between the air and liquid and is among the important → *technological properties* of substances.

📖 [Sudgen, 1924; Wiener, 1948a; Stanton and Jurs, 1992; Gutman, Popovic *et al.*, 1997; Kauffman and Jurs, 2001a; Knotts, Wilding *et al.*, 2001]

• **vapor pressure** ($V_P$)
The vapor pressure of a liquid is the pressure exerted by its vapor when the liquid and vapor are in dynamic equilibrium.

Vapor pressure is an indication of a liquid's evaporation rate. It relates to the tendency of molecules and atoms to escape from a liquid or a solid. A substance with a high vapor pressure at normal temperature is often referred to as volatile. The higher the vapor pressure of a material at a given temperature, the lower the boiling point.

The vapor pressure of any substance increases nonlinearly with temperature according to the Clausius–Clapeyron relation.

📖 [Wiener, 1948b; Pitzer, Lippmann *et al.*, 1955; Balaban and Feroiu, 1990; Basak, Gute *et al.*, 1997; Myrdal and Yalkowsky, 1997; Katritzky, Wang *et al.*, 1998; Liang and Gallagher, 1998; Goll and Jurs, 1999b; Simmons, 1999; Beck, Breindl *et al.*, 2000; Engelhardt McClelland and Jurs, 2000; Basak and Mills, 2001b; Chalk, Beck *et al.*, 2001; Olsen and Nielsen, 2001; Dearden, 2003b; Raevsky, Raevskaja *et al.*, 2007]

➢ **PI index** → Szeged matrices
➢ **Pisanski–Zerovnik index** → Wiener index
➢ **p$K_a$** ≡ *acid dissociation constant* → physico-chemical properties (⊙ equilibrium constants)
➢ **planted tree** → graph (⊙ tree)
➢ **Platt number** ≡ *total edge adjacency index* → edge adjacency matrix
➢ **PLS-based variable selection** → variable selection
➢ **P-matrix** → bond order indices (⊙ graphical bond order)
➢ **Pogliani cis/trans connectivity index** → *cis/trans* descriptors
➢ **Pogliani index** → Zagreb indices
➢ **point-by-point alignment** → alignment rules
➢ **polar effect** → electronic substituent constants
➢ **polar hydrogen factor** → electric polarization descriptors
➢ **polarity/polarizability descriptors** → electric polarization descriptors
➢ **polarity number** → distance matrix
➢ **polarizability** → electric polarization descriptors
➢ **polarizability effect index** → electric polarization descriptors
➢ **polarizability tensor** → electric polarization descriptors
➢ **polarizability volume** → electric polarization descriptors (⊙ mean polarizability)
➢ **polarization** → electric polarization descriptors
➢ **polar surface area** → molecular surface (⊙ solvent-accessible molecular surface)
➢ **Politzer hydrophobic model** → lipophilicity descriptors
➢ **polycenter** → center of a graph

## polymer descriptors

Polymers are large molecules constituted of repeating structural units connected by covalent chemical bonds. Polymers are characterized by some specific → *physico-chemical properties*, → *technological properties* and conformational characteristics such as steric hindrance, characteristic ratio, persistence length, statistical chain segment (or Kuhn segment) length, molar stiffness function (also called molar limiting viscosity number function), intrinsic viscosity, and glass transition temperature [Katritzky, Maran *et al.*, 2000].

Molecular descriptors for polymers with an infinite number of repeating units are often calculated for small sequences (dimers, trimers) or for the single repeating unit.

Polymer descriptors ranges from → *quantum chemicals descriptors* [Holder, Ye *et al.*, 2006a; Yu, Yi *et al.*, 2007] to → *graph invariants* [Gutman, Kolaković *et al.*, 1989b, 1989a; Hosoya, 1991; Bonchev, Mekenyan *et al.*, 1992; Patil, Bora *et al.*, 1995; Gutman and Rosenfeld, 1996; Liu and Zhong, 2005], from the typical descriptors used in → *Linear Solvation Energy Relationships* [Kamlet, Abraham *et al.*, 1984; Taft, Abraham *et al.*, 1985; Kamlet, Doherty *et al.*, 1987a; Moody, Willauer *et al.*, 2005] to quantities computed by → *group contribution methods* [Elbro, Fredeslund *et al.*, 1991].

Polymer descriptors are the → *characteristic ratio*, and, among the → *size descriptors*, the → *Kuhn length*, the → *end-to-end distance*, the → *persistent length*.

Other examples of polymer descriptors are given below.

The **Wiener polymer index** is a normalized Wiener index for infinite polymers defined as [Balaban, Balaban *et al.*, 2001]

$$W_\infty = \frac{d}{3 \cdot (A_{pc} + C_{pc})}$$

where $d$ is the shortest topological distance between two equivalent atoms in two neighboring polymer units, $A_{pc}$ and $C_{pc}$ are the number of atoms and cycles in the polymer unit.

The **mean overcrossing number** $\bar{N}$ is a descriptor for polymer chains accounting for the occurrence of entanglements caused by polymer chains interpenetrating each other (Figure P1). The mean overcrossing number is a → *geometrical descriptor* defined as the number of bond–bond crossings in a regular 2D projection of the chain, averaged over all possible projections and calculated on the → *molecular geometry* [Arteca, 1999]. It is a suitable descriptor of DNA chains, polymer geometrical shape, rheological and dynamic properties of polymer melts and concentrated solutions being explained by the occurrence of entanglements that cause geometrically constrained chain motion.

Moreover, the **average writhe** $\overline{W}_r$ was also defined as the observed overcrossing sum for each given 2D-projection, distinguishing right-handed crossing $(+1)$ from left-handed crossing $(-1)$. By definition, $\bar{N} \geq \overline{W}_r$. Both $\bar{N}$ and $\overline{W}_r$ provide useful information: in a compact random configuration a large value of $\bar{N}$ and a vanishing value of $\overline{W}_r$ are expected, whereas in a configuration with regular dihedral angles (e.g., a compact helix) both $\bar{N}$ and $\overline{W}_r$ are expected to be large.

These two descriptors can be combined to produce an effective polymer shape parameter, called **order parameter** $\varsigma$, such as

$$\varsigma = \bar{N} - \frac{\overline{W}_r}{\bar{N}}$$

which exhibits two regular trends: $\varsigma \to 0$ in a nonentangled regular configuration and $\varsigma \to 1$ in an entangled random configuration.

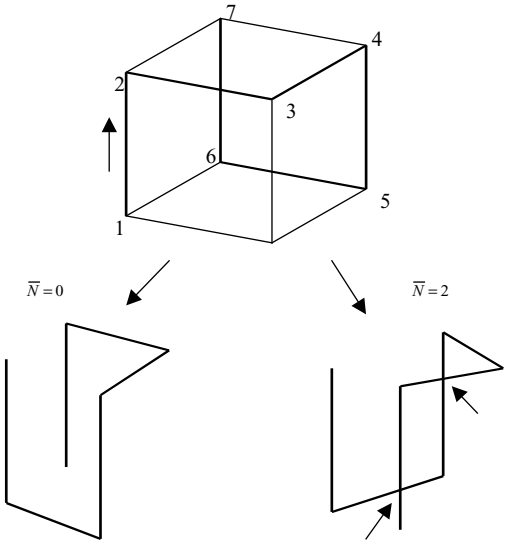



**Figure P1** Example of calculation of the mean overcrossing number.

Another descriptor of the macromolecular topology is the **linking number** $L$ that characterizes the entanglements of molecules having at least two molecular loops [White, 1969]. For two disjoint curves $C_1$ and $C_2$, viewed along a direction in space, the linking number is computed as the sum of the handedness indices of only overcrossings for which curve $C_1$ is underneath $C_2$, ignoring the overcrossings of each curve with itself. Two separate, nonentangled curves yield $L = 0$; the simplest nontrivial link of two loops, $L = 1$.

📖 [Small, 1953; Mekenyan, Dimitrov *et al.*, 1963; Volkenstein, 1963; Fuller, 1971; Bonchev and Mekenyan, 1980; Bonchev, Mekenyan *et al.*, 1981a, 1981b; Kamlet, Doherty *et al.*, 1986a, 1987c; Artemi and Balaban, 1987; Balaban and Artemi, 1987; Arteca and Mezey, 1990; Maranas, 1996; Balaban and Artemi, 1998; Katritzky, Sild *et al.*, 1998a; Katritzky, Sild *et al.*, 1998c; Sundaram and Venkatasubramanian, 1998; Camarda and Maranas, 1999; Zhong, Yang *et al.*, 2002; Arteca, 2003a, 2003b; Camacho-Zuñiga and Ruiz-Treviño, 2003; Edvinsson, Arteca *et al.*, 2003; Adams and Schubert, 2004; Afantitis, Melagraki *et al.*, 2005; Bonchev, Markel *et al.*, 2005; Funar-Timofei, Kurunczi *et al.*, 2005; Liu and Zhong, 2005; Shevade, Homer *et al.*, 2006; Yu, Wang *et al.*, 2006; Xu, Liu *et al.*, 2007]

- **polynomial** → algebraic operators
- **population analysis** → quantum-chemical descriptors
- **population trace** → DARC/PELCO analysis
- **positive predictive value** → classification parameters
- **potential of a charge distribution** → charge descriptors
- **Potential Pharmacophore Point pairs** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **power matrices** → matrices of molecules

- **power matrix** → algebraic operators (⊙ product of matrices)
- **PPFS** ≡ *Property and Pharmacophore Features Score* → scoring functions
- ***P′/P* index** → bond order indices (⊙ graphical bond order)
- **PPP eigenvalues** → spectral indices
- **PPP pairs** ≡ *Potential Pharmacophore Point pairs* → substructure descriptors (⊙ pharmacophore-based descriptors)
- **PPP-triangle descriptors** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **Pratt measure** → statistical indices (⊙ concentration indices)
- **precision** → classification parameters
- **prediction error sum of squares** → regression parameters
- **predictive residual sum of squares** → regression parameters
- **predictive square error** → regression parameters
- **predictor variables** ≡ *independent variables* → data set
- **prime ID number** → ID numbers
- **principal axes of a molecule** → principal moments of inertia
- **principal components** → Principal Component Analysis

## Principal Component Analysis (PCA)

A fundamental chemometric technique for → *exploratory data analysis*, transforming the $p$ variables in the data matrix $\mathbf{X}$ ($n \times p$), where $n$ is the number of objects, into linear combinations of the common factors $\mathbf{T}$ ($n \times M$), called **principal components** and denoted by $\mathbf{t_m}$:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{L}^{\mathbf{T}}$$

where $\mathbf{T}$ is the **score matrix**, $\mathbf{L}$ ($p \times M$) the **loading matrix**, and $M$ the number of significant principal components ($M \leq p$). The columns of the loading matrix represent the eigenvectors $\mathbf{l}_m$; the eigenvector coefficients $\ell_{jm}$ $(-1 \leq \ell_{jm} \leq +1)$, called *loadings*, represent the importance of each original variable (the rows of the loading matrix) in the considered eigenvector [Jolliffe, 1986; Jackson, 1991; Basilevsky, 1994].

The principal components are calculated according to the maximum variance criterion, that is, each successive component is an orthogonal linear combination of the original variables such that it covers the maximum of the variance not accounted for by the previous components. The eigenvalue $\lambda_m$ associated with each $m$th component represents the variance explained by that component. Moreover, the sum of the variances of all the components equals the variance of the original variables.

The principal components are as linear combinations of the $p$ original variables:

$$\mathbf{T} = \mathbf{X} \cdot \mathbf{L}, \quad \text{that is,} \quad t_{im} = x_{i1} \cdot \ell_{1m} + x_{i2} \cdot \ell_{2m} + \cdots + x_{ip} \cdot \ell_{pm} = \sum_{j=1}^{p} x_{ij} \cdot \ell_{jm}$$

where $\ell_{jm}$ are the coefficients of the linear combinations (i.e., the *loadings*) and $t_{im}$ is the PCA score of the $i$th object (e.g., molecule, amino acid, etc.) in the $m$th principal component.

Mathematically, PCA consists in the diagonalization of the → *correlation matrix* (or covariance matrix) of the data matrix $\mathbf{X}$ with size $p \times p$ (the number of variables).

The main advantages of principal components are that

(1) each component is orthogonal to all the remaining components, that is, the information carried by this component is unique;
(2) each component represents a *macrovariable* of the data;
(3) components associated with the lowest eigenvalues do not usually contain useful information (noise, spurious information, etc.).

Once $M$ significant components have been chosen, each $i$th object is represented by the $M$-dimensional score vector:

$$\{t_{i1}; t_{i2}; \ldots; t_{iM}\}$$

often called **z-scores** (or **z-scales**) and denoted as $\{z_{i1}; z_{i2}; \ldots; z_{iM}\}$ or **principal properties** PP and denoted as $\{\mathrm{PP}_{i1}; \ldots; \mathrm{PP}_{iM}\}$.

Principal properties PPs (or $z$-scores) are → *vectorial descriptors* of compounds, which summarize the main information of the original molecular descriptors or are empirical scales describing the physico-chemical properties of the training set objects [Alunni, Clementi *et al.*, 1983; Carlson, 1992; Clementi, Cruciani *et al.*, 1993a]. The number of significant PPs and their meaning depend closely on the original variables used to perform PCA. Since the PPs derived from PCA are orthogonal to each other and their number is usually small, they are suitable for design problems [Skagerberg, Bonelli *et al.*, 1989; Eriksson, Johansson *et al.*, 1997].

Principal properties can be calculated for both whole molecules and substituent groups, fragments, amino acids, and so on. For example, the $i$th substituent can be represented by four PPs, each having a different meaning such as $\mathrm{PP}_1$ = steric, $\mathrm{PP}_2$ = lipophilic, $\mathrm{PP}_3$ = electrostatic, $\mathrm{PP}_4$ = H-bonding properties of the substituent, respectively.

→ *BC(DEF) parameters* are principal properties of a data matrix given by six physico-chemical properties describing 114 diverse liquid-state compounds.

Principal properties calculated on molecular → *interaction energy values* obtained by → *grid-based QSAR techniques* are usually referred to as **3D principal properties** (**3D-PP**) [van de Waterbeemd, Clementi *et al.*, 1993]. They were originally proposed for a theoretical description of the amino acids [Norinder, 1991; Cocchi and Johansson, 1993]. **3D-PP** were also calculated from → *ACC transforms*. Several other principal properties were proposed as the → *amino acid descriptors*.

When different data sets of descriptors are used separately to derive the principal properties of the same compounds, **disjoint principal properties** (**DPP**) are obtained as the whole set of significant **PP**s derived from each block of descriptors:

$$\{\mathbf{PP}_1^{\mathrm{A}}, \mathbf{PP}_2^{\mathrm{A}}, \ldots, \mathbf{PP}_{M_{\mathrm{A}}}^{\mathrm{A}}; \mathbf{PP}_1^{\mathrm{B}}, \mathbf{PP}_2^{\mathrm{B}}, \ldots, \mathbf{PP}_{M_{\mathrm{B}}}^{\mathrm{B}}; \mathbf{PP}_1^{\mathrm{C}}, \mathbf{PP}_2^{\mathrm{C}}, \ldots, \mathbf{PP}_{M_C}^{C}\}$$

where A, B, C represent three different blocks of variables on which PCA was performed and $M_{\mathrm{A}}$, $M_{\mathrm{B}}$, and $M_{\mathrm{C}}$ the corresponding numbers of significant principal components [van de Waterbeemd, Costantino *et al.*, 1995].

📖 [Weiner and Weiner, 1973; Dunn III and Wold, 1978, 1980; Dunn III, Wold *et al.*, 1978; Wold, 1978; Lukovits and Lopata, 1980; Streich, Dove *et al.*, 1980; Lukovits, 1983; McCabe, 1984; Maria, Gal *et al.*, 1987; Eriksson, Jonsson *et al.*, 1988, 1989, 1990; van de Waterbeemd,

El Tayar *et al.*, 1989; Hemken and Lehmann, 1992; Ridings, Manallack *et al.*, 1992; Suzuki, Ohtaguchi *et al.*, 1992a; Tysklind, Lundgren *et al.*, 1992; Caruso, Musumarra *et al.*, 1993; Cristante, Selves *et al.*, 1993; Ordorica, Velazquez *et al.*, 1993; Rodríguez Delgado *et al.*, 1993Rodríguez Delgado, Sánchez *et al.*, 1993; Bazylak, 1994; Franke, Gruska *et al.*, 1994; Norinder, 1994; Azzaoui and Morinallory, 1995; Bjorsvik and Priebe, 1995; Cocchi, Menziani *et al.*, 1995; Clementi, Cruciani *et al.*, 1996; Gibson, McGuire *et al.*, 1996; Kimura, Miyashita *et al.*, 1996; Bjorsvik, Hansen *et al.*, 1997; Young, Profeta *et al.*, 1997; Balasubramanian and Basak, 1998; Langer and Hoffmann, 1998a; Kuanar, Kuanar *et al.*, 1999a; Vendrame, Braga *et al.*, 1999; Xue, Godden *et al.*, 1999b]

- **principal component analysis feature selection** → variable reduction
- **principal inertia axes** ≡ *principal axes of a molecule* → principal moments of inertia

■ **principal moments of inertia** ($I_A$, $I_B$, $I_C$) (≡ *inertia principal moments*)
They are physical quantities related to the rotational dynamics of a molecule. The **moment of inertia** about any axis is defined as

$$I = \sum_{i=1}^{A} m_i \cdot r_i^2$$

where $A$ is the atom number, $m_i$ and $r_i$ are the atomic mass and the perpendicular distance from the chosen axis of the $i$th atom of the molecule, respectively. For any rectangular coordinate system, with axes X, Y, Z, three moments of inertia are defined as

$$I_{XX} = \sum_{i=1}^{A} m_i \cdot (y_i^2 + z_i^2) \qquad I_{YY} = \sum_{i=1}^{A} m_i \cdot (x_i^2 + z_i^2) \qquad I_{ZZ} = \sum_{i=1}^{A} m_i \cdot (x_i^2 + y_i^2)$$

where ($x$, $y$, $z$) are the coordinates of the atoms.

The corresponding cross-terms are called **products of inertia** and are defined as

$$I_{XY} = Y_{YX} = \sum_{i=1}^{A} m_i \cdot x_i \cdot y_i \qquad I_{XZ} = Y_{ZX} = \sum_{i=1}^{A} m_i \cdot x_i \cdot z_i \qquad I_{YZ} = Y_{ZY} = \sum_{i=1}^{A} m_i \cdot y_i \cdot z_i$$

Therefore the **inertia matrix**, denoted by **I**, is a square symmetric matrix $3 \times 3$, collecting the three moment of inertia and six products of inertia.

Principal moments of inertia are the moments of inertia corresponding to that particular and unique orientation of the axes for which one of the three moments has a maximum value, another a minimum value, and the third is either equal to one of the others or intermediate in value to the other two. The corresponding axes are called **principal axes of a molecule** (or **principal inertia axes**). Moreover, the products of inertia all reduce to zero and the corresponding inertia matrix is diagonal. Conventionally, principal moments of inertia are labeled as

$$I_A \leq I_B \leq I_C$$

In general, the three principal moments of inertia have different values, but, depending on the molecular symmetry, they show characteristic equalities such as those shown in Table P5.

A number of → *shape descriptors* is defined in terms of principal moments of inertia. Moreover, principal moments of inertia are used to provide a unique reference framework for the calculation of the → *shadow indices*, and, in general, are used to define → *alignment rules* of the molecules. They constitute the basic starting point for the calculation of → *WHIM descriptors* and → *CoMMA method*.

**Table P5** Principal moments for some selected symmetries.

| Symmetry | Principal moments | Example |
|---|---|---|
| Spherical top | $I_A = I_B = I_C$ | $CCl_4$ |
| Symmetric top | $I_A = I_B \neq I_C$ | $NH_3$ |
| Asymmetric top | $I_A \neq I_B \neq I_C$ | $CH_2FCl$ |
| Linear symmetry | $0 = I_A \neq I_B = I_C$ | HC≡CH |
| Planar symmetry | $I_A + I_B = I_C$ | $C_6H_6$ |

- **principal properties** → Principal Component Analysis
- **privileged pharmacophore keys** → substructure descriptors (⊙ pharmacophore-based descriptors)

## **Probabilistic Receptor Potential** (PRP)

This is a 3D-QSAR method designed to predict, in a qualitative manner, the types of receptor atoms to which a compound would prefer to bind [Labute, 2001].

To this end, molecules with different binding activities are aligned and common hydrogen-bond and hydrophobic regions are determined. Then, the type of interactions that most likely occur at different regions around the compounds are evaluated.

- **probability matrices** ≡ *stochastic matrices* → algebraic operators
- **probe** → grid-based QSAR techniques
- **products of inertia** → principal moments of inertia
- **product of matrices** → algebraic operators
- **product of row sums index** ≡ *PRS index* → distance matrix
- **proference** → DARC/PELCO analysis
- **Property and Pharmacophore Features Score** → scoring functions
- **Property and Pharmacophore Features fingerprints** → scoring functions (⊙ Property and Pharmacophore Features Score)
- **Property-Encoded Surface Translator descriptors** ≡ *PEST descriptors* → TAE descriptor methodology

## **property filters**

A property filter is a set of general and objective rules based on limits on structural features and physico-chemical properties that are shared by drugs or lead compounds. These rules are extracted from large collections of chemicals, containing both generic chemicals and drugs. By comparing a collection of known drugs with a collection of nondrugs, distribution of structural features and properties of compounds are analyzed by different methods to identify those features and value ranges of properties qualifying a compound to be a drug.

To focus drug discovery toward effective and orally adsorbable compounds, properties considered are usually related to *Absorption, Distribution, Metabolism, Excretion* (→ *ADME properties*).

Property filters are largely used in screening of virtual libraries and design of combinatorial libraries, allowing selection from large chemical database of compounds with desired properties to be potential drugs or, alternatively, removal of existing compounds with undesired properties [Clark and Pickett, 2000; Oprea, 2003; Leach, Hann *et al.*, 2006]. When filters are used to extract good drug candidates, they are usually referred to as **drug-like indices**. When they are applied to identify those chemicals that is likely to fail the development process, the term **alert indices** is more appropriate. When filters are only based on limits on functional groups they are properly called **functional group filters** [Muegge, 2003; Walters and Murcko, 2002] or **chemical filters** [Oprea, Gottfries *et al.*, 2000].

Different authors are using the term "drug-like" with slightly different meaning. Muegge says that "drug-likeness is a general descriptor of the potential of a small molecule to become a drug. It is not a unified descriptor but a global characteristic of a compound possessing many specific characteristics such as good solubility, membrane permeability, half-life, and having a pharmacophore pattern to interact specifically with a target protein. In reality, highly potent compounds against a drug target may not be efficacious because of pharmacokinetic problems; they may be toxic or unfavorably interact with other drugs" [Muegge, 2003].

Lipinski defines drug-like "*those compounds that have sufficiently acceptable ADME properties and sufficiently acceptable toxicity properties to survive through the completion of human Phase I clinical trials*" [Lipinski, 2000]. Walters and Murcko define drug-like compounds as those "*molecules which contain functional groups and/or have physical properties consistent with the majority of known drugs*" [Walters and Murcko, 2002].

There is a large variety of molecular descriptors used to address drug-likeness: they range from constitutional and counting descriptors to topological descriptors, from physico-chemical properties to pharmacophore description, from thermodynamic considerations to the synthetic accessibility, from presence of functional groups to ADME/Tox properties.

Several chemoinfomatic approaches were proposed to evaluate drug-likeness of compounds; these include simple counting rules, such as property filters, and more complex regression and classification models, obtained by machine learning algorithms based on → *artificial neural networks* and recursive partitioning. These models have been used to derive descriptor weights and → *scoring functions* that classify compounds as drug or nondrug.

Moreover, to improve the chances of finding a drug candidate, it has been suggested to select small rational libraries from large libraries, or, in other words, to select a set of compounds with properties representative of the large library [Ashton, Jaye *et al.*, 1996]. This set of representative compounds can be selected by means of clustering or cell-based methods. → *Cell-based methods* require one or more quantitative molecular properties accounting for ligand–receptor binding interactions and properties involved in the transport of the drug to its target. Unlike common clustering methods, cell-based methods are more suitable to identify missing diversity in a chemical library and to highlight underrepresented or unrepresented regions of the overall chemical space.

Property filters usually are binary variables assuming a value equal to 1, if the molecule shows a specific property (drug-likeness, ADME properties, and toxicities) and equal to zero otherwise. These filters are not comprised of many molecular descriptors and a threshold or a range of values is associated to each descriptor together with a condition on the descriptor value: if the

conditions are fulfilled for all the descriptors, the studied property is considered as potentially present in the molecule. Usually, a few violations are allowed.

In the following, some property filters are reported. They are divided into drug-like indices and lead-like indices depending on whether they address drug-likeness or lead-likeness of compounds. In the section functional group filters, a survey of the most common functional groups for database filtering is given. All the property filters that allow a drug-likeness ranking of compounds instead of a simple yes/no response are reported elsewhere under → *scoring functions*.

- **drug-like indices**

The **Lipinski drug-like index** (or **rule-of-five**, RO5) is the first drug-like filter proposed to predict oral bioavailability of compounds that have achieved phase II clinical status [Lipinski, Lombardo *et al.*, 1997, 2005]. This filter predicts that poor absorption or permeation is more likely when more than one violation is registered for the four following rules: molecular weight (MW) $\leq 500$, log $P \leq 5$, number of hydrogen-bond acceptors ($HBA$) $\leq 10$; number of hydrogen-bond donors ($HBD$) $\leq 5$.

The Lipinski rules were derived from an analysis of 2245 drugs from the WDI database; they identify compounds lying in a region of property space where the probability of useful oral activity is very high. A compound that fails the filter, that is, two or more properties are out of range, will likely be poorly bioavailable because of poor absorption or permeation.

As the rule-of-five was designed to predict compound bioavailability, it is not really able to distinguish between drugs and nondrugs [Frimurer, Bywater *et al.*, 2000; Oprea, 2000]. Moreover, there are some limitations of the rule-of-five [Keller, Pichota *et al.*, 2006]: (1) RO5 applies only to compounds that are delivered by the oral route (not applicable for substrates of transporter and natural products); (2) RO5 applies only to compounds that are absorbed by passive mechanisms [Lipinski, Lombardo *et al.*, 2001]; (3) important RO5 violations come from antibiotics, antifungals, vitamins, and cardiac glycosides [Walters and Murcko, 2002]; (4) compliant compounds are not necessarily good drugs; (5) RO5 says nothing about specific chemical structural features found in drugs or nondrugs [Lipinski, 2004].

Bhal *et al.* proposed a revised rule-of-five by using the logarithm of the → *octanol–water distribution coefficient* (log $D_{\mathrm{pH}}$), at pH 5.5 (log $D_{5.5}$), instead of log $P$ because log $D$ is a better descriptor for lipophilicity accounting for the ionization of compounds under physiological conditions [Bhal, Kassam *et al.*, 2007]. The idea underpinning this replacement is that since ionization of molecules results in decreased lipophilicity with respect to the neutral state, it is necessary to take into account the ionic state of the compound when describing the lipophilicity of potential drugs.

To achieve a better distinguishing between drugs and nondrugs, other property filters are defined which are extensions of the rule-of-five. Some of them are collected in Table P6 and briefly commented in the text below.

The drug-like filter proposed by Chen *et al.* is applied after a first structural screening aimed at excluding compounds containing atoms different from C, H, O, N, S, P, F, Cl, Br, or I [Chen, Zheng *et al.*, 2005]. Moreover, the three descriptors added to those of the RO5 are → *combined descriptors* defined as ratio of → *count descriptors*: C3p is the ratio of the number of C($sp^3$) atoms over the total number of nonhalogen heavy atoms; h-p is the ratio of the number of hydrogen atoms over the total number of nonhalogen heavy atoms; Unsat-p is the ratio of molecular unsaturation, as defined in → *multiple bond descriptors*, over the number of nonhalogen heavy atoms. The same authors also proposed a simple filter based only on two molecular descriptors

**Table P6** Lipinski rule-of-five (R05) and related drug-like filters.

| Descriptor | *RO5* | Oprea *et al.* | Chen *et al.* | Monge *et al.* | Walters *et al.* | Rishton |
|---|---|---|---|---|---|---|
| MW | ≤500 | [200; 450] | [78; 500] | [100; 800] | [200; 500] | ≤500 |
| log *P* | ≤5 | [−2; 4.5] | [−0.5; 5] | ≤7 | [−5; 5] | ≤5 |
| *HBA* | ≤10 | [1; 8] | [2; 10] | ≤10 | ≤10 | ≤10 |
| *HBD* | ≤5 | ≤5 | ≤5 | ≤5 | ≤5 | ≤5 |
| *RBN* | | [1; 9] | | ≤15 | ≤8 | ≤10 |
| *NRG* | | ≤5 | | ≤6 | | |
| *PSA* (Å$^2$) | | | | | | ≤140 |
| C3p | | | [0.15; 0.8] | | | |
| h-p | | | [0.6; 1.6] | | | |
| Unsat-p | | | [0.10; 0.45] | | | |
| Charge | | | | | [−2; +2] | |
| Halogens | | | | ≤7 | | |
| O + N | | | | ≥1 | | |

MW, molecular weight; *HBA*, number of hydrogen-bond acceptors; *HBD*, number of hydrogen-bond donors; *RBN*, number of rotatable bonds; *NRG*, number of rings (cyclomatic number); *PSA*, partial surface area. Noncited descriptors are defined in the text. Data from [Oprea, Gottfries *et al.*, 2000; Oprea, 2000; Chen, Zheng *et al.*, 2005; Monge, Arrault *et al.*, 2006; Walters and Murcko, 2002; Rishton, 2003].

that are independent of the molecular size [Zheng, Luo *et al.*, 2005]: one is the unsaturation-related descriptor Unsat-p and the other is a descriptor related to the proportion of heteroatoms NO_C3, defined as the ratio of the total number of oxygen and nitrogen atoms over the number of carbon atoms with sp$^3$ hybridization. The filter for drug-like compounds is then,

$$\text{Unsat-p} \leq 0.43 \quad \text{and} \quad 0.10 \leq \text{NO\_}C3 \leq 1.8$$

The filter of Monge *et al.* includes some additional rules based on molecular structural features. In particular: (a) compounds with atoms other than C, H, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li are not allowed to pass the filter; (b) no reactive functions; (c) no perfluorinated chains (e.g., $-CF_2CF_2CF_3$); (d) no rings with more than seven members; (e) alkyl chains $\leq -(CH_2)_6CH_3$. This filter was derived from the analysis of 2.6 million compounds collected from 32 diverse chemical databases.

The property filter of Walters *et al.* is implemented in the program REOS where a set of more than 200 functional group filters is also available to enable one to remove compounds with toxic, reactive, and otherwise undesirable moieties.

The filter proposed by Rishton is based on data taken from the literature.

Another property filter designed to predict oral bioavailability was proposed by [Veber, Johnson *et al.*, 2002] by substituting the four Lipinski rules with the following two rules: (a) number of rotatable bonds ≤10, and (b) polar surface area (*PSA*) ≤ 140 Å$^2$ or the sum of H-bond acceptors and H-bond donors ≤ 12.

Eight **GVW drug-like indices** have been proposed by Ghose–Viswanadhan–Wendoloski [Ghose, Viswanadhan *et al.*, 1999] to help streamline the design of combinatorial chemistry libraries for drug design and develop guidelines for prioritizing large sets of compounds for biological testing. They are based on a consensus definition and have been derived from analysis

of the distribution of some physico-chemical properties (log *P*, molar refractivity, molecular weight, number of atoms) and chemical constitutions of known drug molecules available in the Comprehensive Medicinal Chemistry (CMC) database and seven drug classes defined by disease state.

Among the eight proposed indices is a general drug-like index that has been derived from the analysis of the whole CMC database and seven specific drug-like indices derived from the property distributions within the single drug classes (Table P7).

log *P* (→ *ALOGP*) and → *molar refractivity* (AMR) are calculated by using the atomic contribution method of Ghose, Crippen, and Viswanadhan. The drug-like indices are dummy variables taking value equal to 1 when all the criteria of the consensus definition of a drug-like molecule are satisfied, 0 otherwise. Specifically, a drug-like index equals 1 when log *P*, molar refractivity, molecular weight (MW), and number of atoms (*A*) of a compound are in the property range reported in Table P7; moreover, the compound must be a combination of some of the following functional groups: a benzene ring, a heterocyclic ring (both aliphatic and aromatic), an aliphatic amine, a carboxamide group, an alcoholic hydroxyl group, a carboxy ester, and a keto group. For example, according to the CMC-80 index, an organic compound is a drug-like molecule if: the calculated ALOGP is between −0.4 and 5.6, the molar refractivity AMR between 40 and 130, the molecular weight MW between 160 and 480, the total number of atoms *A* between 20 and 70, and it includes at least one of the above mentioned functional groups.

Two property ranges have been proposed: the *qualifying range* that covers approximately 80% of the drugs studied and the *preferred range* that is the smallest range within the qualifying range occupied by approximately 50% of the drugs. If large compound databases are screened by means of the indices based on the qualifying range (80%), the chance of missing drug-like

**Table P7** Value ranges of the descriptors used in defining GVW drug-like indices.

| Drug class | *P*% | ALOGP | AMR | MW | *A* |
|---|---|---|---|---|---|
| CMC | 80 | [−0.4; 5.6] | [40; 130] | [160; 480] | [20; 70] |
| CMC | 50 | [1.3; 4.1] | [70; 110] | [230; 390] | [30; 55] |
| Antiinflammatory | 80 | [1.4; 4.5] | [59; 119] | [212; 447] | [24; 59] |
| Antiinflammatory | 50 | [2.6; 4.2] | [67; 97] | [260; 380] | [28; 40] |
| Antidepressant | 80 | [1.4; 4.9] | [62; 114] | [210; 380] | [32; 56] |
| Antidepressant | 50 | [2.1; 4.0] | [75; 95] | [260; 330] | [37; 48] |
| Antipsychotic | 80 | [2.3; 5.2] | [85; 131] | [274; 464] | [40; 63] |
| Antipsychotic | 50 | [3.3; 5.0] | [94; 120] | [322; 422] | [49; 61] |
| Antihypertensive | 80 | [−0.5; 4.5] | [54; 128] | [206; 506] | [28; 66] |
| Antihypertensive | 50 | [1.0; 3.4] | [68; 116] | [281; 433] | [36; 58] |
| Hypnotic | 80 | [0.5; 3.9] | [43; 97] | [162; 360] | [20; 45] |
| Hypnotic | 50 | [1.3; 3.5] | [43; 73] | [212; 306] | [29; 38] |
| Antineoplastic | 80 | [−1.5; 4.7] | [43; 128] | [180; 475] | [21; 63] |
| Antineoplastic | 50 | [0.0; 3.7] | [60; 107] | [258; 388] | [30; 55] |
| Antiinfective | 80 | [−0.3; 5.1] | [44; 144] | [145; 455] | [12; 64] |
| Antiinfective | 50 | [0.8; 3.8] | [68; 138] | [192; 392] | [12; 42] |

ALOGP, Ghose–Crippen–Viswanadhan log *P*; AMR, Ghose–Crippen–Viswanadhan molar refractivity; MW, molecular weight; *A*, number of atoms; *P*%, the percentage of covering.

compounds is less than 20%. To make the search/design for new drugs more efficient the indices based on the preferred range (50%) may be used, even if the chance of missing good compounds increases in this case.

Note that as these indices depend on ALOGP, their values are provided only for compounds having C, H, O, N, S, Se, P, B, Si, and halogens.

The **rule-of-unity**, proposed by Yalkowski *et al.* [Sanghvi, Ni *et al.*, 2003; Yalkowsky, Johnson *et al.*, 2006], is a drug-like filter based on a single **absorption parameter** Π calculated by the ratio of the → *octanol–water partition coefficient*, $K_{ow}$, over the *luminal oversaturation number* $O_{Lumen}$, that is,

$$\Pi = \frac{K_{ow}}{O_{Lumen}} = \frac{K_{ow}}{\max\left(1, \frac{4 \cdot Dose}{S_W}\right)}$$

The absorption parameter was defined to predict whether or not at least half of the administered drug will be absorbed.

The **luminal oversaturation number** is defined as the maximum of either unity or four times the dose in grams per 0.250 l of water divided by the aqueous solubility, $S_w$, of the drug in grams per liter [Sanghvi, Ni *et al.*, 2003].

The luminal oversaturation number is a dimensionless number that cannot be less than unity and distinguishes between drugs that are soluble in the gastrointestinal contents from drugs that are not. The former will dissolve readily, whereas the latter will exist as suspensions that will maintain a saturated solution in the gut until sufficient absorption has taken place so that no suspended particles remain. For the calculation of solubility, the **general solubility equation** of Jain–Yalkowsky is used [Yalkowsky, 1999; Jain and Yalkowsky, 2001]:

$$\log S_W = 0.5 - \log K_{ow} - 0.01 \times (\mathrm{MP} - 25)$$

where MP is the → *melting point*.

Drugs with a Π absorption parameter greater than unity tend to be well absorbed (i.e., absorbed fraction > 0.5), while drugs with Π values of less than or equal to 1 are poorly absorbed (absorbed fraction < 0.5). Thus, absorption is most efficient and hence drug-likeness more likely when the absorption parameter Π is greater than unity. This most often occurs when the partition coefficient is greater than unity and/or the oversaturation number is equal to unity.

- **lead-like indices**

The term "drug-like" is used for compounds resembling existing drugs, while the term "lead-like" for compounds possessing the structural and physico-chemical profile of a quality lead [Verheij, 2006]. The concept of lead-like is more restrictive for some terms with respect to the concept of drug-like [Monge, Arrault *et al.*, 2006], depending on the fact that optimization of a lead compound often results in an increase of molecular weight, log*P* and complexity and in a decrease of solubility [Teague, Davis *et al.*, 1999; Hann, Leach *et al.*, 2001; Oprea, 2002a].

Leads should display the following properties to be considered for further development [Oprea, Davis *et al.*, 2001]: (1) relative simple chemical features; (2) membership to a well-established structure–activity relationship series, wherein compounds with similar (sub)-structure exhibit similar target binding affinity; (3) favorable patent situations; and (4) good → *ADME properties*. Moreover, in a strict sense, the definition of leads requires the presence of at least one marketed drug, derived from that particular lead structure.

On an average, compared to drugs, leads have lower molecular complexity (lower molecular weight, less rings and rotatable bonds), lower polarizability, are less hydrophobic (their log*P* is 0.5–1.0 units less than that of drugs), and have lower drug-like scores [Hann, Leach *et al.*, 2001; Oprea, Davis *et al.*, 2001]. Therefore, in general, physico-chemical property values used as a measure of lead-likeness should be lower than those traditionally used for drug-likeness. Moreover, structural features need to be accounted for in defining lead-likeness since there are various different types of structures that yield false positive hits, such as reactive structures or those that irreversibly bind to the target [Rishton, 2003; Lipinski, 2004].

A collection of lead-like indices is reported in Table P8.

**Table P8** Lead-like filters.

| | **Congreve *et al.*** | **Oprea *et al.*** | **Hann–Oprea** | **Verheji** | **Monge *et al.*** | **Wenlock *et al.*** |
|---|---|---|---|---|---|---|
| MW | <300 | ≤450 | ≤460 | ≤450 | ≤460 | ≤473 |
| log *P* | ≤3 | [−3.5; 4.5] | [−4; 4.2] | [−2.0; 4.5] | [−4.0; 4.2] | [−2.0; 5.5] |
| *HBA* | ≤3 | ≤8 | ≤9 | ≤10 | ≤9 | ≤7 |
| *HBD* | ≤3 | ≤5 | ≤5 | ≤5 | ≤5 | ≤4 |
| *RBN* | ≤3 | ≤10 | ≤10 | ≤10 | ≤10 | ≤10 |
| *NRG* | — | ≤4 | ≤4 | — | ≤4 | ≤4 |
| log $D_{7.4}$ | — | [−4; 4] | — | — | — | ≤4.3 |
| *PSA* (Å$^2$) | ≤60 | — | — | ≤150 | — | — |
| log $S_w$ | — | — | ≥−5 | ≥−6 | — | — |
| Halogens | — | — | — | * | ≤7 | — |
| N + O | — | — | — | — | ≥1 | — |

MW, molecular weight; *HBA*, number of hydrogen-bond acceptors; *HBD*, number of hydrogen-bond donors; *RBN*, number of rotatable bonds; *NRG*, number of rings (cyclomatic number); log $D_{7.4}$, log of the distribution coefficient at pH 7.4; *PSA*, partial surface area; log $S_w$, water solubility; Halogens, number of halogen atoms; N + O, total number of nitrogen and oxygen atoms. Data from [Congreve, Carr *et al.*, 2003; Oprea, Davis *et al.*, 2001; Hann and Oprea, 2004; Verheij, 2006; Monge, Arrault *et al.*, 2006; Wenlock, Austin *et al.*, 2003].

The filter proposed by Congreve *et al.* was called the **rule-of-three** (RO3) because, by analogy with the Lipinski → *rule-of-five*, the limits on molecular properties are all multiples of three instead of five.

The filter proposed by Verheij for lead-like compound selection is based on seven molecular descriptors representing molecular properties involved in early discovery, such as oral availability and permeability [Verheij, 2006]. Cutoff values of the descriptors were derived from [Lipinski, Lombardo *et al.*, 1997; Lipinski, 2000; Hann, Leach *et al.*, 2001; Oprea, 2002a; Veber, Johnson *et al.*, 2002]. Moreover, the polar surface area is estimated by the model of the → *topological polar surface area* (*TPSA*). The filter of Monge *et al.* is an extension of the filter of Hann and Oprea, which includes some additional structural rules. To the limits on the number of halogen atoms and the total number of oxygen and nitrogen atoms, the filter also includes the following rules: (a) no atoms other than C, H, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, or Li; (b) no reactive functions; (c) no perfluorinated chains (e.g., $-CF_2CF_2CF_3$); (d) no rings with more than seven members; (e) alkyl chains $\leq -(CH_2)_6CH_3$.

The filter of Wenlock *et al.* is derived from a statistical analysis of a set of marketed oral drugs that are compounds with acceptable physico-chemical properties that have successfully enabled them to overcome the obstacles of development for their desired therapeutic indication.

- **functional group filters** (≡ *chemical filters*)

Functional group filters are applied to exclude from a chemical database those structures that possess undesired functionalities. These can be structures having more than one aldehyde group, structures containing metals, reactive alkyl halides, peroxides, carbazides.

In general, these filters are designed to recognize those functional groups that tend to be toxic or unstable under physiological conditions. A survey of reactive structures that should be avoided in selection of drug or lead candidates is reported by [Rishton, 2003] (Table P9).

**Table P9** List of functional groups responsible for electrophilic protein-reactive false positive from Rishton [Rishton, 2003].

| | | |
|---|---|---|
| Sulfonyl halides | Acyl halides | Alkyl halides |
| Anhydrides | Halopyrimidines | α-Halocarbonyl compounds |
| 1,2-Dicarbonyl compounds | Aldehydes | Aliphatic ketones |
| Perhalo ketones | Aliphatic esters | Imines |
| Epoxides | Aziridines | Thioesters |
| Sulfonate esters | Phosphonate esters | Heteroatom–heteroatom |
| Michael acceptors | β-Heterosubstituted carbonyl compounds | single bonds |

Examples of structural filters implemented in the program REOS are listed in Table P10 [Walters and Murcko, 2002].

**Table P10** List of REOS functional group filters from [Walters and Murcko, 2002].

| | | |
|---|---|---|
| Sulfonyl halides | Nitro groups | Aldehydes |
| Primary alkyl halides | Epoxides | Aziridines |
| Sulfonate esters | Phosphonate esters | Long aliphatic chains |
| Peroxides | 1,2-Dicarbonyl compounds | Acyl halides |

To remove potentially toxic compounds, functional group filters primarily draw from mutagenicity, carcinogenicity, and acute toxicity database [Muegge, 2003].

The **structural alerts** (SA) are chemical filters highlighting molecular substructures or reactive groups that are mainly related to the carcinogenic and mutagenic properties of the chemicals, and represent a sort of "codification" of a long series of studies aimed at highlighting the mechanisms of action of the mutagenic and carcinogenic chemicals [Benigni and Bosa, 2006]. A review about carcinogenic and mutagenic effects and related QSAR models was published by [Frierson, Klopman *et al.*, 2006].

A very effective representation of the structural alerts has been provided by Ashby [Ashby, 1985; Ashby and Tennant, 1988] in the form of a hypothetical poly-carcinogen chemical comprised of most of the known SAs (Figure P2). In Table P11, the structural alert groups are collected.

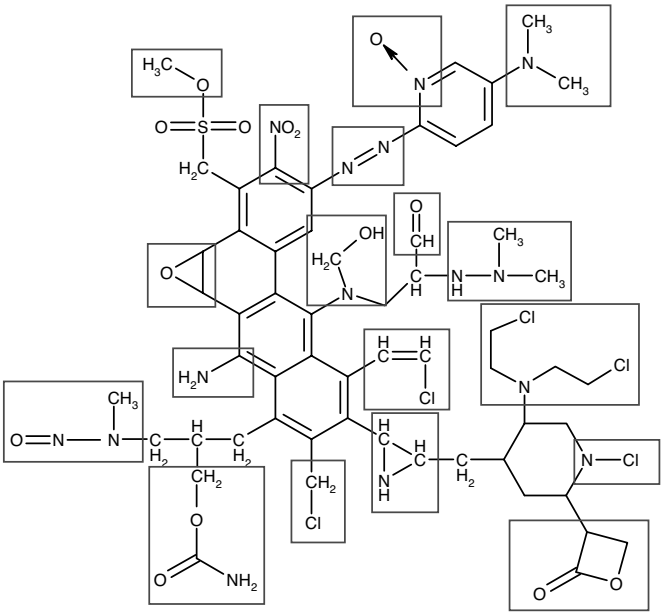


**Figure P2** The hypothetical poly-carcinogen chemical proposed by Ashby [Ashby,1985; Ashby and Tennant, 1988]

**Table P11** Structural alerts proposed by Ashby [Ashby, 1985; Ashby and Tennant, 1988].

| Structural alert | Structural alert |
|---|---|
| Aromatic nitro groups | Alkyl esters of either phosphoric or sulfonic acids |
| Aromatic rings *N*-oxides | Aromatic mono- and dialkylamino groups |
| Alkyl hydrazines | Aromatic azo groups (because of possible reduction to aromatic amines) |
| Alkyl aldehydes | Aromatic and aliphatic aziridinyl derivatives |
| *N*-methyl derivatives | Aromatic and aliphatic substituted primary alkyl halides |
| Monoalkenes | Aromatic amines (including their *N*-hydroxy derivatives and the derived esters |
| β-Haloethyl mustards | Propriolactones and propriosultones |
| *N*-Chloroamines | Derivatives of urethane (carbamates) |
| Alkyl *N*-nitrosoamines | Aliphatic and aromatic epoxides |

Each of the SAs is a "code" for a well-characterized chemical class, with its own specific mechanism of action. However, there are also general physico-chemical factors that may influence the potential reactivity of a chemical, that is, one could expect to observe compounds with structurally alerting features but that are biologically inactive because of a number of reasons, such as molecular weight, solubility, reactivity, and so on.

Starting from eight general toxicophores from the Ashby compilation, a list of 29 toxicophores containing new substructures was proposed to classify compounds according to their mutagenicity (Table P12) [Kazius, McGuire *et al.*, 2005].

**Table P12** Extended list of structural alerts according to [Kazius, McGuire *et al.*, 2005].

| Structural alerts | Structural alerts |
|---|---|
| Specific aromatic nitro | Unsubstituted heteroatom-bonded heteroatom |
| Specific aromatic amine | Nitrogen and sulfur mustard |
| Aromatic nitroso | Polycyclic aromatic systems (PAH) |
| Alkyl nitrite | Bay-region in PAH |
| Nitrosoamine | K-region in PAH |
| Epoxide | Aliphatic *N*-nitro |
| Aziridine | α,β-Unsaturated aldehydes (including R-carbonyl aldehydes) |
| Azide | Diazonium |
| Diazo | β-Propriolactone |
| Triazene | α,β-Unsaturated alcoxy group |
| Aromatic azo | 1-Aryl-2-monoalkyl hydrazine |
| Carboxylic acid halide | Aromatic methylamine |
| Aromatic hydroxylamine | Ester derivative of aromatic Hydroxylamine |
| Aliphatic halide | Polycyclic planar systems |
| Sulfonate-bonded carbon (alkyl alkane sulfonate or dialkyl sulfate) | |

Structural alerts were also searched for within the framework of the **Threshold Toxicological Concern** (*TTC*), aimed at reducing extensive toxicity evaluations [Benigni and Bosa, 2006]. This approach refers to the establishment of a generic human exposure threshold value for groups of chemicals below which there would be no appreciable risk to human health. The underlying principle is that such a value can be identified for many chemicals, including those of unknown toxicity, when considering their chemical structures and the known toxicity of chemicals that share similar structural characteristics. Moreover, the concept that there are levels of exposure that do not cause adverse effects is strictly related to the possibility of setting → *acceptable daily intakes* for chemicals with known toxicological profiles. A general *TTC* approach, mainly based on carcinogenicity data, was adopted by the US Food and Drug Administration Threshold of Regulation for indirect food additives. An extension of this approach to a range a dietary concentrations was proposed by using QSARs, genotoxicity and short term toxicity data [Cheeseman, Machuga *et al.*, 1999]. This resulted in the identification of eight more complex, less generalized structural alerts, that include a majority of the most potent of the 709 carcinogens (Table P13). This study shows that the inclusion of structural alerts as criteria for substances proposed for approval under a threshold of regulation process, can significantly increase the safety

**Table P13** Structural alerts in the *TTC* approach according to [Cheeseman, Machuga *et al.*, 1999].

| Structural alerts | Structural alerts |
|---|---|
| *N*-nitroso compounds | α-Nitro furyl compounds |
| Endocrine disruptors | Hydrazines/triazine/azides/azoxy compounds |
| Strained heteronuclear rings | Polycyclic amines |
| Heavy metal compounds | Organophosphorous compounds |

assurance margin. Substances that do not belong to any of the structural alert classes are likely to have much lower carcinogenic potencies, and therefore may qualify for a higher threshold level.

📖 [Gayoso and Kimri, 1990b, 1990a; Bemis and Murcko, 1996; Stahl and Böhm, 1998; Clark, 1999b, 1999a; Kelder, Grootenhuis *et al.*, 1999; Walters, Ajay *et al.*, 1999; Egan, Merz Jr. *et al.*, 2000; Sakaeda, Okamura *et al.*, 2001; Borodina, Filimonov *et al.*, 2002; Egan and Lauri, 2002; Norinder and Haeberlein, 2002; Engkvist, Wrede *et al.*, 2003; Fichert, Yazdanian *et al.*, 2003; Olah, Bologa *et al.*, 2004b; Vieth, Siegel *et al.*, 2004; te Heesen *et al.*, 2007te Heesen, Schlitter *et al.*, 2007]

- **properties matrix** → topoelectric matrices
- **protein folding degree index** → biodescriptors (⊙ peptide sequences)
- **protein sequences** → biodescriptors (⊙ peptide sequences)
- **protein TOMOCOMD descriptors** → TOMOCOMD descriptors
- **proteo-chemometrics approach** → Structure/Response Correlations
- **PRP** ≡ *Probabilistic Receptor Potential*
- **proteomics maps** → biodescriptors
- **PRS index** → distance matrix
- **pruning of the graph** → centric indices (⊙ Balaban centric index)
- **pruning partition** → centric indices (⊙ Balaban centric index)
- **pseudocenter** → center of a graph
- **pseudoconnectivity indices** → electrotopological state indices
- **pseudograph** → graph

## P_VSA descriptors

These are molecular descriptors defined as the amount of van der Waals surface area (VSA) having a property P in a certain range [Labute, 2000]. These descriptors correspond to a partition of the molecular surface area conditioned by the atomic values of the property P.

To generate P_VSA descriptors, first, the van der Waals surface area $\mathrm{VSA}_i$ of each atom is estimated according to the following:

$$\mathrm{VSA}_i = 4 \cdot \pi \cdot R_i^2 - \pi \cdot R_i \cdot \sum_{j=1}^{A} a_{ij} \cdot \left( \frac{R_j^2 - (R_i - g_{ij})^2}{g_{ij}} \right)$$

where $R$ is the atomic van der Waals radius, the summation goes over all the atoms, but accounts only for contributions from atoms bonded to the $i$th atom, $a_{ij}$ being the elements of the → *adjacency matrix*. The quantity $g_{ij}$ is calculated as

$$g_{ij} = \min\{\max\{|R_i - R_j|, b_{ij}\}, (R_i + R_j)\}$$

where the term $b_{ij}$ is the ideal length of the bond formed by atoms $i$ and $j$, calculated according to the formula:

$$b_{ij} = r_{ij}^* - c_{ij}$$

where $r_{ij}^*$ is a reference bond length and $c_{ij}$ a correction term depending on the → *bond multiplicity*: 0 for single bond, 0.1 for aromatic, 0.2 for double, and 0.3 for triple bonds.

In Tables P14 and P15, the van der Waals radii and the reference bond lengths used for P_VSA calculations are collected.

**Table P14** van der Waals radii (in Angstrom) used for P_VSA calculations.

| Atom-type | *R* | Atom-type | *R* |
|---|---|---|---|
| H (−O) | 0.8 | O (other) | 1.779 |
| H (−N, −P) | 0.7 | F | 1.496 |
| H (other) | 1.485 | P | 2.287 |
| C | 1.950 | S | 2.185 |
| N | 1.950 | Cl | 2.044 |
| O (oxide) | 1.810 | Br | 2.166 |
| O (acid) | 2.152 | I | 2.358 |

**Table P15** Reference bond lengths (in Angstrom) used for P_VSA calculations. The symbol ∼ indicates any kind of bond.

| Bond-type | $r^*$ | Bond-type | $r^*$ | Bond-type | $r^*$ |
|---|---|---|---|---|---|
| C∼C | 1.540 | H−N | 1.010 | N−N | 1.450 |
| C−H | 1.060 | H−O | 0.970 | N−O | 1.460 |
| C−N | 1.470 | H−P | 1.410 | N−P | 1.600 |
| C−O | 1.430 | H−S | 1.310 | N−S | 1.760 |
| C−P | 1.850 | H−F | 0.870 | O−O | 1.470 |
| C−S | 1.810 | H−Cl | 1.220 | O−P | 1.570 |
| C−F | 1.350 | H−Br | 1.440 | O−S | 1.570 |
| C−Cl | 1.800 | H−I | 1.630 | P−P | 2.260 |
| C−Br | 1.970 | | | P−S | 2.070 |
| C−I | 2.120 | | | S−S | 2.050 |

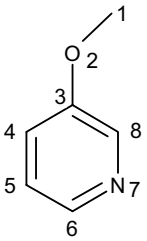
Let $P_i$ be a property of the $i$th atom. Then, the P_VSA descriptors are defined as

$$\text{P\_VSA}_k = \sum_{i=1}^{A} \text{VSA}_i \cdot \delta(\text{P}_i \in [a_{k-1}, a_k]) \quad k = 1, 2, \ldots n$$

where the summation goes over all the atoms, $\text{VSA}_i$ is the van der Waals surface area of the $i$th atom, $\delta(\text{P}_i \in [a_{k-1}, a_k])$ is the Dirac delta function that is equal to 1 for atoms with property value in the specified range, and zero otherwise; $a_0 \leq a_k < a_n$ are interval boundaries such that $[a_0, a_n)$ bounds all values of the property P in any molecule of the data set.

**Example P2**

A hypothetical atomic property is used to partition the van der Waals surface area into six different regions so that a six-dimensional P_VSA vector results.




| Atom | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $P_i$ | 2.4 | 1.2 | 4.5 | 5.9 | 5.7 | 3.1 | 0.2 | 3.9 |
| $VSA_i$ | 9.2 | 6.3 | 2.2 | 4.5 | 4.5 | 4.4 | 4.6 | 4.4 |

$$
\begin{aligned}
P\_VSA(0,2) &= VSA_2 + VSA_7 = 6.3 + 4.6 &&= 10.9\\
P\_VSA(2,3) &= VSA_1 &&= 9.2\\
P\_VSA(3,4) &= VSA_6 + VSA_8 = 4.4 + 4.4 &&= 8.8\\
P\_VSA(4,5) &= VSA_3 &&= 2.2\\
P\_VSA(5,6) &= VSA_4 + VSA_5 &&= 9.0\\
P\_VSA(6,7) &= - &&= 0
\end{aligned}
$$

P_VSA descriptors were calculated from several properties, such as atomic weight (m_VSA), atom polarizabilties (p_VSA), atom-type counts (a–*n*X_VSA), log *P* (*S* log P_VSA), molar refractivity (SMR_VSA), connectivity (δ_VSA), van der Waals volume (vdw_VSA), van der Waals surface (vsa_VSA), and van der Waals density (molecular weight divided by van der Waals volume, den_VSA)), hydrogen-bond donor (HBD_VSA) and acceptor (HBA_VSA), polar atom (hydrogen-bond donors plus hydrogen-bond acceptors (POL_VSA), hydrophobic atom (hyd_VSA), and partial charges (PEOE_VSA).

**Table P16** P_VSA descriptors in terms of log *P* [Labute, 2000], molar refractivity [Wildman and Crippen, 1999], and partial charges [Gasteiger and Marsili, 1980]. Property interval boundaries were optimized using data from a database of 44 795 small organic compounds.

| Property | No. | Interval boundaries for the calculation of P_VSA descriptors |
|---|---|---|
| log *P* | 10 | $(-\propto, -0.4, -0.2, 0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, +\propto)$ |
| Molar refractivity | 8 | $(0, 0.11, 0.26, 0.35, 0.39, 0.44, 0.485, 0.56, +\propto)$ |
| Partial charges | 14 | $(-\propto, -0.3, -0.25, -0.20, -0.15, -0.10, -0.05, 0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, +\propto)$ |

This methodology can be easily extended replacing atomic properties with any → *local vertex invariant* $\mathcal{L}_i$ and, in particular, with local vertex invariants obtained by atomic

properties using the Randić-like formula, that is, taking into account all the bonds incident to the atom.

📖 [Baurin, Mozziconacci *et al.*, 2004; Burton, Ijjaali *et al.*, 2006; Dubus, Ijjaali *et al.*, 2006; Ijjaali, Petitet *et al.*, 2007; Klon and Diller, 2007; Moorthy, Karthikeyan *et al.*, 2007]

➢ **P weighting scheme** → weighting schemes