



- **Labeled Hydrogen-Filled molecular Graph**  $\equiv$  *H-filled molecular graph*  $\rightarrow$  molecular graph
- **Labeled Hydrogen-Suppressed molecular Graph**  $\equiv$  *H-depleted molecular graph*  $\rightarrow$  molecular graph
- **labyrinthicity**  $\rightarrow$  walk counts
- **Laffort solute descriptors**  $\rightarrow$  Linear Solvation Energy Relationships
- **Lagrange distance**  $\rightarrow$  similarity/diversity ( $\odot$  Table S7)
- **Lance-Williams distance**  $\rightarrow$  similarity/diversity ( $\odot$  Table S7)
- **Laplacian graph energy**  $\rightarrow$  spectral indices

■ **Laplacian matrix (L)** ( $\equiv$  *admittance matrix, Kirchhoff matrix, combinatorial Laplacian matrix*)

This is a square  $A \times A$  symmetric matrix,  $A$  being the number of vertices in the  $\rightarrow$  *molecular graph*, obtained as the difference between the  $\rightarrow$  *vertex degree matrix V* and the  $\rightarrow$  *adjacency matrix A* [Mohar, 1989b, 1989a]:

$$\mathbf{L} = \mathbf{V} - \mathbf{A} = \mathbf{V}^{1/2} \cdot (\mathbf{I} - \mathbf{H}) \cdot \mathbf{V}^{1/2}$$

where  $\mathbf{V}$  is the diagonal  $\rightarrow$  *vertex degree matrix* of dimension  $A \times A$  whose diagonal entries are the  $\rightarrow$  *vertex degrees*  $\delta_i$ . In the last expression,  $\mathbf{I}$  indicates the  $\rightarrow$  *identity matrix* and  $\mathbf{H}$  is a matrix derived from the  $\rightarrow$  *random walk Markov matrix MM* by a similarity transformation. The matrix  $\mathbf{I} - \mathbf{H}$  is sometimes called the **normalized Laplacian matrix** [Chung, 1997]. Note that the negative half-sum of elements of the normalized Laplacian matrix is the  $\rightarrow$  *Randić connectivity index*  ${}^1\chi$  [Klein, Palacios *et al.*, 2004]:

$${}^1\chi = -\frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [(\mathbf{I} - \mathbf{H})]_{ij}$$

The entries of the Laplacian matrix formally are

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

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

The Laplacian matrix is also related to the vertex-edge and edge-vertex  $\rightarrow$  *incidence matrices*.

The diagonalization of the Laplacian matrix gives  $A$  real eigenvalues  $\lambda_i$  that constitute the **Laplacian spectrum** [Mohar, 1991b; Trinajstić, Babic *et al.*, 1994] and are conventionally labeled so that

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_A$$

Among the several properties of the Laplacian eigenvalues, three important ones are

- (a) the Laplacian eigenvalues are nonnegative numbers;
- (b) the last eigenvalue  $\lambda_A$  is always equal to zero;
- (c) the eigenvalue  $\lambda_{A-1}$  is greater than zero if, and only if, the graph  $G$  is connected; therefore, for a molecular graph all the Laplacian eigenvalues except the last are positive numbers.

Moreover, the sum of the positive eigenvalues is equal to twice the number  $B$  of graph edges, that is,

$$\sum_{i=1}^{A-1} \lambda_i \equiv \text{tr}(\mathbf{L}) = 2 \cdot B$$

The sum of the reciprocal  $A - 1$  positive eigenvalues was proposed as a molecular descriptor [Mohar, Babic *et al.*, 1993; Gutman, Yeh *et al.*, 1993] and called the **quasi-Wiener index**  $W^*$  [Marković, Gutman *et al.*, 1995]; it is defined as

$$W^* = A \cdot \sum_{i=1}^{A-1} \frac{1}{\lambda_i}$$

For acyclic graphs, the quasi-Wiener index  $W^*$  coincides with the  $\rightarrow$  *Wiener index*  $W$ , that is,  $W^* = W$ , while for cycle-containing graphs the two descriptors differ. Moreover, it has been demonstrated that the quasi-Wiener index coincides with the  $\rightarrow$  *Kirchhoff number* for any graph [Gutman and Mohar, 1996].

The product of the positive  $A - 1$  eigenvalues of the Laplacian matrix gives the **spanning tree number**  $T^*$  of the molecular graph  $G$  as

$$T^* = \frac{1}{A} \cdot \prod_{i=1}^{A-1} \lambda_i = \frac{|a|}{A}$$

where the  $\rightarrow$  *spanning tree* is a connected acyclic subgraph containing all the vertices of  $G$  [Trinajstić, Babic *et al.*, 1994]. The term  $a$  in the second equality is the coefficient of the linear term in the  $\rightarrow$  *Laplacian polynomial* [Nikolić, Trinajstić *et al.*, 1996b]. The number of spanning trees of a graph is used as a measure of  $\rightarrow$  *molecular complexity* for polycyclic graphs; it increases with the complexity of the molecular structure. It has to be noted that some specific algorithms have been proposed to calculate the number of spanning trees in molecular graphs of cata-condensed systems [John, Mallion *et al.*, 1998].

Moreover, the **spanning-tree density** (STD) and the **reciprocal spanning-tree density** (RSTD) were defined as [Mallion and Trinajstić, 2003]

$$\text{STD} = \frac{T^*}{{}^eN} \quad \text{STD} \leq 1 \quad \text{RSTD} = \frac{{}^eN}{T^*} \quad \text{RSTD} \geq 1$$

where  ${}^eN$  is the number of ways of choosing any  $A - 1$  edges belonging to the set  $\mathcal{E}(G)$  of graph edges. RSTD was proposed as a measure of *intricacy* of a graph, that is, the bigger RSTD is, the more intricate  $G$ .

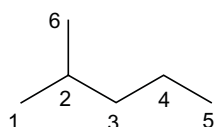
Also derived from the Laplacian matrix are the **Mohar indices**  $TI_1$  and  $TI_2$ , defined as

$$TI_1 = 2 \cdot A \cdot \log\left(\frac{B}{A}\right) \cdot \sum_{i=1}^{A-1} \frac{1}{\lambda_i} = 2 \cdot \log\left(\frac{B}{A}\right) \cdot W^* \quad TI_2 = \frac{4}{A \cdot \lambda_{A-1}}$$

where  $\lambda_{A-1}$  is the first nonzero eigenvalue and  $W^*$  the quasi-Wiener index [Trinajstić, Babić *et al.*, 1994]. Being  $W^* = W$  for acyclic graphs, the first Mohar index  $TI_1$  is closely related to the Wiener index for acyclic graphs.

### Example L1

Laplacian matrix  $L$ , its eigenvalues and some related indices for 2-methylpentane.



$$W^* = 6 \times 5.33305 = 31.998 \approx 32 = W$$

$$T^* = \frac{6.000}{6} = 1$$

$$TI_1 = -5.0676 \quad TI_2 = 2.0519$$

$L =$	Atom	1	2	3	4	5	6
	1	1	-1	0	0	0	0
	2	-1	3	-1	0	0	-1
	3	0	-1	3	-1	0	0
	4	0	0	-1	2	-1	0
	5	0	0	0	-1	1	0
	6	0	-1	0	0	0	1

ID	Eigenvalues
1	4.2143
2	3.0000
3	1.4608
4	1.0000
5	0.3249
6	0.0000

With some analogy to the Laplacian matrix is the **second path matrix**, denoted by  $S$  and defined as [John and Diudea, 2004]

$$S = A^2 - V$$

where  $A^2$  is the square adjacency matrix and  $V$  is the diagonal matrix of the vertex degrees.

The sum of the entries of the matrix  $S$  coincides with the  $\rightarrow$  *Platt number*  $F$  and twice the  $\rightarrow$  *Gordon–Scantlebury index*  $N_{GS}$ :

$$\sum_{i=1}^A \sum_{j=1}^A [S]_{ij} = F = 2 \cdot N_{GS}$$

Moreover, the **quasi-Euclidean matrix**, denoted as  $\rho_{qE}$ , was defined as [Ivanciuc, Ivanciuc *et al.*, 2001b; Zhu and Klein, 1996]

$$[\rho_{qE}]_{ij} = \begin{cases} [\Gamma^2]_{ii} + [\Gamma^2]_{jj} - 2 \cdot [\Gamma^2]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where  $\Gamma$  is the generalized inverse of the Laplacian matrix  $L$ . The off-diagonal elements  $[\rho_{qE}]_{ij}$  of this matrix are called **quasi-Euclidean distances**.

📖 [Ivanciuc, 1993; Gutman, Lee *et al.*, 1994; Nikolić, Trinajstić *et al.*, 1996b; Chan, Lam *et al.*, 1997; Xiao, 2004]

- **Laplacian polynomial** → characteristic polynomial-based descriptors
- **Laplacian spectrum** → Laplacian matrix
- **lateral validation** → validation techniques
- **lattice representation**  $\equiv$  *stereoelectronic representation* → molecular descriptors

■ **layer matrices (LM)** ( $\equiv$  *shell matrices*)

A layer matrix **LM** of a  $\rightarrow$  *molecular graph*  $G$  is a rectangular unsymmetrical matrix  $A \times (D + 1)$ ,  $A$  being the number of graph vertices and  $D$  the  $\rightarrow$  *topological diameter*. The entry  $i$ - $k$  ( $lm_{ik}$ ) is the sum of the weights of the vertices located in the concentric shell (layer) at  $\rightarrow$  *topological distance*  $k$  around the vertex  $v_i$  [Diudea, Minailiuc *et al.*, 1991; Diudea, 1994; Skorobogatov and Dobrynin, 1988]. The  $k$ th layer of the vertex  $v_i$  is the set  $V_{ik}(G)$  of vertices defined as the following:

$$V_{ik}(G) = \{a | a \in V(G); d_{ia} = k\}$$

where  $d_{ia}$  is the topological distance of the  $a$ th vertex from  $v_i$ .

The entries of the layer matrix are formally defined as

$$lm_{ik} = \sum_{a \in V_{ik}} w_a \quad \text{or} \quad lm_{ik} = \sum_{j=1}^A w_j \cdot \delta(d_{ij}; k)$$

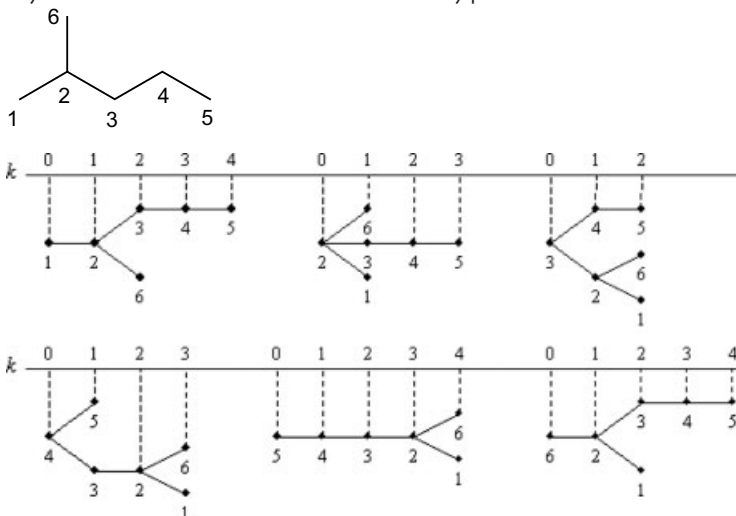
where  $w$  is the  $\rightarrow$  *weighting scheme* for graph vertices, and  $\delta(d_{ij}; k)$  the Dirac delta function equal to one when the distance  $d_{ij}$  between vertices  $v_i$  and  $v_j$  is equal to  $k$ , and zero otherwise.

The columns of the layer matrix are  $D + 1$ ,  $D$  being the topological diameter and the case  $k = 0$  is also considered, meaning that the property of the focused  $i$ th vertex is also considered.

The weights  $w$  for graph vertices can be any chemical or topological atomic properties. Examples of chemical  $\rightarrow$  *atomic properties* are  $\rightarrow$  *van der Waals volume*, atomic mass,  $\rightarrow$  *polarizability*; examples of  $\rightarrow$  *local vertex invariants* are  $\rightarrow$  *vertex degree*,  $\rightarrow$  *path degree*,  $\rightarrow$  *walk degree*.

### Example L2

Layer structures for the six vertices of 2-methylpentane.



Based on the different atomic properties, several layer matrices can be obtained; the most common are defined below.

• **cardinality layer matrix (LC)**

The simplest layer matrix obtained weighting all vertices by a weight equal to one [Diudea, 1994]. Therefore, the entry  $i$ - $k$  of the matrix is the number  $lc_{ik}$  of vertices located at distance  $k$  from the focused  $i$ th vertex. For the cardinality layer matrix, the following relations hold:

$$\sum_{k=0}^D lc_{ik} = \sum_{i=1}^A lc_{i0} = A \quad \text{and} \quad \sum_{i=1}^A lc_{i1} = 2 \cdot B$$

where  $A$  is the number of graph vertices,  $B$  the number of graph edges and  $D$  the molecule diameter that is the largest distance in the graph.

The cardinality layer matrix was originally called  **$\lambda$  matrix** [Skorobogatov and Dobrynin, 1988] and **F matrix** [Diudea and Pârnu, 1988]. Moreover, for acyclic graphs, the cardinality layer matrix coincides with the  $\rightarrow$  *path-layer matrix*.

By this matrix, the  $\rightarrow$  *information layer index* is calculated. Moreover, four different local vertex invariants derived from the cardinality layer matrix have been proposed [Wang, Milne *et al.*, 1994] as the following:

$$\begin{aligned} \gamma_i &= \log \left( \sum_{k=1}^D lc_{ik} \cdot 2^{-k} \right) & \gamma_i &= \log \left( \sum_{k=1}^D lc_{ik} \cdot 4^{-k} \right) \\ \gamma_i &= \log \left( \sum_{k=1}^D lc_{ik} / (k+1) \right) & \gamma_i &= \log \left( \sum_{k=1}^D lc_{ik} / (k+1)^{3/2} \right) \end{aligned}$$

where  $D$  is the maximum topological distance in the graph.

• **branching layer matrix (LB)**

A layer matrix obtained weighting all the vertices in the graph by their  $\rightarrow$  *vertex degrees*  $\delta$  [Diudea, Minailiuc *et al.*, 1991]. Therefore, the entry  $i$ - $k$  of the matrix is the sum of the vertex degrees over all vertices in the  $k$ th layer around the focused  $i$ th vertex:

$$lb_{ik} = \sum_{j=1}^A \delta_j \cdot \delta(d_{ij}; k)$$

where  $\delta_j$  is the vertex degree of the  $j$ th atom,  $d_{ij}$  is the distance between vertices  $v_i$  and  $v_j$ , and  $\delta(d_{ij}, k)$  the Dirac delta function equal to one when the distance  $d_{ij}$  is equal to  $k$  and zero otherwise. For acyclic graphs, the elements of the branching layer matrix coincide with the  $\rightarrow$  *valence shells* of the graph vertices proposed by Randić to compute weighted path counts [Randić, 2001g].

Each  $i$ th row of the matrix **LB** expresses the global state of vertex degrees from the viewpoint of vertex  $i$ , that is, the distribution of sums of vertex degrees in shells around the  $i$ th vertex. The

sum of each  $i$ th row element is a constant equal to  $2B$ , twice the number of graph edges. Moreover, more branched and  $\rightarrow$  *central vertices* show higher values in the first layers within the corresponding rows, while less branched and  $\rightarrow$  *terminal vertices* have higher values in the far layers. The vertex degrees of the atoms are in the first column ( $k = 0$ ).

- **connectivity valence layer matrix (LCV)**

Similar to the branching layer matrix, this matrix is obtained weighting the vertices by their  $\rightarrow$  *valence vertex degree*  $\delta^v$  instead of the simple vertex degree  $\delta$  [Hu and Xu, 1996].

- **edge layer matrix (LE)**

Analogously to the branching layer matrix, the edge layer matrix **LE** is obtained weighting all vertices by the number of distinct edges incident to the vertices of the  $k$ th layer around the vertex  $v_i$ , without counting any edge already counted in a preceding layer [Diudea, Minailiuc *et al.*, 1991]. The sum of the elements of each  $i$ th row is a constant equal to the number of edges  $B$ , while the sum of the elements of each  $k$ th column is a constant equal to  $2B$ . The  $\rightarrow$  *vertex degrees* of the atoms are in the first column ( $k = 0$ ).

- **connectivity bond layer matrix (LCB)**

A layer matrix whose entry  $i-k$  is defined as the sum of the  $\rightarrow$  *conventional bond order*  $\pi^*$  of the edges connecting the vertices situated in the  $k$ th layer with the vertices of the  $(k - 1)$ th layer with respect to the focused  $i$ th vertex [Hu and Xu, 1996].

- **sum layer matrix (LS)**

To increase the discriminating power of atomic and molecular descriptors derived from the layer matrices, the sum layer matrix **LS** was also defined, and its entries are the sums of the corresponding entries of the branching layer matrix **LB** ( $lb_{ik}$ ) and edge layer matrix **LE** ( $le_{ik}$ ) [Diudea, Minailiuc *et al.*, 1991]:

$$ls_{ik} = lb_{ik} + le_{ik}$$

The sum of the elements of each  $i$ th row is a constant equal to  $3B$ ,  $B$  being the number of graph edges.

- **distance sum layer matrix (LDS)**

A layer matrix obtained weighting the vertices by their  $\rightarrow$  *vertex distance degree*  $\sigma$ , that is, the row sum of the  $\rightarrow$  *distance matrix* **D** [Balaban and Diudea, 1993]. Therefore, the entry  $i-k$  of the layer matrix is the sum of the vertex distance degrees of the vertices located at distance  $k$  from the focused  $i$ th vertex. It is obvious that the entries of the first column ( $k = 0$ ) are only the vertex distance degrees. Moreover, the sums over each row in **LDS** are all equal to twice the  $\rightarrow$  *Wiener index*, that is, the following relation holds:

$$\sum_{k=0}^D ls_{ik} = \sum_{i=1}^A ls_{i0} = 2 \cdot W$$

where  $A$  is the number of graph vertices and  $W$  the Wiener index.

- **geometric sum layer matrix (LGS)**

A layer matrix defined by analogy with the distance sum layer matrix deriving the weighting scheme for vertices from the  $\rightarrow$  *geometry matrix* **G** instead of the  $\rightarrow$  *distance matrix* **D** [Diudea, Horvath *et al.*, 1995b]. Therefore, the entry  $i-k$  of this layer matrix is the sum of the  $\rightarrow$  *geometric distance degrees*  $^G\sigma$  of the vertices located at distance  $k$  from the focused  $i$ th vertex, where the geometric distances are obtained by methods of  $\rightarrow$  *computational chemistry*. The sums over each row in **LGS** as well as the zero-column sum are all equal to twice the  $\rightarrow$  *3D-Wiener index*.

- **path degree layer matrix (LPD)**

A layer matrix obtained weighting all vertices by the  $\rightarrow$  *path degree*  $\xi$ . Therefore, the entry  $i-k$  of the matrix is the sum of the path degrees of all vertices located at distance  $k$  from the focused  $i$ th vertex. The half-sum of the elements in the first column ( $k=0$ ) is equal to the half-sum of the elements in each row of matrix **LPD** and corresponds to a molecular descriptor repurposed as the  $\rightarrow$  *all-path Wiener index*  $W^{AP}$ , that is, the following relations hold:

$$\sum_{k=0}^D lpd_{ik} = \sum_{i=1}^A lpd_{i0} = 2 \cdot W^{AP}$$

where  $A$  is the number of graph vertices.

Note that for acyclic graphs, the path degree layer matrix **LPD** coincides with the distance sum layer matrix **LDS**.

- **walk degree layer matrix ( $LW^{(m)}$ )**

A layer matrix obtained weighting the vertices by their  $\rightarrow$  *walk degree* (i.e., the  $\rightarrow$  *atomic walk count* of length  $m$ ,  $awc_i^{(m)}$ ). Therefore, the entry  $i-k$  of the layer matrix ( $lw_{ik}$ ) is the sum of the walk degrees of the vertices located at distance  $k$  from the focused  $i$ th vertex [Diudea, Topan *et al.*, 1994]. Different matrices  $LW^{(m)}$  can be calculated according to the chosen order  $m$  of the walk degrees  $awc_i^{(m)}$ . The walk degree layer matrix  $LW^{(1)}$  coincides with the branching layer matrix **LB**. The elements of the first column ( $k=0$ ) in the matrix  $LW^{(m)}$  represent only the walk degrees of order  $m$ .

Moreover, the half-sum of both the entries in the first column ( $k=0$ ) and in each row is the total number of walks of length  $m$  in the graph, as can be seen from the following relationships:

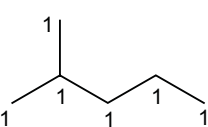
$$\frac{1}{2} \cdot \sum_{k=0}^D lw_{ik}^{(m)} = \frac{1}{2} \cdot \sum_{i=1}^A lw_{i0}^{(m)} = \frac{1}{2} \cdot \sum_{i=1}^A awc_i^{(m)} = mwc^{(m)}$$

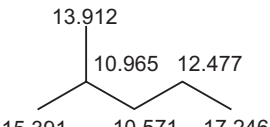
where  $A$  is the number of graph vertices,  $awc_i^{(m)}$  the atomic walk count, and  $mwc^{(m)}$  is the  $m$ th order  $\rightarrow$  *molecular walk count*.

Analogous layer matrices can be obtained using, as the weighting scheme,  $\rightarrow$  *weighted walk degrees* instead of the simple vertex walk degrees.

**Example L3**

Cardinality layer matrix **LC** and geometric sum layer matrix **LGS** for 2-methylpentane. Matrix rows represent vertices (*i*), while matrix columns represent layers (*k*). Vertices in the graphs are labeled according to the unitary weighting scheme for cardinality layer matrix and geometric distance degrees for geometric sum layer matrix.





LC =

<i>i/k</i>	0	1	2	3	4
1	1	1	2	1	1
2	1	3	1	1	0
3	1	2	3	0	0
4	1	2	1	2	0
5	1	1	1	1	2
6	1	1	2	1	1

LGS =

<i>i/k</i>	0	1	2	3	4
1	15.391	10.955	24.483	12.477	17.246
2	10.955	39.874	12.477	17.246	0
3	10.571	23,432	46.549	0	0
4	12.477	27.817	10.955	29.303	0
5	17.246	12.477	10.571	10.955	29.303
6	13.912	10.955	25.962	12.477	17.246

Derived from layer matrices, two main types of  $\rightarrow$  local vertex invariants are defined on the basis of two types of operators, the **centric operator**  $c_i$  and the **centrocomplexity operator**  $x_i$  [Diudea, 1994; Diudea, Horvath *et al.*, 1995b]:

$$c_i(\mathbf{LM}) = \left[ \sum_{k=1}^D (lm_{ik})^{k/d} \right]^{-1} \quad \text{and} \quad x_i(\mathbf{LM}) = \left[ \frac{1}{w_i} \cdot \sum_{k=0}^D lm_{ik} \cdot 10^{-zk} \pm l_i \right]^{\pm 1} \cdot t_i$$

where

$$l_i = f_i \cdot \left( \frac{lm_{i0}}{10} + \frac{lm_{i1}}{100} \right) \quad \text{and} \quad f_i = \sum_{j=1}^A a_{ij} \cdot (\pi_{ij}^* - 1) \quad \pi_{ij}^* = 0 \text{ if } (i, j) \notin E(\mathcal{G})$$

where  $d$  is a specified topological distance larger than the topological diameter  $D$  (for example, 10);  $w_i$  is the atomic property of the  $i$ th vertex,  $z$  is the number of figures of the largest entry  $lm_{ik}$  value;  $l_i$  is a local parameter accounting for multiple bonds,  $f_i$  being the  $\rightarrow$  atomic multigraph factor obtained by summing up the conventional bond orders  $\pi_{ij}^*$  of the vertices  $j$  adjacent to the  $i$ th vertex;  $t_i$  is a weighting factor accounting for heteroatoms by means of atomic numbers, electronegativities, covalent radii, and so on.

A third type of operator generating local invariants from layer matrices is defined as follows [Balaban and Diudea, 1993]:

$$x_{ji}(\mathbf{LM}) = \sum_{j=1}^A a_{ij} \cdot \left[ \frac{\sum_{k=0}^D lm_{ik} \cdot 10^{-zk}}{t_i \cdot (1 + f_i)} \cdot \frac{\sum_{k=0}^D lm_{jk} \cdot 10^{-zk}}{t_j \cdot (1 + f_j)} \right]^{-1/2}$$

where the sum runs over all vertices  $j$  adjacent to the  $i$ th vertex, that is, located at distance one from  $v_i$ .



Centrocomplexity invariant values should measure the location of the considered vertex with respect to a vertex “of importance”, that is, a vertex of highest branching degree, electro-negativity, and so on, while centric invariant values should measure the centrality of a vertex, that is, its location with respect to the  $\rightarrow$  graph center.

The local indices obtained by applying centrocomplexity operator to the branching layer matrix **LB** are called **regressive vertex degrees** [Diudea, Minailiuc *et al.*, 1991]. Such indices are an extension of the concept of  $\rightarrow$  vertex degree, taking into account the contributions of distant vertices from the focused  $i$ th vertex; these contributions decrease with increasing distance, slightly augmenting the value of the classical vertex degree. High values of regressive vertex degrees should correspond to vertices of highest degree, closest to branching sites and to the graph center. The two types of originally proposed centrocomplexity operators were defined as

$$RVD_i \equiv x_i^{R1}(\mathbf{LB}) = \sum_{k=0}^D lb_{ik} \cdot k^{-3} = \delta_i + \sum_{k=1}^D lb_{ik} \cdot k^{-3}$$

$$RVD_i \equiv x_i^{R2}(\mathbf{LB}) = \sum_{k=0}^D lb_{ik} \cdot 10^{-k} = \delta_i + \sum_{k=1}^D lb_{ik} \cdot 10^{-k}$$

where **LB** is the branching layer matrix and  $\delta$  the vertex degree.

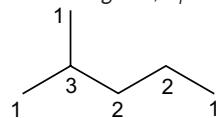
#### Example L4

Branching layer matrix **LB** and regressive vertex degrees for 2-methylpentane.

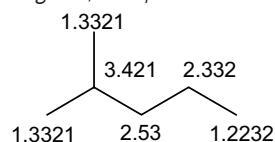
**LB** =

$i/k$	0	1	2	3	4
1	1	3	3	2	1
2	3	4	2	1	0
3	2	5	3	0	0
4	2	3	3	2	0
5	1	2	2	3	2
6	1	3	3	2	1

Vertex degrees,  $\delta_i$



Regressive vertex degrees,  $RVD_i$



$$RVD_1 = RVD_6 = \sum_{k=0}^4 lb_{1k} \cdot 10^{-k} = 1 + 0.3 + 0.03 + 0.002 + 0.0001 = 1.3321$$

$$RVD_2 = \sum_{k=0}^4 lb_{2k} \cdot 10^{-k} = 3 + 0.4 + 0.02 + 0.001 = 3.421$$

$$RVD_3 = \sum_{k=0}^4 lb_{3k} \cdot 10^{-k} = 2 + 0.5 + 0.03 = 2.53$$

$$RVD_4 = \sum_{k=0}^4 lb_{4k} \cdot 10^{-k} = 2 + 0.3 + 0.03 + 0.002 = 2.332$$

$$RVD_5 = \sum_{k=0}^4 lb_{5k} \cdot 10^{-k} = 1 + 0.2 + 0.02 + 0.003 + 0.0002 = 1.2232$$

By analogy with the regressive vertex degrees, the **regressive distance sums** (or **regressive incremental distance sums**) are local invariants obtained by applying the centrocomplexity operator  $x_i$  to the distance sum layer matrix **LDS** [Balaban and Diudea, 1993]. Also in this case, a simplified form of the centrocomplexity operator has been proposed:

$$RDS_i \equiv x_i(\mathbf{LDS}) = \sum_{k=0}^D (lds_{ik} \cdot 10^{-zk}) = \sigma_i + \sum_{k=1}^D (lds_{ik} \cdot 10^{-zk})$$

where  $\sigma_i$  is the distance sum (i.e., vertex distance sum) of the  $i$ th vertex and  $z$  is the number of figures of the largest entry  $lds_{ik}$  value.

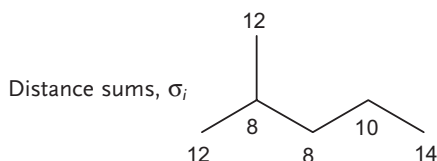
To obtain greater discrimination between the terminal and central vertices, the **regressive decremental distance sums** were proposed [Balaban, 1995b]. They are calculated from the distance sum layer matrix **LDS** by the following:

$$RDDS_i = \sigma_i - \sum_{k=1}^D (lds_{ik} \cdot 10^{-zk})$$

where  $\sigma_i$  is the  $\rightarrow$  distance sum of the  $i$ th vertex. In this way, the progressively attenuated contributions due to more distant vertices are subtracted from the distance degree of the focused vertex.

#### Example L5

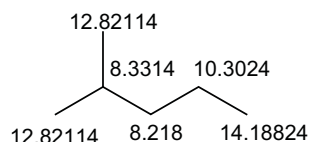
Distance sum layer matrix **LDS**, regressive distance sums, and regressive decremental distance sums for 2-methylpentane.



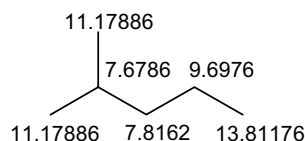
**LDS** =

$i/k$	0	1	2	3	4
1	12	8	20	10	14
2	8	32	10	14	0
3	8	18	38	0	0
4	10	22	8	24	0
5	14	10	8	8	24
6	12	8	20	10	14

Regressive distance sums,  $RDS_i$



Regressive decremental distance sums,  $RDDS_i$



$$RDS_1 = RDS_6 = \sum_{k=0}^4 lds_{1k} \cdot 10^{-k} = 12 + 0.8 + 0.020 + 0.0010 + .00014 = 12.82114$$

$$RDS_2 = \sum_{k=0}^4 lds_{2k} \cdot 10^{-k} = 8 + 0.32 + 0.010 + 0.0014 = 8.3314$$

$$\begin{aligned}
 RDS_3 &= \sum_{k=0}^4 lds_{3k} \cdot 10^{-k} = 8 + 0.18 + 0.038 = 8.218 \\
 RDS_4 &= \sum_{k=0}^4 lds_{4k} \cdot 10^{-k} = 10 + 0.22 + 0.08 + 0.0024 = 10.3024 \\
 RDS_5 &= \sum_{k=0}^4 lds_{5k} \cdot 10^{-k} = 14 + 0.10 + 0.08 + 0.008 + 0.00024 = 14.18824 \\
 RDDS_1 &= RDDS_6 = \sigma_1 - \sum_{k=0}^4 lds_{1k} \cdot 10^{-k} = 12 - 0.8 - 0.020 - 0.0010 - 0.00014 = 11.17886 \\
 RDDS_2 &= \sigma_2 - \sum_{k=0}^4 lds_{2k} \cdot 10^{-k} = 8 - 0.32 - 0.010 - 0.0014 = 7.6786 \\
 RDDS_3 &= \sigma_3 - \sum_{k=0}^4 lds_{3k} \cdot 10^{-k} = 8 - 0.18 - 0.038 = 7.8162 \\
 RDDS_4 &= \sigma_4 - \sum_{k=0}^4 lds_{4k} \cdot 10^{-k} = 10 - 0.22 - 0.08 - 0.0024 = 9.6976 \\
 RDDS_5 &= \sigma_5 - \sum_{k=0}^4 lds_{5k} \cdot 10^{-k} = 14 - 0.10 - 0.08 - 0.00024 = 13.81176
 \end{aligned}$$

The sums of the local vertex invariants  $x_i$  and  $c_i$  over all of the vertices give the corresponding molecular descriptors, called **centrocomplexity topological index**  $X$  and **centric topological index**  $C$ , respectively:

$$X(\mathbf{LM}) = \sum_{i=1}^A x_i(\mathbf{LM}) \quad \text{and} \quad C(\mathbf{LM}) = \sum_{i=1}^A c_i(\mathbf{LM})$$

where  $\mathbf{LM}$  represents any layer matrix. These descriptors are related to  $\rightarrow$  *molecular complexity*. Normalized centrocomplexity and centric local invariants  $x'_i$  and  $c'_i$  are obtained dividing each local invariant by the corresponding global topological index.

📖 [Diudea and Bal, 1990; Diudea and Kacso, 1991, 1992; Dobrynin, 1993; Ivanciuc, Balaban *et al.*, 1993b, 1992; Wang, Milne *et al.*, 1994; Dobrynin and Mel'nikov, 2001; Diudea, Jäntschi *et al.*, 2002; Diudea, 2002a; Diudea and Ursu, 2003; Konstantinova and Vidyuk, 2003]

- **LCD descriptors**  $\rightarrow$  molecular descriptors ( $\odot$  invariance properties of molecular descriptors)
- **LDOS**  $\equiv$  *Local Density Of States*  $\rightarrow$  quantum-chemical descriptors ( $\odot$  EIM descriptors)
- **LEACH index**  $\rightarrow$  environmental indices ( $\odot$  leaching indices)
- **leaching indices**  $\rightarrow$  environmental indices
- **lead compound**  $\rightarrow$  drug design
- **leading eigenvalue**  $\rightarrow$  spectral indices
- **leading eigenvalue of the distance matrix**  $\rightarrow$  spectral indices ( $\odot$  eigenvalues of the distance matrix)
- **leading eigenvalue of (A + D)**  $\rightarrow$  spectral indices

- **lead-like indices** → property filters
- **learning set**  $\equiv$  *training set* → data set
- **leave-more-out technique** → validation techniques (⊙ cross-validation)
- **leave-one-out technique** → validation techniques (⊙ cross-validation)
- **length-to-breadth ratio** → shape descriptors (⊙ Kaliszan shape parameter)
- **Lennard–Jones 6–12 potential function** → molecular interaction fields (⊙ steric interaction fields)
- **Leo–Hansch hydrophobic fragmental constants** → lipophilicity descriptors
- **Lethal Concentration** → biological activity indices (⊙ toxicological indices)
- **Lethal Dose** → biological activity indices (⊙ toxicological indices)
- **Lethal Time** → biological activity indices (⊙ toxicological indices)
- **level pattern indices** → spectral indices (⊙ eigenvalues of the adjacency matrix)
- **leverage matrix** → chemometrics (⊙ regression analysis)
- **lhaf(D) index** → algebraic operators (⊙ determinant)
- **ligand** → drug design
- **L index** → combined descriptors
- **$L_Z$  index** → Hosoya Z matrix
- **LIN index** → environmental indices (⊙ leaching indices)
- **linear aromatic substituent reactivity relationships**  $\equiv$  *Yukawa–Tsuno equation* → electronic substituent constants (⊙ resonance electronic constants)
- **Linear Free Energy Relationships** → extrathermodynamic approach
- **linear graph** → graph
- **linear indices** → TOMOCOMD descriptors
- **linearity index** → shape descriptors
- **linear notation systems** → molecular descriptors
- **linear similarity index** → quantum similarity

### ■ Linear Solvation Energy Relationships (LSERs)

Linear solvation energy relationships constitute the basis on which effects of solvent–solute interactions on → *physico-chemical properties* and reactivity parameters are studied. In general, a property  $\mathcal{P}$  of a species A in a solvent S can be expressed as

$$\mathcal{P}_{A,S} = \sum_j \varphi_j(A, S)$$

where  $\varphi$  are complex functions of both solvents and solutes [Kamlet, Abboud *et al.*, 1981]. By assuming that these functions can be factorized into two contributions separately dependent on solute and solvent, the property can be represented as

$$\mathcal{P}_{A,S} = \sum_j f_j(A) \cdot g_j(S)$$

where  $f$  are functions of the solute and  $g$  functions of the solvent.

The underlying philosophy of the linear solvation energy relationships is based on the possibility to study these two functions, after a proper choice of the reference systems and properties. Moreover, it has been recognized that solution properties  $\mathcal{P}$  mainly depend on three factors: a cavity term, a polar term, and hydrogen-bond term:

$$P = \text{intercept} + \text{cavity term} + \text{dipolarity/polarizability term} + \text{hydrogen-bond term}$$

Therefore, a typical linear solvation energy relationship is expressed as [Kamlet, Doherty *et al.*, 1987a]

$$P_{A,S} = b_0 + b_1 \cdot (\delta_H^2)_1 \cdot V_2 + b_2 \cdot \Pi_1^* \cdot \Pi_2^* + b_3 \cdot \alpha_1 \cdot \beta_2 + b_4 \cdot \beta_1 \cdot \alpha_2$$

where  $b$  are estimated regression coefficients, and the subscripts 1 and 2 in the solvent/solute property parameters refer to the solvent S and the solute A, respectively. This equation is usually known as **Abraham's general equation** or **solvatochromic equation** even if it is extended to cover some nonspectroscopic properties, and the parameters of polarity/dipolarizability and hydrogen-bonding as **solvatochromic parameters**. The term *solvatochromic* is derived from the origin of this approach referring to the effect solvent has on the color of an indicator that is used for quantitative determination of some molecular attributes (*solvatochromic parameters*).

From the general solvatochromic equation, two special cases can be encountered. When dealing with effects of different solvents on properties of a specific solute, the general equation is explicitly on solvent parameters:

$$P_{A,S_i} = b_0 + b_1 \cdot (\delta_H^2)_i + b_2 \cdot \Pi_{1,i}^* + b_3 \cdot \alpha_{1,i} + b_4 \cdot \beta_{1,i}$$

This equation has been used in several correlations of solvent effects on solute properties such as reaction rates and equilibrium constants of solvolyses, energy of electronic transitions, solvent induced shifts in UV/visible, IR, and NMR spectroscopy, fluorescence lifetimes, formation constants of hydrogen-bonded and Lewis acid/base complexes [Kamlet, Doherty *et al.*, 1986c].

Conversely, when dealing with solubilities, lipophilicity, or other properties of a set of different solutes in a specific solvent, the general equation is explicitly on the solute parameters:

$$P_{A_i,S} = b_0 + b_1 \cdot V_i + b_2 \cdot \Pi_{2,i}^* + b_3 \cdot \alpha_{2,i} + b_4 \cdot \beta_{2,i}$$

This equation has been mainly used in correlations of aqueous solubility of compounds, octanol/water partition coefficients and some other  $\rightarrow$  *partition coefficients* together with some biological properties [Kamlet, Abraham *et al.*, 1984; Kamlet, Doherty *et al.*, 1986a, 1987a, 1987c, 1988c].

Other two general linear solvation energy relationships for solute physico-chemical properties in a fixed phase [Abraham, Whiting *et al.*, 1990b, 1991a, 1991b; Abraham, 1993d; Abraham, Andonian-Haftvan *et al.*, 1994] are

$$\log(P_{A_i,S}) = b_0 + b_1 \cdot V_{X,i} + b_2 \cdot R_{2,i} + b_3 \cdot \pi_{2,i}^H + b_4 \cdot \alpha_{2,i}^H + b_5 \cdot \beta_{2,i}^H$$

$$\log(P_{A_i,S}) = b_0 + b_1 \cdot L_i^{16} + b_2 \cdot R_{2,i} + b_3 \cdot \pi_{2,i}^H + b_4 \cdot \alpha_{2,i}^H + b_5 \cdot \beta_{2,i}^H$$

where the first can be applied to processes within condensed phases and the second to processes involving gas-condensed phase transfer. The solute descriptors in these relationships are called **Abraham-Klamt descriptors** and were successively denoted as [Zissimos, Abraham *et al.*, 2002c]

$$A \equiv \alpha^H; \quad B \equiv \beta^H; \quad S \equiv \pi^H; \quad E \equiv R; \quad V \equiv V_X$$

Similar equations but based on other solute descriptors were proposed in literature with the aim of better chromatographic data [Abraham, Ibrahim *et al.*, 2004]. In particular, five solute descriptors, here called **Laffort solute descriptors** (Table L1), were defined by Laffort *et al.* using GLC retention data on five stationary phases for 240 compounds [Laffort and Patte, 1976; Patte, Etcheto *et al.*, 1982]. These solute descriptors were used to fit a number of physico-chemical and biochemical properties. Note that in the first paper [Laffort and Patte, 1976], the five solute descriptors were obtained by  $\rightarrow$  *Principal Component Analysis* on the data obtained from 25 stationary phases for 75 compounds, thus their numerical values differ from those obtained in the later paper.

Other five solute descriptors, here called **Wilson solute descriptors** (Table L1), were proposed by Wilson *et al.* using ten different HPLC stationary phases, all with acetonitrile 50% as the mobile phase, initially for 67 compounds and then extended to a larger class of compounds [Wilson, Nelson *et al.*, 2002a, 2002b; Wilson, Dolan *et al.*, 2002].

**Weckwerth solute descriptors** (Table L1) are five solute parameters based on  $\rightarrow$  *Kovats retention index* on seven of GC stationary phases for 53 compounds [Weckwerth, Vitha *et al.*, 2001].

**Table L1** Laffort, Wilson, and Weckwerth solute descriptors.

Laffort solute descriptors	Wilson solute descriptors	Weckwerth solute descriptors
Volume-sensitive apolar factor ( $\alpha$ )	Hydrophobicity ( $\eta'$ )	Solute volume ( $V$ )
Orientation factor, proportional to dipole moment of simple molecules ( $\omega$ )	Steric parameter ( $\sigma'$ )	Solute polarizability ( $P$ )
Electron factor, related to dispersion interactions ( $\epsilon$ )	Basicity ( $\beta'$ )	Solute dipolarity ( $D$ )
Hydrogen bond acidity ( $\pi$ )	Acidity ( $\alpha'$ )	Hydrogen bond acidity ( $A$ )
Hydrogen bond basicity ( $\beta$ )	Cation-exchange parameter ( $\kappa'$ )	Hydrogen bond basicity ( $B$ )

The descriptors of the solvatochromic equation are specified below.

- **cavity term**

The cavity term is a measure of the endoergic cavity-forming process, that is, the free energy necessary to separate the solvent molecules, overcoming solvent–solvent cohesive interactions, and provide a suitably size cavity for the solute. The magnitude of the cavity term depends on the  $\rightarrow$  *Hildebrand solubility parameter*  $\delta_H$  and  $\rightarrow$  *volume descriptors* of the solute. The solute volume can be measured in different ways, such as by  $\rightarrow$  *van der Waals volume*  $V^{vdw}$  [Leahy, 1986],  $\rightarrow$  *molar volume*  $\bar{V}$  or  $\rightarrow$  *Mc Gowan's characteristic volume*  $V_X$ ; in some cases, also  $\rightarrow$  *molecular weight*  $MW$  has been used. Usually, the volumes are divided by 100 ( $\bar{V}/100$ ) to obtain a more homogeneous scale with respect to the other parameters.

The parameter  $L^{16}$  is the  $\rightarrow$  *Ostwald solubility coefficient* on *n*-hexadecane at 298 K; it includes both general dispersion interactions and endoergic cavity term and was proposed for modeling properties of solutes in processes involving gas-condensed phase transfer such as gas-liquid chromatographic parameters [Abraham, Grellier *et al.*, 1987].

• **dipolarity/polarizability term** ( $\equiv$  *dipole term*)

This term is a measure of the exoergic balance (i.e., release of energy) of solute–solvent and solute–solute dipolarity/polarizability interactions. This term, denoted by  $\Pi^*$ , describes the ability of the compound to stabilize a neighboring charge or dipole by virtue of nonspecific dielectric interactions and is in general given by  $\rightarrow$  *electric polarization descriptors* such as  $\rightarrow$  *dipole moment* or other empirical  $\rightarrow$  *polarity/polarizability descriptors* [Abraham, Grellier *et al.*, 1988]. Other specific polarity parameters empirically derived for linear solvation energy relationships are reported below.

Several **solvent polarity scales** were proposed to quantify the polar effects of solvents on physical properties and reactivity parameters in solution, such as rate of solvolyses, energy of electronic transitions, solvent induced shifts in IR, or NMR spectroscopy. Most of the polarity scales were derived by an empirical approach based on the principles of the  $\rightarrow$  *linear free energies relationships* applied to a chosen reference property and system where hydrogen bonding effects are assumed negligible [Reichardt, 1965, 1990; Kamlet, Abboud *et al.*, 1981, 1983].

The most important scales of solvent polarity are:

**$\pi^*$  polarity scale.** A solvent polarity parameter (also denoted as  $\pi^*_s$ ) proposed by Kamlet, Abboud and Taft [Kamlet, Abboud *et al.*, 1977; Kamlet, Carr *et al.*, 1981], based on the solvatochromic shifts on the frequency maxima of the  $\pi \rightarrow \pi^*$  transitions of seven different benzene derivatives. The  $\pi^*$  values are averaged on the seven compounds to prevent the inclusion of specific effects or spectral anomalies and are normalized so that  $\pi^*$  equals zero for cyclohexane and one for dimethylsulfoxide. This scale is one of the most comprehensive for the number of considered solvents and is widely used. Moreover, for solution properties  $\phi$  involving different relative contributions of polarity and polarizability,  $\pi^*$  values can be corrected as  $(\pi^* - d\delta_H)$ , where  $\delta_H$  is the  $\rightarrow$  *Hildebrand solubility parameter*. The term  $d$  is calculated by dividing the difference in  $\phi$  at  $\pi^* = 0.7$ , as obtained separately for nonpolychlorinated aliphatic and aromatic solvents, by the average of the slopes of the solvatochromic equations for aliphatic and aromatic solvents. The Hildebrand parameter is assumed to be  $\delta_H = 0.0$  for nonpolychlorinated aliphatic solutes,  $\delta_H = 0.5$  for polychlorinated aliphatics, and  $\delta_H = 1.0$  for aromatic solutes; the term  $d$  ranges from zero, for maximal polarizability contributions to the studied property, to  $-0.40$  for minimal contributions. The  $\rightarrow$  *excess molar refractivity*  $R_2$  has also been used as polarizability correction term instead of the Hildebrand parameter  $\delta_H$  [Abraham, Lieb *et al.*, 1991].

**Y polarity scale.** A solvent polarity scale proposed by Grunwald and Winstein [Grunwald and Winstein, 1948] based on solvolytic rate  $k_0$  of *t*-butyl chloride in 80% aqueous ethanol at 298 K. The Y-polarity value for a given solvent is calculated by

$$Y = \log k_s - \log k_0$$

where  $k_s$  is the solvolytic rate of *t*-butyl chloride in the considered solvent. Y scale was proposed as measure of an empirical “ionizing power” of solvents.

**$E_T$  polarity scale.** A solvent polarity scale proposed by Dimroth, Reichardt, and coworkers [Dimroth, Reichardt *et al.*, 1963; Reichardt, 1965] based on the solvatochromic band shifts of the 4-(2,4,6-triphenylpyridinium)-2,6-diphenylphenoxide and its trimethyl derivative. This scale is one of the most comprehensive for the number of considered solvents and is widely used.

**$E_K$  polarity scale.** A solvent polarity scale proposed by Walther [Walther, 1974] based on the hypsochromic shift of the longest wavelength absorption of a molybdenum complex.

**Z polarity scale.** A solvent polarity scale proposed by Kosower [Kosower 1958a, 1958b] based on the energy of the electronic transition of the 1-ethyl-4-carbomethoxypyridinium iodide that is strongly solvent-dependent. This is a measure of an internal charge transfer process. The original set of Z values being quite small, it was successively extended by means of other indicators (Table L2).

**Table L2** Empirical parameters of the solvent polarity from different sources.

Solvent	$Y$	$Z$	$E_T$	$E_K$	$\pi^*$
H <sub>2</sub> O	3.493	94.6	63.1		1.09
HCOOH	2.054				0.65
HCONH <sub>2</sub>	0.604	83.3	56.6		
CH <sub>3</sub> COOH	−1.64	79.2	51.1	55.0	0.62
CH <sub>3</sub> OH	−1.09	83.6	55.5	56.3	0.60
CH <sub>3</sub> NO <sub>2</sub>			46.3		0.85
CH <sub>3</sub> CN		71.3	46.0		0.75
C <sub>2</sub> H <sub>5</sub> OH	−2.03	79.6	51.9	55.3	0.54
C <sub>5</sub> H <sub>5</sub> N		64.0	40.2	57.0	0.87
CCl <sub>4</sub>			32.5	49.9	0.294
CH <sub>2</sub> Cl <sub>2</sub>		64.2	41.1	53.9	0.802
C <sub>6</sub> H <sub>6</sub>		54.0	34.5	53.4	0.588
C <sub>6</sub> H <sub>5</sub> Cl			37.5	53.9	0.709
C <sub>6</sub> H <sub>5</sub> Br			37.5	53.9	0.794
C <sub>6</sub> H <sub>5</sub> CN			42.0		0.933
C <sub>6</sub> H <sub>5</sub> NO <sub>2</sub>			42.0		1.006
Cyclo-C <sub>6</sub> H <sub>12</sub>			31.2	49.0	0.000

A wide variety of correlations among solvent polarity scales was studied [Reichardt and Dimroth, 1968]; however, because of the different reference compounds used to define them, direct comparison should be done with caution [Bentley and von Schleyer, 1977].

The **solute polarity parameter**  $\pi_2^*$  was originally taken as identical with the solvent polarity parameter  $\pi^*$  for nonassociated liquids only [Taft, Abraham *et al.*, 1985]. Then, an alternative solute polarity parameter  $\pi_2^H$  (or  $\sum \pi_2^H$ ) was proposed, which was based on experimental procedures that include, at least in principle, all types of solute molecules [Abraham, Whiting *et al.*, 1991a; Abraham and Whiting, 1992]. Values of this solute parameter were determined by back-calculation solving “inverse” solvation equation systems based on 30–70 stationary phases for each solute.  $\pi_2^H$  values refer to a situation in which a solute molecule is surrounded by an excess of solvent molecules and so they are effective values, more correctly denoted as  $\sum \pi_2^H$ , accounting for combined effects due to polyfunctional groups in the molecule.

#### • hydrogen-bond parameters

They are measures of the exoergic effects (i.e., release of energy) of the complexation between solutes and solvents.

The  $\rightarrow$  *hydrogen bond donor* (HBD) power of a compound is called **hydrogen bond acidity** (or **hydrogen-bond electron-drawing power**) and is denoted by  $\alpha_1$  and  $\alpha_2$  for solvents and solutes, respectively.



The most important scales for **solvent HBD acidity** are here reported.

The  $\alpha$  scale proposed to measure solvent hydrogen bond acidity, that is, the ability of a bulk solvent to act as hydrogen bond donor toward a solute, was derived from 16 diverse properties involving 13 solutes as averaged values [Taft and Kamlet, 1976; Kamlet, Abboud *et al.*, 1983].

The **Gutmann's Acceptor Number (AN)** was proposed [Gutmann, 1978] as quantitative empirical parameter of solvent hydrogen bond acidity based on  $^{31}\text{P}$ -NMR shifts of triethylphosphine oxide at infinite dilution, calculated as  $\text{AN} = -\delta_{\infty}^{\text{corr}} \cdot 2.349$ .

For **solute HBD acidity**, different scales were proposed mainly based on complexation constants and enthalpies of complexation. The most important are here reported.

The  $\alpha_{\text{m}}$  scale was proposed for solute HBD acidity of "monomer" amphihydrogen-bonding compounds acting as non-self-associated solutes [Taft, Abraham *et al.*, 1985; Kamlet, Doherty *et al.*, 1986a]. In particular,  $\alpha_{\text{m}}$  values were derived from  $\log K$  values for complexation with pyridine *N*-oxide in cyclohexane; this set of values was successively extended through various back-calculations using the solvatochromic equation.

The  $\alpha_2^{\text{H}}$  scale was proposed for solute HBD acidity based on  $\log K$  values for 1:1 complexation of series of acids against a given base in dilute solution of  $\text{CCl}_4$  [Abraham, Grellier *et al.*, 1989]. Forty-five linear equations were solved for each considered base by a series of acids:

$$\log K_i = b_0^{\text{B}} + b_1^{\text{B}} \cdot \log K_{\text{A}}^{\text{H}_i}$$

where  $b^{\text{B}}$  are the regression coefficients characterizing each reference base, and  $\log K_{\text{A}}^{\text{H}}$  values are characteristics of hydrogen-bonding acids, and hence represent the solute hydrogen bond acidities. All the equations intersect at a "magic" point where  $\log K = -1.1$  ( $K$  measured on molar scale). The general  $\log K_{\text{A}}^{\text{H}}$  values were then transformed into  $\alpha_2^{\text{H}}$  values suitable for multivariate regression analysis by the following:

$$\alpha_2^{\text{H}} = \frac{\log K_{\text{A}}^{\text{H}} + 1.1}{4.636}$$

A fairly good correlation was found between  $\alpha_2^{\text{H}}$  scale and  $\alpha_{\text{m}}$  scale. Moreover, the set of original  $\alpha_2^{\text{H}}$  values was then enlarged by solving a system of solvatochromic equations on partition coefficients, thus including several new compounds and molecular fragments [Abraham, Chadha *et al.*, 1994b]. The **effective solute hydrogen-bond acidity**  $\sum \alpha_2^{\text{H}}$  was back-calculated by a number of multiple linear regression equations for solutes surrounded by a large excess of solvent and hence undergoing multiple hydrogen-bonding. This hydrogen bond descriptor agrees with  $\alpha_2^{\text{H}}$  values for monofunctional compounds, while for polyfunctional compounds it significantly differs [Abraham, 1993d].

The  $\rightarrow$  *hydrogen bond acceptor* (HBA) power of a compound is called **hydrogen bond basicity** (or **hydrogen-bond electron-acceptor power**) and is denoted by  $\beta_1$  and  $\beta_2$  for solvents and solutes, respectively.

The most important scales for **solvent HBA basicity** are here reported.

The  $\beta$  scale was proposed to measure solvent hydrogen bond basicity, that is, the ability of a bulk solvent to act as hydrogen bond acceptor. This scale was derived by systematic application of the solvatochromic comparison method; the final  $\beta$  values were calculated by averaging 13  $\beta$

parameters for each solvent obtained with different solutes and different physico-chemical properties [Kamlet, Abboud *et al.*, 1981, 1983].

The **Koppel–Paju B scale** was proposed to measure solvent hydrogen bond basicity, based on solvent shifts of the IR stretching frequencies of the free and hydrogen-bonded OH group of phenol in CCl<sub>4</sub> [Koppel and Paju, 1974].

The **Gutmann's Donor Number (DN)** was proposed [Gutmann, 1978] as quantitative empirical parameter for solvent nucleophilicity. For most solvents, it was found to well correlate with  $\beta$  scale.

The most important scale for **solute HBA basicity** are here reported.

The  **$\beta_m$  scale** was proposed for solute HBA basicity of “monomer” amphihydrogen-bonding compounds acting as non-self-associated solutes. In particular,  $\beta_m$  values are taken equal to  $\beta$  values for non-self-associating compounds.

The  **$\beta_2^H$  scale** was proposed for solute HBA basicity based on log  $K$  values for 1:1 complexation of series of bases against a number of reference acids in dilute solution of CCl<sub>4</sub> [Abraham, Grellier *et al.*, 1990]. Thirty-four linear equations were solved for each considered reference acid by a series of bases:

$$\log K_i = b_0^A + b_1^A \cdot \log K_B^H$$

where  $b^A$  are the regression coefficients characterizing each reference acid, and  $\log K_B^H$  values are characteristics of the bases representing the solute hydrogen bond basicities. All the equations intersect at a “magic” point where  $\log K = -1.1$  ( $K$  measured on molar scale). The general  $\log K_B^H$  values were then transformed into  $\beta_2^H$  values suitable for multivariate regression analysis by the following:

$$\beta_2^H = \frac{\log K_B^H + 1.1}{4.636}$$

This transformation was proposed to obtain a basicity scale with a zero-point corresponding to all non-hydrogen-bonding bases, such as alkanes and cycloalkanes. Moreover, on this scale, hexamethylphosphoric triamide basicity is equal to one. The set of original  $\beta_2^H$  values was then enlarged by solving a system of solvatochromic equations on partition coefficients, thus including several new compounds and molecular fragments [Abraham, Chadha *et al.*, 1994b]. The **effective solute hydrogen-bond basicity**  $\sum \beta_2^H$  was back-calculated by a number of multiple linear regression equations for solutes surrounded by a large excess of solvent and hence undergoing multiple hydrogen-bonding. This hydrogen bond descriptor agrees with  $\beta_2^H$  values for monofunctional compounds, while for poly-functional compounds it significantly differs [Abraham, Whiting *et al.*, 1991a; Abraham and Whiting, 1992; Abraham, 1993d].

For most solutes, the effective hydrogen-bond basicity is constant over all the solvent systems; however, in the case of some specific solutes, including anilines and pyridines, the effective solute hydrogen-bond basicity varies with the solvent system. Therefore, the descriptor  $\sum \beta_2^H$  is preferably used for partition between water and rather nonaqueous solvent systems, while an alternative  $\sum \beta_2^0$  can be used for partition between water and aqueous solvent systems [Abraham and Rafols, 1995].

**Table L3** LSER parameter values for some solutes. Symbols defined in the text and data taken from [Kamlet, Doherty *et al.*, 1987a; Abraham, Andonian-Haftvan *et al.*, 1994].

Solute	$\bar{V}/100$	$V_x$	$\log L^{16}$	$R_2$	$\sum \pi_2^H$	$\alpha_m$	$\beta_m$	$\sum \alpha_2^H$	$\sum \beta_2^H$
Diethyl ether	1.046	0.7309	2.015	0.041	0.25	0.00	0.47	0.00	0.45
Di- <i>n</i> -butyl ether	1.693	1.2945	3.924	0.000	0.25	0.00	0.46	0.00	0.45
1-Propanol	0.757	0.5900	2.031	0.236	0.42	0.33	0.45	0.37	0.48
2-Propanol	0.765	0.5900	1.764	0.212	0.36	0.33	0.51	0.33	0.56
1-Butanol	0.915	0.7309	2.601	0.224	0.42	0.33	0.45	0.37	0.48
2-Butanol	0.917	0.7309	2.338	0.217	0.36	0.33	0.51	0.33	0.56
1-Pentanol	1.086	0.8718	3.106	0.219	0.42	0.33	0.45	0.37	0.48
Cyclopentanol	1.009	0.7630	3.241	0.427	0.54	0.33	0.51	0.32	0.56
Trichloroethene	0.897	0.7146	2.997	0.524	0.40	0.00	0.10	0.08	0.03
1,1,1-Trichloroethane	0.989	0.7576	2.733	0.369	0.41	0.00	0.10	0.00	0.09
Tetrachloromethane	0.968	0.7391	2.823	0.458	0.38	0.00	0.10	0.00	0.00
1,2-Dichloroethane	0.787	0.6352	2.573	0.416	0.64	0.00	0.10	0.10	0.11
Butanone	0.895	0.6879	2.287	0.166	0.70	0.00	0.48	0.00	0.51
Cyclopentanone	0.986	0.7202	3.221	0.373	0.86	0.00	0.52	0.00	0.52
Cyclohexanone	1.136	0.8611	3.792	0.403	0.86	0.00	0.53	0.00	0.56
Acetonitrile	0.521	0.4042	1.739	0.237	0.90	0.15	0.35	0.04	0.33
Benzene	0.989	0.7164	2.786	0.610	0.52	0.00	0.10	0.00	0.14
Phenol	0.989	0.7751	3.766	0.805	0.89	0.61	0.33	0.60	0.31
<i>m</i> -Cresol	1.163	0.9160	4.329	0.822	0.88	0.55	0.33	0.57	0.34
Nitrobenzene	1.127	0.8906	4.557	0.871	1.11	0.00	0.30	0.00	0.28
2-Nitrotoluene	1.217	1.0315	4.878	0.866	1.11	0.00	0.30	0.00	0.28
Benzonitrile	1.120	0.8711	4.039	0.742	1.11	0.00	0.35	0.00	0.33
Tetrahydrofuran	0.911	0.6223	2.636	0.289	0.52	0.55	0.00	0.00	0.48

A **modified LSER** version, called **MLSER**, is based on  $\rightarrow$  *quantum-chemical descriptors* and defined as [Wang, Zhai *et al.*, 2005]

$$P_i = b_0 + b_1 \cdot \alpha + b_2 \cdot \mu + b_3 \cdot \epsilon_{\text{HOMO}} + b_4 \cdot q^- + b_5 \cdot \epsilon_{\text{LUMO}} + b_6 \cdot q^+$$

where  $\alpha$  is the  $\rightarrow$  *molecular polarizability*,  $\mu$  the  $\rightarrow$  *dipole moment*,  $\epsilon_{\text{HOMO}}$  and  $\epsilon_{\text{LUMO}}$  the energies of the highest occupied and lowest unoccupied orbitals, respectively, and  $q^-$  and  $q^+$  the negative and positive charges, respectively.

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

- **linear subfragment descriptors**  $\rightarrow$  substructure descriptors
- **line graph**  $\rightarrow$  graph
- **line graph connectivity indices**  $\rightarrow$  iterated line graph sequence
- **line graph Randić connectivity index**  $\rightarrow$  iterated line graph sequence
- **link**  $\equiv$  *connection*  $\rightarrow$  edge adjacency matrix
- **linking number**  $\rightarrow$  polymer descriptors
- **Lipinski drug-like index**  $\rightarrow$  property filters ( $\odot$  drug-like indices)
- **lipole**  $\rightarrow$  lipophilicity descriptors
- **lipophilicity**  $\rightarrow$  lipophilicity descriptors

### ■ lipophilicity descriptors

**Lipophilicity**, denoted as  $P$ , is the measure of the partitioning of a compound between a lipidic and an aqueous phase that depends on solute bulk, polar, and hydrogen-bonding effects [Taylor, 1990].

Compounds for which  $P > 1$  or  $\log P > 0$  are *lipophilic*, and compounds for which  $P < 1$  or  $\log P < 0$  are *hydrophilic*.

The most widely used molecular descriptor of lipophilicity is the  $\rightarrow$  *octanol–water partition coefficient*  $K_{ow}$  (**log  $K_{ow}$** , or also **log  $P$**  when no further specifications are given) that is the partition coefficient between 1-octanol and water:

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

Other  $\rightarrow$  *partition coefficients* related to lipophilicity are defined for  $n$ -alkane systems, such as the partition coefficient between  $n$ -heptane–water.

Lipophilicity can be factorized into two main terms as [Carrupt, Testa *et al.*, 1997]

$$\text{lipophilicity} = \text{hydrophobicity} - \text{polarity}$$

where **hydrophobicity** refers to nonpolar interactions (such as dispersion forces, hydrophobic interactions, etc.) of the solute with organic and aqueous phases and **polarity** to polar interactions (such as ion–dipole interactions, hydrogen-bonds, induction and orientation forces, etc.). As it can be observed, in this case, the term hydrophobicity is not synonymous with lipophilicity, but is a component of it.

Usually, hydrophobicity is encoded by  $\rightarrow$  *steric descriptors* such as molar or  $\rightarrow$  *molecular volume*, which account satisfactorily for nonpolar interactions; polarity can be described by polar terms that are negatively related to lipophilicity. An important factorization of lipophilicity is provided by the  $\rightarrow$  *solvatochromic parameters*. Moreover, a measure of the global polarity of a given solute was proposed by Testa and coworkers [Testa and Seiler, 1981; El Tayar and Testa, 1993; Vallat, Gaillard *et al.*, 1995] and called **interactive polar parameter  $\Lambda$**  (or **Testa lipophobic constant**). It is defined as the difference between the experimental lipophilicity measure and that estimated for an hypothetical  $n$ -alkane of the same molecular volume  $V$  as

$$\Lambda = (\log P)_{\text{exp}} - (\log P)_{\text{calc}}$$

where the calculated  $\log P$  is obtained from the following equation:

$$(\log P)_{\text{calc}} = 0.0309 \cdot V + 0.346$$

The interactive parameter  $\Lambda$  should by definition encode the same information as the solvatochromic polar parameters; it takes negative values for lipophobic fragments [van de Waterbeemd and Testa, 1987; El Tayar, Testa *et al.*, 1992, 1993].

For apolar compounds, an analogous linear relationship should be expected between  $\log P$  and  $\rightarrow$  *van der Waals volume*  $V^{vdw}$  that accounts for steric contributions to  $\log P$  [Moriguchi, 1975; Moriguchi, Kanada *et al.*, 1976, 1977]. An improved relation is obtained by incorporating into  $V^{vdw}$  a correction accounting for  $\rightarrow$  *molecular branching*. Moreover, to extend the relation to polar compounds, a correction factor called **Moriguchi polar parameter** (or **hydrophilic effect**)  $V_H$  was proposed as in the following equation:

$$\log P = 2.71 \times (V^{vdw} - V_H) + 0.12$$

Thus, the hydrophilic effect  $V_H$  is calculated as

$$V_H = V^{vdw} - \frac{\log P - 0.12}{2.71}$$

and it accounts for intramolecular hydrophobic bonding; moreover, it was found to well correlate with the interactive polar parameter  $\Lambda$ .

There are several methods developed for the calculation of  $\log P$  from molecular structure, based on  $\rightarrow$  *substituent constants*,  $\rightarrow$  *fragmental constants*,  $\rightarrow$  *electronic descriptors*,  $\rightarrow$  *steric descriptors*,  $\rightarrow$  *connectivity indices*,  $\rightarrow$  *surface areas*,  $\rightarrow$  *volume descriptors*,  $\rightarrow$  *chromatographic descriptors*.

Important reviews and books about lipophilicity are: [Leo, 1990; Hansch, Leo *et al.*, 1995; Carrupt, Testa *et al.*, 1997; Reinhard and Drefahl, 1999; Testa, Crivori *et al.*, 2000; Mannhold, 2003; Caron and Ermondi, 2008; Mannhold and Ostermann, 2008; Pliška, Testa *et al.*, 2008; van de Waterbeemd and Mannhold, 2008].

The most popular approaches to  $\log P$  calculation are listed below.

• **Hansch–Fujita hydrophobic substituent constants** ( $\equiv$  *hydrophobic substituent constants*,  $\pi$ )

The lipophilicity is calculated by analogy with the  $\rightarrow$  *Hammett equation* as

$$\log \frac{P_X}{P_H} = \rho \cdot \pi_X$$

where  $P_X$  and  $P_H$  are the partition coefficients of a X-substituted and unsubstituted compounds, respectively;  $\pi_X$  is the hydrophobic constant of the substituent X; the  $\rho$  constant reflects the characteristics of the solvent system and it is assumed equal to one for octanol/water solvent system [Fujita, Iwasa *et al.*, 1964].

These hydrophobic substituent constants are commonly used in  $\rightarrow$  *Hansch analysis* to encode the lipophilic behavior of the substituents; the lipophilicity of the whole molecule is obtained by adding to the lipophilicity of the unsubstituted parent compound ( $\log P_H$ ) the lipophilic contributions of the substituents:

$$\log P(\text{molecule}) = \log P_H + \sum_{s=1}^S \pi_{X_s}$$

where  $S$  is the number of substitution sites and  $\pi_{X_s}$  are the hydrophobic constants of the substituents in the molecule. Distinct values of the  $\pi$  constants were defined for aromatic and aliphatic compounds. The Hansch–Fujita hydrophobic constants are still widely used in QSAR studies, but not for calculating  $\log P$  values.

The major drawback of this approach is that  $\pi$  values depend on their electronic environment. When electronic interactions of the substituent X with other substituents in the compound are possible, more realistic  $\pi$  values have to be used. In particular,  $\pi^-$  lipophilic constant, also known as **Norrington lipophilic constant** [Norrington, Hyde *et al.*, 1975] measures the lipophilic contribution of strong electron-releasing groups such as  $-\text{OH}$ ,  $-\text{NH}_2$ ,  $-\text{NHR}$ , or  $-\text{NR}_1\text{R}_2$  when they are attached to a conjugated system (usually phenol or aniline);  $\pi^+$  lipophilic constant measures the lipophilic contribution of strong electron-attracting groups such as cyano or nitro groups, conjugated with the functional group. The use of this last constant is very rare.

Based on the decomposition of  $\log P$  into enthalpic  $P_h$  and entropic  $P_s$  contributions,

$$\log P = \frac{-\Delta G_p^0}{2.303 \cdot RT} = \frac{-\Delta H_p^0}{2.303 \cdot RT} + \frac{\Delta S_p^0}{2.303 \cdot R} = P_h + P_s$$

the hydrophobic substituent constant  $\pi$  was decomposed [Da, Ito *et al.*, 1992; Da, Yanagi *et al.*, 1993] into the **enthalpic hydrophobic substituent constant**  $\pi_h$  and **entropic hydrophobic substituent constant**  $\pi_s$ , respectively:

$$\pi = \pi_h + \pi_s$$

$$\pi_h = (P_h)_X - (P_h)_H \quad P_h = \frac{-\Delta H_p^0}{2.303 \cdot RT}$$

$$\pi_s = (P_s)_X - (P_s)_H \quad P_s = \frac{+\Delta S_p^0}{2.303 \cdot R}$$

where  $\Delta G_p^0$ ,  $\Delta H_p^0$ , and  $\Delta S_p^0$  are the Gibbs free energy, enthalpy, and entropy of transfer for partition, respectively;  $R$  is the gas constant and  $T$  the absolute temperature; the subscripts X and H denote the substituted and unsubstituted compound, respectively. The enthalpic contribution  $P_h$  can be interpreted as a new hydrophobic parameter that reflects the heat evolved when a solute is transferred from water to nonaqueous phase. Similarly, the entropic contribution  $P_s$  can be interpreted to reflect the change of randomness induced in the solution when a solute is transferred from water to nonaqueous phase.

📖 [Hansch, Maloney *et al.*, 1962; Hansch, Muir *et al.*, 1963; Hansch and Anderson, 1967; Martin and Lynn, 1971; Hansch, Leo *et al.*, 1971, 1972, 1973; Hansch and Dunn III, 1972; Fujita and Nishioka, 1976; Fujita, 1983; Leo, 1993; Gago, Pastor *et al.*, 1994; Amić, Davidović-Amić *et al.*, 1998]

#### • Nys-Rekker hydrophobic fragmental constants ( $f$ )

Also simply called **hydrophobic fragmental constants**, they are measures of the absolute lipophilicity contribution of specific molecular fragments to the lipophilicity of the molecule [Nys and Rekker, 1973, 1974; Rekker, 1977a, 1977b; Rekker and De Kort, 1979].

The  $\log P$  of a molecule is calculated by summing up the fragmental contributions and applying the appropriate correction factors as

$$\log P = b_0 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

where  $f_i$  and  $N_i$  are the hydrophobic constant and the number of occurrences of  $i$ th fragment in the considered compound,  $N_j$  is the number of occurrences of the  $j$ th correction factor.  $c_j$  is the value of the considered correction factor describing some special structural features (proximity effects, hydrogen atoms attached to polar groups, aryl-aryl conjugation, etc.); in practice, it can be calculated as

$$c_j = k_j \cdot 0.219$$

where 0.219 is the so-called “magic constant” and  $k_j$  is an integer value characterizing the  $j$ th correction factor.

Different sets of fragmental constants were derived by multiple regression analysis for fragments depending on their attachment to an aliphatic or aromatic carbon atom. In this approach, the effects of intramolecular interactions on lipophilicity are taken into account by the correction factors and implicitly in the definition of the molecular fragments. However, group interactions are evaluated more by the topological distance between the groups rather than by the chemical nature of the groups and their geometry.

A drawback of this approach is that, for the same compound, different selections of fragments give different  $\log P$  values. Moreover, for complex compounds, the decomposition of the molecular structure into appropriate fragments is not unique and is a difficult task.

Calculation of  $\log P$  based on revised Nys–Rekker fragmental constants is provided by some software, such as PROLOGP and SANALOGP.

📖 [Mayer, van de Waterbeemd *et al.*, 1982; Takeuchi, Kuroda *et al.*, 1990; Rekker, 1992; Rekker and Mannhold, 1992; Rekker and de Vries, 1993; Mannhold, Rekker *et al.*, 1998; Rekker, Mannhold *et al.*, 1998]

#### • Leo-Hansch hydrophobic fragmental constants ( $f'$ )

These are hydrophobic constants calculated for molecular fragments by a “constructionist approach” that consists of determining very accurately the  $\log P$  values of simple compounds usually having a single functional group and then calculating fundamental hydrophobic fragmental constants from these values [Hansch and Leo, 1979; Leo, 1987, 1993].  $\log P$  of compound is calculated by using the Rekker’s additive scheme, based on different fragmental constants  $f''$  and correction factors  $F'$ . The decomposition of the molecular structure into fragments is performed by using a unique and simple set of rules, thus obtaining a unique solution; the fragments are either atoms or polyatomic groups. Correction factors were derived from compounds with more than one substituent to better approximate experimental  $\log P$  values. They take into account proximity effects due to multiple halogenation and groups giving hydrogen-bonds, intramolecular hydrogen-bonds involving oxygen and nitrogen atoms, electronic effects in aromatic systems, unsaturation, branching, chains, rings. Over 200 fragmental constants and 14 correction factors have been determined.

The software version of the Leo-Hansch fragmental method is known as **CLOGP** or **Calculated LOGP** [Chou and Jurs, 1979].

📖 [Leo and Hansch, 1971; Leo, Jow *et al.*, 1975; Lyman, Reehl *et al.*, 1982; Mayer, van de Waterbeemd *et al.*, 1982; Abraham and Leo, 1987; Leo, 1991, 1993]

#### • Klopman hydrophobic models

The first model for the prediction of  $\log P$  proposed by Klopman and Iroff [Klopman and Iroff, 1981] was based on the assumption that partition coefficients of molecules depend on the charge densities on each atom of the molecule. The following equation was proposed including both atom and group counting descriptors and charge density descriptors:

$$\log P = b_0 + \sum_i b_i \cdot N_i + \sum_i b'_i \cdot q_i + \sum_i b''_i \cdot q_i^2 + \sum_j c_j \cdot N_j$$

where the first three summations run over the different types of atoms,  $b$  are the estimated regression coefficients,  $N_i$  the number of occurrences of the  $i$ th atom-type,  $q_i$  the charge density

on the  $i$ th atom-type;  $N_j$  are the occurrences or the presence/absence of some specific functional groups (acid/ester, nitrile, amide groups) whose influence on molecule lipophilicity is described by selected correction factors  $c_j$ .

Another model was proposed based on atomic composition of the molecule only, ignoring the influence of the charge density descriptors on the  $\log P$  calculation [Klopman, Namboodiri *et al.*, 1985]. The best-fitted proposed model is

$$\log P = -0.206 + 0.332 \cdot N_C + 0.071 \cdot N_H - 0.860 \cdot N_O - 1.124 \cdot N_N + 0.688 \cdot N_{Cl} \\ + 0.981 \cdot (N_{ac} + N_{est}) - 0.138 \cdot N_{Ph} + 2.969 \cdot N_{NO_2} + 1.053 \cdot I_{aliph}$$

$$n = 195; \quad r^2 = 0.949; \quad s = 0.293; \quad F = 33$$

where  $N_C$ ,  $N_H$ ,  $N_O$ ,  $N_N$ , and  $N_{Cl}$  are the numbers of carbon, hydrogen, oxygen, nitrogen, and chlorine atoms, respectively;  $N_{ac}$  and  $N_{est}$  are the numbers of acid and ester groups, respectively;  $N_{Ph}$ , and  $N_{NO_2}$  are the numbers of phenyl rings and nitrile groups;  $I_{aliph}$  is an indicator variable for aliphatic hydrocarbons.

The regression coefficients of the model relative to the atomic counting descriptors can be viewed as individual  $\log P$  contributions of each atom; these are the **Klopman hydrophobic atomic constants**. A better evaluation of these atomic contributions was proposed by classifying the atoms also accounting for their environment represented by the first neighbors [Klopman and Wang, 1991]. Moreover, this method uses for the evaluation of  $\log P$  only those atom-centered groups that are identified by stepwise regression as the most significant groups determining  $\log P$ .

A further developed model (**KLOGP** or **Klopman LOGP**), was proposed as

$$\log P = b_0 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

where  $N_i$  is the number of occurrences of the  $i$ th atom-centered fragment in the molecule and  $N_j$  are the number of occurrences of particular fragments accounting for the interactions between groups whose influence on molecule lipophilicity is described by calculated correction factors  $c_j$  [Klopman, Li *et al.*, 1994]. Basic atom-centered groups are of two types: (a) atomic groups defined by their chemical element, hybridization state, and the number of attached hydrogen atoms; (b) functional groups containing at least one heteroatom. Correction factors are supplementary hydrophobic constants relative to specific substructures with more than two nonhydrogen atoms.

The regression coefficients  $b_i$  are the Klopman hydrophobic atomic constants measuring the hydrophobic contributions of atom-types in the same way as the hydrophobic fragmental constants  $f_i$  defined in Rekker and Leo–Hansch approaches. The best evaluation of 64 atomic constants plus 30 correction factors was obtained by a training set of 1663 compounds,  $r^2 = 0.93$ ,  $s = 0.38$ ,  $F = 218$

The automated recognition of fragments and correction factors is performed by **CASE approach** (*Computer Automated Structure Evaluation*). Basically, CASE is an artificial intelligence system capable of identifying structural fragments that may be associated with the properties of the training molecules, such as biological activity and physico-chemical properties [Klopman, 1984]. **MULTICASE** (or **MCASE**) is the most recent upgraded software version [Klopman, 1992, 1998; Saiakhov, Stefan *et al.*, 2000; Klopman, Zhu *et al.*, 2003].



- **Suzuki–Kudo hydrophobic fragmental constants**

The contribution method of Suzuki and Kudo [Suzuki and Kudo, 1990; Suzuki, 1991] is based on hydrophobic fragmental constants  $f_i$  and it is defined as

$$\log P = b_0 + \sum_i f_i \cdot N_i$$

where  $N_i$  is the number of occurrences of the  $i$ th fragment in the molecule. A first set of 415 basic hydrophobic constants was proposed, representing the lipophilic contributions of groups, each described by its structural environment. Several basic groups were first defined as  $\text{CH}_3$ ,  $\text{CH}_2$ ,  $\text{CO}$ ,  $\text{SO}_2$ , and so on, which were further distinguished according to their neighboring atoms with their connectivities. Groups of atoms such as cyano and nitro are considered as univalent heteroatoms. In addition, extended fragments based on the basic fragments plus some other functional groups were selected together with some user-defined fragments. A training set of 1465 compounds plus a test set of 221 compounds were used to evaluate the hydrophobic constants by multivariate regression analysis. The software version of Suzuki–Kudo method is **CHEMICALC** (*Combined Handling of Estimation Methods Intended for Completely Automated Log P Calculation*).

- **Broto–Moreau–Vanduycke hydrophobic atomic constants**

The Broto–Moreau–Vanduycke contribution method is based on hydrophobic atomic constants  $a_k$  measuring the lipophilic contributions of atoms, each described by its nature, neighboring atoms and associated connectivities, thus implicitly considering some proximity effects and interactions in conjugated systems [Broto, Moreau *et al.*, 1984b]. Hydrogen atoms and correction factors are not explicitly considered. The model is defined as

$$\log P = b_0 + \sum_i a_i \cdot N_i$$

where  $N_i$  is the number of occurrences of the  $i$ th atom-type in the molecule. The atomic constants  $a_i$  were estimated by multivariate regression analysis.

Atom-types are classified according to their structural environment; carbon atom-types are differentiated by their bonds to nonhydrogen atoms; heteroatoms are differentiated by their bonds to nonhydrogen first neighbors and the nature of the neighbors, moreover if the neighbor atom is a carbon atom its bond environment is also accounted for. A conjugation contribution is considered as correction factor for  $sp^2$  carbon atoms in conjugated systems.

Using a training set of 1868 compounds, a set of 222 atomic constants was proposed and the best-fitted model gave a standard error about 0.4 log unit.

The software program **SMILOGP** for the calculation of log  $P$  is based on the Broto–Moreau–Vanduycke hydrophobic constants and the SMILES notation for the recognition of molecule atom-centered fragments [Convard, Dubost *et al.*, 1994].

- **Ghose–Crippen hydrophobic atomic constants**

These are hydrophobic atomic constants  $a_k$  measuring the lipophilic contribution of atoms in the molecule, each described by its neighboring atoms [Ghose and Crippen, 1986; Ghose, Pritchett *et al.*, 1988; Viswanadhan, Ghose *et al.*, 1989]. The model for log  $P$  calculation is defined as

$$\log P = \sum_k a_k \cdot N_k$$

where  $N_k$  is the number of occurrences of the  $k$ th atom-type.

A set of 120  $\rightarrow$  *atom-centered fragment descriptors* that are the **Ghose–Crippen descriptors**, was proposed (Table L4). Atom-centered fragments were defined for hydrogen atoms, carbon atoms, and heteroatoms. Hydrogen and halogen atoms are classified by the hybridization and oxidation state of the carbon atom to which they are bonded; for hydrogens, heteroatoms attached to a carbon atom in  $\alpha$ -position are further considered. Carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms.

The corresponding hydrophobic constants were evaluated by multivariate regression analysis using a training set of 8364 compounds,  $r^2 = 0.95$ ,  $Q^2 = 0.90$  and  $RMSE = 0.555$  [Ghose, Viswanadhan *et al.*, 1998]. The  $\log P$  estimated by Ghose–Crippen method is actually called **ALOGP** [Viswanadhan, Reddy *et al.*, 1993]. As in the Broto–Moreau–Vanduycke approach, correction factors are avoided, while hydrogen atoms are instead considered.

Calculation of ALOGP is provided by a number of software programs, such as DRAGON, MOLCAD, PROLOGP, and TSAR.

A smaller set of atom-centered fragments was later proposed to avoid ambiguity sometimes occurring in atom-type assignment. It is comprehensive for the common elements in organic molecules, and also includes metals and noble gases [Wildman and Crippen, 1999].

Ghose–Crippen descriptors were successfully used also to model  $\rightarrow$  *molar refractivity* [Ghose and Crippen, 1987] and solvation free energies [Viswanadhan, Ghose *et al.*, 1999] by  $\rightarrow$  *group contribution methods*.

**Table L4** Ghose–Crippen atomic contributions to  $\log P$  and molar refractivity (MR).

ID	Description	$\log P$	MR	ID	Description	$\log P$	MR
C-001	CH <sub>3</sub> R/CH <sub>4</sub>	-1.5603	2.968	O-061	O <sup>-a</sup>	1.052	1.945
C-002	CH <sub>2</sub> R <sub>2</sub>	-1.012	2.9116	O-062	O <sup>-</sup> (negatively charged)	-0.7941	—
C-003	CHR <sub>3</sub>	-0.6681	2.8028	O-063	R—O—O—R	0.4165	—
C-004	CR <sub>4</sub>	-0.3698	2.6205	Se-064	Any—Se—Any	0.6601	—
C-005	CH <sub>3</sub> X	-1.788	3.015	Se-065	=Se	—	—
C-006	CH <sub>2</sub> RX	-1.2486	2.9244	N-066	Al—NH <sub>2</sub>	-0.5427	2.6221
C-007	CH <sub>2</sub> X <sub>2</sub>	-1.0305	2.6329	N-067	Al <sub>2</sub> —NH	-0.3168	2.5
C-008	CHR <sub>2</sub> X	-0.6805	2.504	N-068	Al <sub>3</sub> —N	0.0132	2.898
C-009	CHRX <sub>2</sub>	-0.3858	2.377	N-069	Ar—NH <sub>2</sub> /X—NH <sub>2</sub>	-0.3883	3.6841
C-010	CHX <sub>3</sub>	0.7555	2.559	N-070	Ar—NH—Al	-0.0389	4.2808
C-011	CR <sub>3</sub> X	-0.2849	2.303	N-071	Ar—NAl <sub>2</sub>	0.1087	3.6189
C-012	CR <sub>2</sub> X <sub>2</sub>	0.02	2.3006	N-072	RCO—N</>N—X=X	-0.5113	2.5
C-013	CRX <sub>3</sub>	0.7894	2.9627	N-073	Ar <sub>2</sub> NH/Ar <sub>3</sub> N/ Ar <sub>2</sub> N—Al/R..N..R <sup>b</sup>	0.1259	2.7956
C-014	CX <sub>4</sub>	1.6422	2.3038	N-074	R#N/R=N—	0.1349	2.7
C-015	=CH <sub>2</sub>	-0.7866	3.2001	N-075	R—N—R <sup>c</sup> /R—N—X	-0.1624	4.2063
C-016	=CHR	-0.3962	4.2654	N-076	Ar—NO <sub>2</sub> /R—N(—R)— O <sup>d</sup> /RO—NO	-2.0585	4.0184
C-017	=CR <sub>2</sub>	0.0383	3.9392	N-077	Al—NO <sub>2</sub>	-1.915	3.0009
C-018	=CHX	-0.8051	3.6005	N-078	Ar—N=X/X—N=X	0.4208	4.7142
C-019	=CRX	-0.2129	4.487	N-079	N <sup>+</sup> (positively charged)	-1.4439	—

(Continued)

Table L4 (Continued)

ID	Description	log P	MR	ID	Description	log P	MR
C-020	=CX <sub>2</sub>	0.2432	3.2001	U-080	Undefined	—	—
C-021	#CH	0.4697	3.4825	F-081	F <sup>e</sup> attached to C <sup>1</sup> (sp <sup>3</sup> )	0.4797	0.8725
C-022	#CR/R=C=R	0.2952	4.2817	F-082	F <sup>e</sup> attached to C <sup>2</sup> (sp <sup>3</sup> )	0.2358	1.1837
C-023	#CX	—	3.9556	F-083	F <sup>e</sup> attached to C <sup>3</sup> (sp <sup>3</sup> )	0.1029	1.1573
C-024	R--CH--R	-0.3251	3.4491	F-084	F <sup>e</sup> attached to C <sup>1</sup> (sp <sup>2</sup> )	0.3566	0.8001
C-025	R--CR--R	0.1492	3.8821	F-085	F <sup>e</sup> attached to C <sup>2</sup> (sp <sup>2</sup> )-C <sup>4</sup> (sp <sup>2</sup> )/ C <sup>1</sup> (sp)/C <sup>4</sup> (sp <sup>3</sup> )/X	0.1988	1.5013
C-026	R--CX--R	0.1539	3.7593	Cl-086	Cl <sup>e</sup> attached to C <sup>1</sup> (sp <sup>3</sup> )	0.7443	5.6156
C-027	R--CH--X	0.0005	2.5009	Cl-087	Cl <sup>e</sup> attached to C <sup>2</sup> (sp <sup>3</sup> )	0.5337	6.1022
C-028	R--CR--X	0.2361	2.5	Cl-088	Cl <sup>e</sup> attached to C <sup>3</sup> (sp <sup>3</sup> )	0.2996	5.9921
C-029	R--CX--X	0.3514	3.0627	Cl-089	Cl <sup>e</sup> attached to C <sup>1</sup> (sp <sup>2</sup> )	0.8155	5.3885
C-030	X--CH--X	0.1814	2.5009	Cl-090	Cl <sup>e</sup> attached to C <sup>2</sup> (sp <sup>2</sup> )-C <sup>4</sup> (sp <sup>2</sup> )/ C <sup>1</sup> (sp)/C <sup>4</sup> (sp <sup>3</sup> )/X	0.4856	6.1363
C-031	X--CR--X	0.0901	—	Br-091	Br <sup>e</sup> attached to C <sup>1</sup> (sp <sup>3</sup> )	0.8888	8.5991
C-032	X--CX--X	0.5142	2.6632	Br-092	Br <sup>e</sup> attached to C <sup>2</sup> (sp <sup>3</sup> )	0.7452	8.9188
C-033	R--CH..X	-0.3723	3.4671	Br-093	Br <sup>e</sup> attached to C <sup>3</sup> (sp <sup>3</sup> )	0.5034	8.8006
C-034	R--CR..X	0.2813	3.6842	Br-094	Br <sup>e</sup> attached to C <sup>1</sup> (sp <sup>2</sup> )	0.8995	8.2065
C-035	R--CX..X	0.1191	2.9372	Br-095	Br <sup>e</sup> attached to C <sup>2</sup> (sp <sup>2</sup> )-C <sup>4</sup> (sp <sup>2</sup> )/ C <sup>1</sup> (sp)/C <sup>4</sup> (sp <sup>3</sup> )/X	0.5946	8.7352
C-036	Al-CH=X	-0.132	4.019	I-096	I <sup>e</sup> attached to C <sup>1</sup> (sp <sup>3</sup> )	1.4201	13.9462
C-037	Ar-CH=X	-0.0244	4.777	I-097	I <sup>e</sup> attached to C <sup>2</sup> (sp <sup>3</sup> )	1.1472	14.0792
C-038	Al-C(=X)-Al	-0.2405	3.9031	I-098	I <sup>e</sup> attached to C <sup>3</sup> (sp <sup>3</sup> )	—	14.073
C-039	Ar-C(=X)-R	-0.0909	3.9964	I-099	I <sup>e</sup> attached to C <sup>1</sup> (sp <sup>2</sup> )	0.7293	12.9918
C-040	R-C(=X)-X/ R-C#X/X=C=X	-0.1002	3.4986	I-100	I <sup>e</sup> attached to C <sup>2</sup> (sp <sup>2</sup> )-C <sup>4</sup> (sp <sup>2</sup> )/ C <sup>1</sup> (sp)/C <sup>4</sup> (sp <sup>3</sup> )/X	0.7173	13.3408
C-041	X-C(=X)-X	0.4182	3.4997	F-101	Fluoride ion	—	—
C-042	X--CH..X	-0.2147	2.7784	Cl-102	Chloride ion	-2.6737	—
C-043	X--CR..X	-0.0009	2.6267	Br-103	Bromide ion	-2.4178	—
C-044	X--CX..X	0.1388	2.5	I-104	Iodide ion	-3.1121	—
U-045	Undefined	—	—	U-105	Undefined	—	—
H-046	H <sup>e</sup> attached to C <sup>0</sup> (sp <sup>3</sup> ) no X attached to next C	0.7341	0.8447	S-106	R-SH	0.6146	7.8916
H-047	H <sup>e</sup> attached to C <sup>1</sup> (sp <sup>3</sup> )/C <sup>0</sup> (sp <sup>2</sup> )	0.6301	0.8939	S-107	R <sub>2</sub> S/RS-SR	0.5906	7.7935
H-048	H <sup>e</sup> attached to C <sup>2</sup> (sp <sup>3</sup> )/C <sup>1</sup> (sp <sup>2</sup> )/ C <sup>0</sup> (sp)	0.518	0.8005	S-108	R=S	0.8758	9.4338
H-049	H <sup>e</sup> attached to C <sup>3</sup> (sp <sup>3</sup> )/C <sup>2</sup> (sp <sup>2</sup> )/ C <sup>3</sup> (sp <sup>2</sup> )/C <sup>3</sup> (sp)	-0.0371	0.832	S-109	R-SO-R	-0.4979	7.7223
H-050	H attached to heteroatom	-0.1036	0.8	S-110	R-SO <sub>2</sub> -R	-0.3786	5.7558

(Continued)

Table L4 (Continued)

ID	Description	log P	MR	ID	Description	log P	MR
H-051	H attached to alpha-C <sup>f</sup>	0.5234	0.8188	Si-111	>Si<	1.5188	—
H-052	H <sup>e</sup> attached to C <sup>0</sup> (sp <sup>3</sup> ) with 1X attached to next C	0.6666	0.9215	B-112	>B— as in boranes	1.0255	—
H-053	H <sup>e</sup> attached to C <sup>0</sup> (sp <sup>3</sup> ) with 2X attached to next C	0.5372	0.9769	U-113	Undefined	—	—
H-054	H <sup>e</sup> attached to C <sup>0</sup> (sp <sup>3</sup> ) with 3X attached to next C	0.6338	0.7701	U-114	Undefined	—	—
H-055	H <sup>e</sup> attached to C <sup>0</sup> (sp <sup>3</sup> ) with 4X attached to next C	0.362	—	P-115	P ylides	—	—
O-056	Alcohol	−0.3567	1.7646	P-116	R <sub>3</sub> –P=X	−0.9359	5.5306
O-057	Phenol/enol/carboxyl OH	−0.0127	1.4778	P-117	X <sub>3</sub> –P=X (phosphate)	−0.1726	5.5152
O-058	=O	−0.0233	1.4429	P-118	PX <sub>3</sub> (phosphite)	−0.7966	6.836
O-059	Al–O–Al	−0.1541	1.6191	P-119	PR <sub>3</sub> (phosphine)	0.6705	10.0101
O-060	Al–O–Ar/ Ar–O–Ar/R..O..R/ R–O–C=X	0.0324	1.3502	P-120	C–P(X) <sub>2</sub> =X (phosphonate)	−0.4801	5.2806

R: any group linked through carbon; X: any electronegative atom (O, N, S, P, Se, halogens); Al and Ar: aliphatic and aromatic groups, respectively; =: a double bond; #: a triple bond; -: an aromatic bond as in benzene or delocalized bonds such as the N–O bond in a nitro group; ..: aromatic single bonds as the C–N bond in pyrrole. Data from [Ghose, Viswanadhan *et al.*, 1998].

<sup>a</sup>As in nitro, N-oxides.

<sup>b</sup>Pyrrole-type structure.

<sup>c</sup>Pyridine-type structure.

<sup>d</sup>Pyridine N-oxide type structure.

<sup>e</sup>The superscript represents the formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the conventional bond orders with electronegative atoms; the C–N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds; the C..X bond order in pyrrole or furan may be considered as 1.

<sup>f</sup>An alpha-C may be defined as a C attached through a single bond with –C=X, –C#X, –C–X.

#### • MLOGP ( $\equiv$ Moriguchi model based on structural parameters)

This is a model described by a regression equation based on 13 structural parameters and defined as [Moriguchi, Hirono *et al.*, 1992b, 1994]

$$\log P = -1.014 + 1.244 \cdot (F_{CX})^{0.6} - 1.017 \cdot (N_O + N_N)^{0.9} + 0.406 \cdot F_{PRX} - 0.145 \cdot N_{UNS}^{0.8} \\ + 0.511 \cdot I_{HB} + 0.268 \cdot N_{POL} - 2.215 \cdot F_{AMP} + 0.912 \cdot I_{ALK} - 0.392 \cdot I_{RNG} - 3.684 \cdot F_{QN} \\ + 0.474 \cdot N_{NO_2} + 1.582 \cdot F_{NCS} + 0.773 \cdot I_{BL}$$

$$n = 1230; \quad r^2 = 0.91; \quad s = 0.411; \quad F = 900.4$$

The meaning of the structural parameters and corresponding regression coefficients are reported in Table L5.

**Table L5** Regression coefficients of the Moriguchi model.

Symbol	Descriptor	$b_i$
$b_0$	Intercept	−1.014
$F_{CX}$	Summation of number of carbon and halogen atoms weighted by C = 1.0; F = 0.5; Cl = 1.0; Br = 1.5; I = 2.0	1.244
$N_O + N_N$	Total number of nitrogen and oxygen atoms	−1.017
$F_{PRX}$	Proximity effect of N/O: X – Y = 2; X – A – Y = 1 (X, Y: N/O; A: C, S, or P) with correction −1 for carbox-amide/sulfonamide	0.406
$N_{UNS}$	Total number of unsaturated bonds (not those in NO <sub>2</sub> )	−0.145
$I_{HB}$	Dummy variable for the presence of intramolecular H-bonds	0.511
$N_{POL}$	Number of polar substituents	0.268
$F_{AMP}$	Amphoteric property: $\alpha$ -amino = 1.0; aminobenzoic acid or pyridinecarboxylic acid = 0.5	−2.215
$I_{ALK}$	Dummy variable for alkane, alkene, cycloalkane, cycloalkene (hydrocarbons with 0 or 1 double bond)	0.912
$I_{RNG}$	Dummy variable for the presence of ring structures (not benzene and its condensed rings)	−0.392
$F_{QN}$	Quaternary nitrogen = 1.0; N-oxide = 0.5	−3.684
$N_{NO_2}$	Number of nitro groups	0.474
$F_{NCS}$	–N=C=S group = 1.0; –S–CN group = 0.5	1.582
$I_{\beta L}$	Dummy variable for the presence of $\beta$ -lactam	0.773

#### • Moriguchi model based on surface area

This is a model for predicting lipophilicity of compounds based on the  $\rightarrow$  *solvent-accessible surface area* SASA generated by a solvent probe of 1.4 Å radius and a set of parameters encoding hydrophilic effects of polar groups [Iwase, Komatsu *et al.*, 1985]:

$$\log P = -1.06 + 1.90 \cdot SASA - 1.00 \cdot \sum_k S_{H_k}$$

$$n = 138; \quad r^2 = 0.99; \quad s = 0.13; \quad F = 7284$$

where  $S_H$  are measures of the surface area of polar groups contributing negatively to  $\log P$  of the compounds. The latter parameters can be considered as fragmental correction factors whose values are derived separately for polar groups in aliphatic and aromatic systems.

#### • Dunn model based on surface area

This is a model for predicting  $\log P$  values in different solvent systems [Dunn III, Koehler *et al.*, 1987; Koehler, Grigoras *et al.*, 1988], defined by the equation

$$\log P_{solv} = a_{solv} \cdot ISA - b_{solv} \cdot f(HSA)$$

where  $ISA$  is the  $\rightarrow$  *isotropic surface area*, related to the solute surface accessible to nonspecific solvent interactions, and  $HSA$  the solvent-accessible  $\rightarrow$  *hydrated surface area*, associated with

hydration of polar functional groups.  $f(HSA)$  is the *hydrated fraction surface area*, that is,  $HSA/SASA$ , encoding the polar component of the lipophilicity as the  $S_H$  parameter in the  $\rightarrow$  Moriguchi model based on surface area.

#### • Camilleri model based on surface area

This is a model based on the factorization of the solvent-accessible surface area into 12 contributions relative to 12 molecular fragments [Camilleri, Watts *et al.*, 1988]:

$$\log P = b_0 + \sum_{i=1}^{12} b_i \cdot N_i$$

where  $b$  are the regression coefficients associated with the surface area contributions and  $N_i$  the number of occurrences of the  $i$ th molecular fragment (Table L6).

**Table L6** Regression coefficients of the Camilleri model.

Type	Fragment	$b_i$
$b_0$	Intercept	−23.9
$b_1$	Aromatic hydrocarbon	2.49
$b_2$	Saturated hydrocarbon chains not $A_3$ , $A_6$ , $A_{10}$ , $A_{12}$	2.731
$b_3$	Single saturated carbon atom attached to a nonhydrocarbon group plus hydrogens	−2.237
$b_4$	OH group	−1.809
$b_5$	Oxygen atom of OR group not $A_{11}$	−0.042
$b_6$	Hydrocarbon part of OR group not $A_{12}$	0.963
$b_7$	Cl atom	3.634
$b_8$	NH <sub>2</sub> or NH group	−3.197
$b_9$	C(=O)R group	−0.712
$b_{10}$	Hydrocarbon chain part in C(=O)R group	0.697
$b_{11}$	Oxygen atom of OR group in C(=O)OR group	−8.54
$b_{12}$	Hydrocarbon part of OR group in C(=O)OR group	3.526

#### • Politzer hydrophobic model

This is a model obtained by applying the  $\rightarrow$  GIPF approach (*General Interaction Properties Function approach*) proposed by Politzer and coworkers [Brinck, Murray *et al.*, 1993; Murray, Brinck *et al.*, 1993, 1994] as general method to estimate  $\rightarrow$  physico-chemical properties in terms of  $\rightarrow$  molecular electrostatic potential (MEP) properties calculated at the  $\rightarrow$  molecular surface.

The Politzer hydrophobic model was proposed as the following:

$$\log P = -0.504 + 0.0300 \cdot SA - 0.00472 \cdot (N_N + 2N_O) \cdot \sigma_-^2 - 0.000963 \cdot SA \cdot \Pi$$

$$n = 70; \quad r^2 = 0.97; \quad s = 0.277$$

where  $SA$  is the molecular surface area,  $N_N$  and  $N_O$  are the numbers of nitrogen and oxygen atoms, respectively,  $\sigma_-^2$  is the variance of the negative regions of the molecular surface potential,  $\Pi$  is the  $\rightarrow$  local polarity index.

Improvements of the Politzer hydrophobic model were later proposed using additional  $\rightarrow$  quantum-chemical descriptors derived from the molecular electrostatic potential, dipole moment,

and ionization energies. These descriptors were searched for to give the best estimates of the cavity term, polarity/dipolarizability term, and hydrogen-bond parameters defined in  $\rightarrow$  *Linear Solvation Energy Relationships* [Haeberlein and Brinck, 1997].

• **KOWWIN** ( $\equiv$  *Meylan–Howard hydrophobic model; AFC method*)

The Meylan–Howard hydrophobic model is derived from an atom/fragment contribution method providing 150 hydrophobic atomic and fragmental constants  $f_i$  measuring the lipophilic contributions of atoms and fragments in the molecule, together with 250 correction factors [Meylan and Howard, 1995, 1996, 2000; KOWWIN – Syracuse Research Corporation, 2008].

The model is defined as

$$\log P = 0.229 + \sum_i f_i \cdot N_i + \sum_j c_j \cdot N_j$$

$$n = 2351; \quad r^2 = 0.982; \quad s = 0.216$$

where  $N_i$  is the number of occurrences of the  $i$ th atom-type or fragment, and  $N_j$  is the number of occurrences of the  $j$ th correction factor  $c_j$ .

The hydrophobic constants  $f_i$  have been evaluated by a first linear regression analysis of 1120 compounds, without considering correction factors. The correction factors were then derived from a linear regression on additional 1231 compounds, correlating the differences between experimental  $\log P$  and the  $\log P$  estimated by the first regression model.

• **VLOGP** ( $\equiv$  *Gombar hydrophobic model*)

This is a model for the assessment of  $\log P$  based on 363 molecular descriptors derived from the molecular topology and obtained from 6675 diverse chemicals, with  $r^2 = 0.986$  and  $s = 0.20$  [Gombar and Enslein, 1996; Gombar, 1999]. Among the molecular descriptors considered are  $\rightarrow$  *molecular weight*,  $\rightarrow$  *electrotopological state indices*,  $\rightarrow$  *Kier shape descriptors* of order 1–7, and some  $\rightarrow$  *symmetry descriptors*. In particular, several descriptors are defined as the sum of the E-states of the atoms involved in the whole molecule or in predefined molecular fragments.

• **BLOGP** ( $\equiv$  *Bodor LOGP, Bodor hydrophobic model*)

This is a nonlinear 18-parameter model based on 10 molecular descriptors calculated by semiempirical quantum-chemistry methods, starting from optimized 3D geometries [Bodor, Gabanyi *et al.*, 1989; Bodor and Huang, 1992a, 1994]:

$$\begin{aligned} \log P = & 9.552 + 0.005286 \cdot \text{MW} + 0.08325 \cdot N_C + 1.0392 \cdot I_{\text{alk}} - 0.05726 \cdot \mu \\ & - 7.6661 \cdot O - 5.5961 \cdot O^2 + 2.1059 \cdot O^4 + 0.05984 \cdot SA - 0.0001141 \cdot SA^2 \\ & - 0.2741 \cdot Q - 8.5144 \cdot Q_N + 31.243 \cdot Q_N^2 - 17.377 \cdot Q_N^4 \\ & - 4.6249 \cdot Q_O + 20.346 \cdot Q_O^2 - 5.4195 \cdot Q_O^4 - 5.0040 \cdot Q_{\text{ON}} \end{aligned}$$

$$n = 302, \quad r^2 = 0.96, \quad s = 0.306, \quad F = 368$$

The BLOGP molecular descriptors are shown in Table L7.

Table L7 Molecular descriptors and regression coefficients of the BLOGP model.

Symbol	Descriptor	$b_i$
$b_0$	Intercept	9.5524
MW	Molecular weight	0.005 286
$N_C$	Number of carbon atoms	0.083 25
$I_{alk}$	Indicator variable for the presence of alkanes	1.0392
$\mu$	Dipole moment	−0.057 26
$O$	Ovality index	−7.6661
$O^2$	Second power of the ovality index	−5.5961
$O^4$	Fourth power of the ovality index	2.1059
SA	van der Waals surface area	0.059 84
$SA^2$	Second power of the van der Waals surface area	−0.000 1141
$Q$	Total absolute atomic charge	−0.2741
$Q_N$	Square root of the sum of the squared charges of nitrogen atoms	−8.5144
$Q_N^2$	Second power of $Q_N$	31.243
$Q_N^4$	Fourth power of $Q_N$	−17.377
$Q_O$	Square root of the sum of the squared charges of oxygen atoms	−4.6249
$Q_O^2$	Second power of $Q_O$	20.346
$Q_O^4$	Second power of $Q_O$	−5.4195
$Q_{ON}$	Sum of the absolute charges of oxygen and nitrogen atoms	−5.0040

The model shows high correlation between the independent variables and some nonsignificant regression coefficients.

A model based on the same set of  $\rightarrow$  *quantum-chemical descriptors* was also proposed for aqueous solubility by a neural network approach [Bodor, Harget *et al.*, 1991; Bodor, Huang *et al.*, 1992, 1994].

#### • Kantola–Villar–Loew hydrophobic models

This is a lipophilicity model based on atomic charges, surface areas, dipole moments, and a set of adjustable parameters depending only on the atomic number [Kantola, Villar *et al.*, 1991]. The parameter values are determined to reproduce experimental  $\log P$  values using the following general model:

$$\log P = \sum_{i=1}^A [\alpha_i \cdot SA_i + \beta_i \cdot SA_i \cdot q_i^2 + \gamma_i \cdot q_i] + \delta \cdot \mu$$

where SA are atomic contributions to the  $\rightarrow$  *solvent-accessible surface area*,  $q$  are atomic charges,  $\mu$  the dipole moment. Setting to zero some of the parameters  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\delta$ , different submodels are obtained. The model descriptors are calculated by  $\rightarrow$  *computational chemistry* methods, thus resulting in conformationally dependent hydrophobicity values.

#### • molecular lipophilicity potential model

This is a  $\log P$  model based on the  $\rightarrow$  *molecular lipophilicity potential* (MLP) defined as

$$\log P = -0.10 + 0.00286 \cdot \sum MLP^+ + 0.00152 \cdot \sum MLP^-$$



$$n = 114, \quad r^2 = 0.94, \quad s = 0.37, \quad F = 926$$

where descriptors  $\Sigma MLP^+$  and  $\Sigma MLP^-$  are the sum of the positive and negative  $MLP$  values, respectively [Gaillard, Carrupt *et al.*, 1994b]. They represent the hydrophobic and polar contributions of the molecule.

The specific expression for  $MLP$  used in this model is the following:

$$MLP_i = \sum_k a_k \cdot \frac{1 + \exp(b \cdot c)}{1 + \exp[b \cdot (r_{ik} - c)]}$$

where  $MLP_i$  is the molecular lipophilicity potential at the  $i$ th grid point,  $a_k$  are the  $\rightarrow$  Broto–Moreau–Vanduycke hydrophobic atomic constants,  $r_{ik}$  the  $\rightarrow$  geometric distance between the  $k$ th fragment and the  $i$ th grid point;  $b$  and  $c$  are the two parameters defining the shape of the Fermi-type function used to calculate  $MLP$  values ( $b = 1.33$  and  $c = 3.25$ ).  $MLP$  values are calculated using the  $\rightarrow$  solvent-accessible surface area as integration space.

#### • ACD/log P

This is a model for  $\log P$  calculation proposed by Petrauskas and Kolovanov [Petrauskas and Kolovanov, 2000] as a modification of the  $\rightarrow$  CLOGP, aimed at reducing the large number of correction factors involved in the CLOGP calculation.

For example, H-atoms are never detached from carbon atoms, as CLOGP is. This automatically enlarges a list of fragmental increments, but eliminates the need for many structural correction factors: chain and ring flexibility, chain and group branching, double and triple unsaturation, aromatic ring conjugation in biphenyls and fusion in naphthalenes, and so on.

Using a training set of 3601 compounds, the best-fitted model gave  $r^2 = 0.984$  and  $s = 0.21$ .

#### • HLOGP model

The HLOGP model, proposed by Viswanadhan *et al.*, [Viswanadhan, Ghose *et al.*, 2000], uses both smaller atom sized and larger fragments encoded into  $\rightarrow$  molecular holograms. The model was obtained by generating holograms of various lengths for each molecule in the training set and performing Partial Least Squares (PLS) analysis, followed by the selection of model features leading the least standard error.

Using a training set of 265 compounds, the best-fitted model, with a hologram length of 257 and using 18 PLS components, gave  $r^2 = 0.941$  and  $S = 0.58$ .

#### • XLOGP

This is a model for  $\log P$  calculation based on a  $\rightarrow$  group-contribution method proposed by Wang Renxiao *et al.* [Wang, Fu *et al.*, 1997; Wang, Gao *et al.*, 2000]. The model is defined in terms of the  $\rightarrow$  solvent accessible surface area and the atomic charges of 76 atom-types, together with five additional correction factors, as

$$\log P = \sum_i a_i \cdot N_i + \sum_j c_j \cdot N_j$$

where  $a$  and  $c$  are the lipophilic contributions of each atom-type and correction factor, respectively, and  $N_i$  and  $N_j$  are the number of occurrences of the  $i$ th atom-type and  $j$ th correction factor, respectively.

Using the SYBYL atomic codes, atom-types are classified as carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorous, and halogens in neutral organic molecules, according to their

hybridization states and their nearest neighboring atoms. Five correction factors were introduced to account for some intramolecular interactions.

Using a training set of 1831 compounds, the best-fitted model gave  $r^2 = 0.937$ , improving the results obtained without correction factors ( $r^2 = 0.908$ ).

#### • SLOGP

This is the  $\log P$  calculated by a  $\rightarrow$  *group-contribution method* proposed by Hou and Xu [Hou and Xu, 2002; Hou and Xu, 2003a; Hou and Xu, 2003b] based on the calculation of  $\rightarrow$  *solvent accessible surface area* for 100 atom/group types, together with two additional correction factors, and defined as

$$\log P = \sum_i a_i \cdot N_i + \sum_j c_j \cdot N_j$$

where  $a$  and  $c$  are the contributions of each atom-type and correction factor, respectively, and  $N_i$  and  $N_j$  are the number of occurrences of the  $i$ th atom-type and  $j$ th correction factor, respectively. Using the SMARTS atomic codes, atom-types were defined in the same way as it was for the XLOGP model.

Using a training set of 1850 compounds, containing the same 1831 compounds used to develop the XLOGP model, the best-fitted model gave  $r^2 = 0.976$ , with  $s = 0.368$ .

#### • Duchowitz–Castro $\log P$

This is the  $\log P$  calculated by a simple approach based on the contributions of molecule atoms and bonds, as [Duchowicz and Castro, 2000]

$$\log P = c_0 + \sum_i a_i \cdot A_i + \sum_j b_j \cdot B_j$$

where the first summation runs over the atom-types and second one over the bond-types;  $a$  and  $b$  are the regression coefficients of the different types of atoms (A) (C, H, O, N, etc.) and bonds (B) (C–H, C=O, O–H, etc.). This approach has been tested only on small congeneric data sets.

#### • lipole

The lipole of a molecule is a measure of the lipophilic distribution and is calculated from atomic lipophilicity values  $l_i$  as

$$L = \sum_{i=1}^A r_i \cdot l_i$$

where  $r_i$  and  $l_i$  are the distance from the  $\rightarrow$  *center of mass* and the lipophilicity of the  $i$ th atom, respectively;  $A$  is the number of atoms in the molecule [TSAR – Oxford Molecular Ltd., 1999].

#### • topological index of hydrophobicity

This is a topological index based on molecular connectivity defined as

$$\chi_H = {}^1\chi + \sum_k \delta\chi_k$$

where  ${}^1\chi$  is the  $\rightarrow$  *Randić connectivity index* and  $\delta\chi$  are correction factors determined experimentally [Sakhartova and Shatz, 1984; Shatz, Sakhartova *et al.*, 1984]. The corrections are  $\delta\chi(\text{alkanes}) = 0$ ,  $\delta\chi(\text{alkylbenzenes}) = -1.597$ ,  $\delta\chi(\text{ketones}) = -3.076$ .

### • Ferreira–Kiralj hydrophobicity parameters

Two simple structure-based lipophilicity parameters were proposed by Ferreira and Kiralj [Ferreira and Kiralj, 2004] for modeling  $\log P$ . The first parameter is the fraction  $w_C$  of the number of hydrophobic carbon atoms  $N_C^{\text{hyd}}$ , defined as the number of hydrophobic carbon atoms (all carbon atoms except those in C=O, C–O<sup>−</sup> and CN groups) divided by the number of all nonhydrogen atoms:

$$w_C = \frac{N_C^{\text{hyd}}}{A - N_H}$$

where  $A$  is the total number of atoms and  $N_H$  the number of hydrogens, respectively.

The second parameter is the surface fraction of hydrophobic carbon atoms  $S_F$ , calculated analogously to  $w_C$ : instead of atom counts, their CPK atomic surface areas from optimized geometries of compounds (in charged forms at neutral pH) are used.

These parameters were used in addition to other classical  $\log P$  descriptors in Partial Least Squares (PLS) regression for modeling biological activities.

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

- **Li valence vertex degree** → vertex degree
- **L/L quotient matrix** → biodescriptors (⊙ DNA sequences)
- **LMO technique**  $\equiv$  *leave-more-out technique* → validation techniques (⊙ cross-validation)
- **loading matrix** → Principal Component Analysis
- **local Balaban index** → connectivity indices
- **Local Chemical Environments** → scoring functions (⊙ MultiLevel Chemical Compatibility)
- **local connectivity indices** → connectivity indices
- **Local Density Of States** → quantum-chemical descriptors (⊙ EIM descriptors)
- **local dipole index** → charge descriptors
- **Local Edge Invariants** → local invariants
- **local ETA index** → ETA indices
- **local functionality index** → ETA indices
- **local hardness** → quantum-chemical descriptors (⊙ hardness indices)
- **local information on distances**  $\equiv$  *relative vertex distance complexity* → topological information indices

### ■ local invariants

These are numerical quantities derived from the molecular topology and used to characterize local properties of a molecule; these numbers are calculated in such a way as to be independent of any arbitrary atom/bond numbering. Local invariants can be distinguished into **Local Vertex Invariants** (LOVIs) and **Local Edge Invariants** (LOEIs), depending on whether they refer to atoms or bonds. They are usually calculated from the → *H-depleted molecular graphs*.

### • Local Vertex Invariants (LOVIs)

These are numerical quantities associated with graph vertices independently of any arbitrary vertex numbering, used to characterize local properties in a molecule. They can be either purely

topological if heteroatoms are not distinguished from carbon atoms, or chemical if the heteroatoms are assigned distinct values from carbon atoms, even when these are topologically equivalent [Balaban, 1987, 1994a; Filip, Balaban *et al.*, 1987; Ivanciuc, Balaban *et al.*, 1993b]. To account for the presence of heteroatoms, local vertex invariants can be calculated from molecular graphs where vertices are weighted by physico-chemical  $\rightarrow$  *atomic properties*. An ideal set of LOVIs is such that distinct LOVIs are relative to nonequivalent vertices in any graph.

LOVIs of a molecule are usually collected into an  $A$ -dimensional vector,  $A$  being the number of graph vertices, and, sometimes, as diagonal terms into a  $(A \times A)$  diagonal matrix. Examples of matrices collecting LOVIs are the  $\rightarrow$  *vertex degree matrix*,  $\rightarrow$  *vertex Zagreb matrix*, and  $\rightarrow$  *modified vertex Zagreb matrix* and several  $\rightarrow$  *augmented matrices*.

Local vertex invariants are used to calculate several molecular  $\rightarrow$  *topological indices* by applying different operators such as addition of LOVIs, addition of squares of LOVIs, addition of reciprocal geometric means for any pair of adjacent vertices. Moreover, they can be used to obtain  $\rightarrow$  *canonical numbering* of molecular graphs and compare molecules to study  $\rightarrow$  *molecular branching* and *centricity*.

The most well-known LOVIs are  $\rightarrow$  *vertex degree*,  $\rightarrow$  *valence vertex degree*,  $\rightarrow$  *bond vertex degree*, and the other several variants of the vertex degree,  $\rightarrow$  *atom eccentricity*,  $\rightarrow$  *vertex distance degree*,  $\rightarrow$  *walk degree*,  $\rightarrow$  *atomic path counts*,  $\rightarrow$  *atomic ID numbers*,  $\rightarrow$  *path degree*,  $\rightarrow$  *extended connectivity*,  $\rightarrow$  *exponential sum connectivities*,  $\rightarrow$  *graph potentials*, LOVIs calculated by  $\rightarrow$  *MPR approach* and those applying  $\rightarrow$  *centric operator* and  $\rightarrow$  *centrocomplexity operator* to  $\rightarrow$  *layer matrices*.

A general approach to derive a local vertex invariant from a symmetric  $\rightarrow$  *graph-theoretical matrix* ( $A \times A$ ) is to compute the sum of the elements in the  $i$ th row, or  $j$ th column, of the matrix  $\mathbf{M}$ :

$$\mathcal{L}_i \equiv VS_i(\mathbf{M}, w) = \sum_{j=1}^A [(\mathbf{M}, w)]_{ij}$$

where  $VS$  indicates the  $\rightarrow$  *row sum operator* and the resulting LOVI is called **vertex sum**;  $w$  is the  $\rightarrow$  *weighting scheme* used to calculate the molecular matrix  $\mathbf{M}$ .  $\mathcal{L}$  is here adopted as the general symbol for local vertex invariants. For unsymmetrical graph-theoretical matrices  $\mathbf{UM}(w)$ , the **vertex double sum**, denoted as  $VDS_i$ , was defined as local vertex invariant instead of the vertex sum [Ivanciuc, 1999c]:

$$\mathcal{L}_i \equiv VDS_i(\mathbf{M}, w) = \sum_{j=1}^A [\mathbf{M}(w)]_{ij} + \sum_{j=1}^A [\mathbf{M}(w)]_{ji} - [\mathbf{M}(w)]_{ii}$$

where the diagonal element is subtracted because it is added in both summations.

Another common LOVI derived from a graph-theoretical matrix is the maximum value of the matrix entries in the  $i$ th row:

$$\mathcal{L}_i = \max_j ([\mathbf{M}(w)]_{ij})$$

Other typical LOVIs are obtained by adding only matrix entries corresponding to the vertices adjacent to the  $i$ th vertex:

$$\mathcal{L}_i = \sum_{j=1}^A a_{ij} \cdot [\mathbf{M}(w)]_{ij}$$

where  $a_{ij}$  are the elements of the  $\rightarrow$  *adjacency matrix* equal to one for pairs of adjacent vertices, and zero otherwise. Moreover, an extension of this kind of LOVIs is represented by higher order LOVIs calculated as

$${}^k\mathcal{L}_i = \sum_{j=1}^A [\mathbf{M}(w)]_{ij} \cdot \delta(d_{ij}; k)$$

where  $\delta(d_{ij}; k)$  is the Kronecker delta function that is equal to one for pairs of vertices at a topological distance of  $k$ , and zero otherwise.

Extended LOVIs, defined by using the same formula as the  $\rightarrow$  *extended connectivity*, are generated according to the following expression:

$$\mathcal{L}_i^k = \sum_{j=1}^A a_{ij} \cdot \mathcal{L}_j^{k-1} \quad k = 1, 2, \dots$$

where  $a_{ij}$  are the elements of the  $\rightarrow$  *adjacency matrix*, being equal to one for pairs of adjacent vertices, and zero otherwise; at the beginning ( $k = 1$ ), any kind of LOVI is simply the  $\rightarrow$  *vertex degree*  $\delta$ , that is, the number of adjacent vertices.

Other LOVIs can be generated by different combinations of the basic LOVIs or any other atomic property, used as the  $\rightarrow$  *weighting scheme*  $w$  for graph vertices. Let  $w_1$ ,  $w_2$ , and  $w_3$  be three vertex weighting schemes, then a generalized LOVI function is here proposed as [Authors, This book]

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[ \sum_{j=1}^A a_{ij}^{(k)} \cdot \left( \frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

where  $w_1$ ,  $w_2$ , and  $w_3$  are the vertex weightings,  $a^{(k)}$  are the elements of the  $k$ th power of the  $\rightarrow$  *adjacency matrix* corresponding to the number of  $\rightarrow$  *equipose random walks* of length  $k$  (*walk count*) from vertex  $v_i$  to vertex  $v_j$ , and  $f$  is a distance smoothing function used to modulate the role of distances in attenuating contributions from vertices far apart.  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\phi$ , and  $\lambda$  are user-defined real parameters. The distance smoothing functions are those proposed in the definition of the  $\rightarrow$  *interaction graph matrices* and  $\rightarrow$  *perturbation graph matrices*:

$$f_1(d_{ij}, \gamma) = d_{ij}^\gamma \quad f_2(d_{ij}, \gamma) = (d_{ij} + 1)^\gamma \quad f_3(d_{ij}, \gamma) = 2^{\gamma \cdot d_{ij}} \quad f_4(d_{ij}, \gamma, x) = (d_{ij} \cdot x^{(d_{ij}-1)})^\gamma$$

Most of the well-known LOVIs are encompassed by this general definition.

According to this scheme, if  $\lambda = 0$ , a general LOVI for the  $i$ th vertex is defined taking into account only properties of the vertex itself:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta$$

If  $\gamma = 0$ , any information related to the topological distance between vertices is neglected and the LOVI reduces to

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[ \sum_{j=1}^A a_{ij}^{(k)} \cdot w_{3j}^\phi \right]^\lambda$$

Moreover, there are two basic ways to consider the vertex surrounding, depending on the power  $k$  of the adjacency matrix. If  $k = 1$ , only properties of the vertices bonded to the  $i$ th vertex are considered:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[ \sum_{j=1}^A a_{ij} \cdot \left( \frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

In this case,  $a_{ij}$  being the elements of the adjacency matrix, the summation runs over all the pairs of vertices, but the only nonvanishing contributions are from adjacent vertices; moreover, the influence of the smoothing functions is constant, the distance between adjacent vertices being always equal to one. If  $k = 0$ , properties of all the vertices in the graph take part in LOVI definition, each contributing by a quantity tuned by its separation from the considered vertex:

$$\mathcal{L}_i = w_{1i}^\alpha \cdot w_{2i}^\beta \cdot \left[ \sum_{j=1}^A \left( \frac{w_{3j}^\phi}{f(d_{ij}, \gamma)} \right) \right]^\lambda$$

The **VTI indices**, proposed by Ivanciuc [Ivanciuc, 1989], which are combinations of  $\rightarrow$  topological distances  $d$  and  $\rightarrow$  vertex degrees  $\delta$ , are included in this scheme. They are defined as

$$\text{VTI}_i = \delta_i^\beta \cdot \sum_{j=1}^A d_{ij}^\gamma \cdot \delta_j^\phi$$

Eighteen VTI indices were defined by setting  $\beta = 0, \pm 1$ ;  $\gamma = \pm 1$ ;  $\phi = 0, \pm 1$  (Table L8). Among these, combination  $\beta = 0, \gamma = 1, \phi = 0$  (ID 1) gives the  $\rightarrow$  distance sum,  $\beta = 1, \gamma = 1, \phi = 0$  (ID 2) gives the  $\rightarrow$  vertex degree distance,  $\beta = -1, \gamma = 1, \phi = 0$  (ID 3) gives the LOVI  $t_i$  used to calculate the  $\rightarrow J_i$  index, and  $\beta = 0, \gamma = -1, \phi = 0$  (ID 10) gives the  $\rightarrow$  reciprocal distance sum.

Other LOVIs similar to VTI indices are derived as the row sums of  $\rightarrow$  distance-degree matrices for different combinations of  $\beta$ ,  $\gamma$ , and  $\phi$  parameters. These local indices were extensively

**Table L8** List of the 18 VTI indices.

ID	$\gamma$	$\beta$	$\phi$	VTI	ID	$\gamma$	$\beta$	$\phi$	VTI
1	1	0	0	$\sum_j d_{ij} = \sigma_i$	10	-1	0	0	$\sum_j d_{ij}^{-1} = RDS_i$
2	1	1	0	$\delta_i \cdot \sum_j d_{ij} = \delta_i \cdot \sigma_i$	11	-1	1	0	$\delta_i \cdot \sum_j d_{ij}^{-1} = \delta_i \cdot RDS_i$
3	1	-1	0	$\delta_i^{-1} \cdot \sum_j d_{ij} = \sigma_i / \delta_i$	12	-1	-1	0	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} = RDS_i / \delta_i$
4	1	0	1	$\sum_j d_{ij} \cdot \delta_j$	13	-1	0	1	$\sum_j d_{ij}^{-1} \cdot \delta_j$
5	1	0	-1	$\sum_j d_{ij} \cdot \delta_j^{-1}$	14	-1	0	-1	$\sum_j d_{ij}^{-1} \cdot \delta_j^{-1}$
6	1	1	1	$\delta_i \cdot \sum_j d_{ij} \cdot \delta_j$	15	-1	1	1	$\delta_i \cdot \sum_j d_{ij}^{-1} \cdot \delta_j$
7	1	1	-1	$\delta_i \cdot \sum_j d_{ij} \cdot \delta_j^{-1}$	16	-1	1	-1	
8	1	-1	1	$\delta_i^{-1} \cdot \sum_j d_{ij} \cdot \delta_j$	17	-1	-1	1	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} \cdot \delta_j$
9	1	-1	-1	$\delta_i^{-1} \cdot \sum_j d_{ij} \cdot \delta_j^{-1}$	18	-1	-1	-1	$\delta_i^{-1} \cdot \sum_j d_{ij}^{-1} \cdot \delta_j^{-1}$

studied for analyzing  $\rightarrow$  *molecular branching* and used to derive molecular descriptors obtained by setting  $\beta$ ,  $\gamma$ , and  $\phi$  equal to a number of values such as  $-\infty$ ,  $-6$ ,  $-5$ ,  $-4$ ,  $-3$ ,  $-2$ ,  $-1$ ,  $-1/2$ ,  $-1/3$ ,  $-1/4$ ,  $-1/5$ ,  $0$ ,  $1/5$ ,  $1/4$ ,  $1/3$ ,  $1/2$ ,  $1$ ,  $2$ ,  $3$ ,  $4$  [Perdih, 2003; Perdih and Perdih, 2003d, 2004].

Another set of local vertex invariants was proposed by Diudea *et al.* [Diudea, Kacso *et al.*, 1996] using a Randić-like formula as

$$\text{conn}(w)_i = \sum_{j=1}^A a_{ij} \cdot (w_i \cdot w_j)^\alpha$$

where  $w_i$  and  $w_j$  are atomic weightings associated with vertices  $v_i$  and  $v_j$ ; the summation runs over all the vertices and accounts only for contributions from vertices adjacent to  $v_i$ ,  $a_{ij}$  being the elements of the adjacency matrix;  $\alpha$  is a real exponent usually equal to  $-1/2$  and sometimes to  $+1/2$ . By summing all LOVIs over all atoms, the corresponding molecular  $\rightarrow$  *graph invariant* is obtained. If the weighting scheme is the  $\rightarrow$  *vertex degree*, the obtained LOVIs correspond to twice the first order  $\rightarrow$  *local connectivity indices*. Moreover, the  $\rightarrow$  *Randić–Razinger index* and  $\rightarrow$  *local Balaban index* are obtained when the weighting scheme for vertices is the  $\rightarrow$  *walk degree* and the  $\rightarrow$  *vertex distance degree*, respectively [Diudea, Minailiuc *et al.*, 1997a].

Another set of local vertex invariants, denoted as  $\text{EFTI}_i$ , was derived from  $\rightarrow$  *fragment topological indices*, when one nonhydrogen atom at a time is considered:

$$\text{EFTI}_i = \text{TI}(\mathcal{G}) - \text{IFTI}(\mathcal{G}') - \text{IFTI}(i)$$

where  $\text{TI}$  is any topological index, increasing with increase in the number of graph vertices,  $\mathcal{G}'$  is the subgraph obtained by erasing the  $i$ th vertex with its incident edges,  $\text{IFTI}(i)$  is the corresponding topological index calculated for the  $i$ th vertex that is often equal to zero or constant (e.g.,  $\text{IFTI}(i) = 1$  for the  $\rightarrow$  *Hosoya Z index*) [Mekenyan, Bonchev *et al.*, 1988a].

### • Local Edge Invariants (LOEIs)

Analogous to the local vertex invariants, local edge invariants are descriptors of the graph edges used to characterize local properties in a molecule; they are numbers associated with graph edges independently of any arbitrary edge numbering. They can be calculated as the local vertex invariants of the first-order  $\rightarrow$  *line graph* corresponding to the molecular graph.

Edge invariants can also be directly obtained by  $\rightarrow$  *physico-chemical properties* of the bonds used as the  $\rightarrow$  *weighting scheme* for graph edges, such as bond dipole moments,  $\rightarrow$  *bond order indices*, and so on, as well as from the  $\rightarrow$  *edge adjacency matrix* and  $\rightarrow$  *edge distance matrix* by applying specific matrix operators such as the  $\rightarrow$  *row sum operator*.

Moreover, local edge invariants can be calculated by some combination of the local vertex invariants or atomic properties of the two incident vertices. Two general formulas to derive edge invariants from vertex invariants  $\mathcal{L}$  are

$$\mathcal{L}_{ij} = \frac{\mathcal{L}_i + \mathcal{L}_j}{2} \quad \text{and} \quad \mathcal{L}_{ij} = \sqrt{\mathcal{L}_i \cdot \mathcal{L}_j}$$

corresponding to the arithmetic and geometric mean, respectively, of the local vertex invariants of the two vertices  $v_i$  and  $v_j$  connected by the edge.

The two previous expressions can be extended taking into account all the vertices bonded to the two vertices  $v_i$  and  $v_j$  forming the edge  $i-j$ :

$$\mathcal{L}_{ij} = \sum_{k=1}^A a_{ik} \cdot \mathcal{L}_k + \sum_{k=1}^A a_{jk} \cdot \mathcal{L}_k \quad \text{and} \quad \mathcal{L}_{ij} = \prod_{k=1}^A (\mathcal{L}_k)^{a_{ik}} \cdot \prod_{k=1}^A (\mathcal{L}_k)^{a_{jk}}$$

Their average values can be also considered, using the corresponding arithmetic and geometric means.

A generalization of the concept of  $\rightarrow$  *edge connectivity*, defined in terms of  $\rightarrow$  *vertex degrees*, is given by the following expression:

$$\mathcal{L}_{ij} = (VS_i[\mathbf{M}] \cdot VS_j[\mathbf{M}])^\lambda$$

where  $\mathbf{M}$  is a  $\rightarrow$  *graph-theoretical matrix*,  $VS$  the  $\rightarrow$  *vertex sum operator*, and  $\lambda$  is a variable parameter.

To obtain a final topological index that gives greater weight to terminal bonds, if the vertex sum  $VS$  corresponding to terminal vertices of a graph are smaller than the average vertex sums of the interior vertices, a value of  $\lambda = -1/2$  is suggested. On the other hand, when the vertex sums corresponding to terminal vertices of the graph are greater than the average vertex sums of the interior vertices, a value of  $\lambda = 1/2$  should be chosen to generate bond contributions [Randić, Balaban *et al.*, 2001]. For  $\lambda = -1/2$ , the bond contribution above defined is the same as that used in the  $\rightarrow$  *Ivanciuc–Balaban operator*.

Moreover, another class of local edge invariants can be obtained as the average value of the  $\rightarrow$  *vertex sums* of the two incident vertices raised to a variable exponent  $\lambda$  [Randić, Balaban *et al.*, 2001]:

$$\mathcal{L}_{ij} = \left( \frac{VS_i[\mathbf{M}] + VS_j[\mathbf{M}]}{2} \right)^\lambda$$

Another general class of edge invariants was defined as the harmonic mean of the edge invariants of the edges linked to the edge  $i$ – $j$  [Alikhanidi and Takahashi, 2006]:

$$\mathcal{L}_{ij} = \frac{\delta_i + \delta_j - 2}{\sum_{k=1}^A a_{ik} \cdot (1/\mathcal{L}_{ik}) + \sum_{k=1}^A a_{jk} \cdot (1/\mathcal{L}_{jk})} \quad k \neq i \neq j$$

where the numerator is the total number of edges incident to the edge  $i$ – $j$ ;  $\delta$  is the number of edges incident to a vertex, that is, the  $\rightarrow$  *vertex degree*. In the denominator, the first summation accounts for contributions from edges incident to the  $i$ th vertex, while the second one for contributions from edges incident to the  $j$ th vertex.

📖 [Klopman, Raychaudhury *et al.*, 1988; Klopman and Raychaudhury, 1990; Balaban and Balaban, 1991; Balaban, Ciubotariu *et al.*, 1991; Balaban and Balaban, 1992; Balaban, Filip *et al.*, 1992; Bonchev and Kier, 1992; Ivanciuc, Balaban *et al.*, 1992; Kier and Hall, 1992a; Balaban and Diudea, 1993; Bonchev, Kier *et al.*, 1993; Balaban, 1992, 1994c, 1995b; Diudea, Horvath *et al.*, 1995a; Medeleanu and Balaban, 1998]

- **localized effect**  $\equiv$  *polar effect*  $\rightarrow$  electronic substituent constants
- **local polarity index**  $\rightarrow$  electric polarization descriptors
- **local profiles**  $\rightarrow$  molecular profiles



- **local quantum-chemical properties** → quantum-chemical descriptors
- **local simple flexibility index** → flexibility indices (⊙ global flexibility index)
- **local softness** → quantum-chemical descriptors (⊙ softness indices)
- **local spectral moment** → edge adjacency matrix
- **local synthetic invariant** → iterated line graph sequence
- **local vertex invariants** → local invariants
- **LOEIs**  $\equiv$  *LOcal Edge Invariants* → local invariants
- **LOEL**  $\equiv$  *Lowest-Observed-Effect Level* → biological activity indices (⊙ toxicological indices)
- **log  $D_{pH}$**   $\equiv$  *octanol–water distribution coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log  $K_{mw}$**   $\equiv$  *micelle–water partition coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log  $K_{oc}$**   $\equiv$  *soil sorption partition coefficient* → physico-chemical properties (⊙ partition coefficients)
- **log  $K_{ow}$**   $\equiv$  *octanol–water partition coefficient* → lipophilicity descriptors
- **log  $P$**   $\equiv$  *octanol–water partition coefficient* → lipophilicity descriptors
- **London cohesive energy** → Hildebrand solubility parameter
- **lone-pair electrons index** → electronic descriptors
- **lone-pair electrostatic interaction** → electronic descriptors
- **longest walk connectivity index** → connectivity indices (⊙ walk connectivity indices)
- **long hafnian** → algebraic operators (⊙ determinant)
- **LOO technique**  $\equiv$  *leave-one-out technique* → validation techniques (⊙ cross-validation)
- **loops** → graph
- **lopping centric information index** → centric indices
- **Lovasz–Pelikan index** → spectral indices (⊙ eigenvalues of the adjacency matrix)
- **LOVIs**  $\equiv$  *LOcal Vertex Invariants* → local invariants
- **Löwdin population analysis** → quantum-chemical descriptors
- **Lowest-Observed-Effect Level** → biological activity indices (⊙ toxicological indices)
- **lowest unoccupied molecular orbital** → quantum-chemical descriptors
- **lowest unoccupied molecular orbital energy** → quantum-chemical descriptors
- **LUDI energy function** → scoring functions
- **Lu index** → hyper-Wiener-type indices
- **luminal over-saturation number** → property filters (⊙ drug-like indices)