

D

➤ **DAI indices** → atom-type-based topological indices

■ **DARC/PELCO analysis**

DARC/PELCO analysis, that is, a topological method dealing with structural environments, was originally proposed and, then, later refined and broadened by Dubois [Dubois, Laurent *et al.*, 1966, 1973a, 1973b, 1976; Dubois, 1976].

The method is based on a combination of the DARC system (**Description, Acquisition, Retrieval and Computer-aided design**) and the PELCO (**Perturbation of an Environment Limited Concentric and Ordered**) procedure. Moreover, it accounts for the simultaneous representation of all the data set compounds and the population of the compounds structurally contained in them; the data set compounds, which are those compounds for which a molecular property has been experimentally evaluated, generate an ordered multidimensional space that constitutes a → *hyperstructure*.

Each molecule is represented by an ordered → *chromatic graph*, which describes the topological and chemical nature of each site; vertex chromatism corresponds to the chemical nature of atoms, edge chromatism to the bond multiplicity.

The hyperstructure is built following the operations of focalization, organization, ordering, and chromatic evaluation of each data set molecule.

The superimposition of all the ordered graphs of the data set molecules provides the hyperstructure whose central topological vertex corresponds to the focus, and the environment is organized in concentric layers A–B where each vertex corresponds to an atom present in at least one compound.

The *focus* is the → *maximum common substructure* among the data set compounds. The *environment* is organized in concentric layers centered on the focus and is *limited* and *concentric* (ELC). The vertices at an odd distance from the focus belong to layer A; the vertices at an even distance belong to layer B. Each pair of successive A–B layers starting from the focus constitutes an *environment limited to B* (EB). The environment is *ordered* (ELCO) in the sense that each site (vertex) is located unambiguously by means of a topological coordinate (A_i or B_{ij}); each topological coordinate gives a development direction for the environment.

More specifically, from the first concentric environment of the hyperstructure, the different substitution sites give rise to the main development directions starting from the focus. Note that each hyperstructure vertex actually is a topochemical site, which can contain more than one atom and can be labeled differently, in accordance with the chemical nature of the atoms in it.

The hyperstructure can be mathematically represented by the **DARC/PELCO matrix**, which has n rows, the number of data set compounds, and N_S columns, the topochromatic sites of the hyperstructure. Each row of the DARC/PELCO matrix is called **topochromatic vector**, denoted as I_i , and directly accounts for the overall topology of the one molecule. The **DARC/PELCO descriptors** of a molecule are the elements of its topochromatic vector and are binary variables I_{is} equal to 1 if the s th topochromatic site of the hyperstructure contains an atom of the i th molecule and zero otherwise.

The **DARC/PELCO model** is defined as

$$\hat{y}_i = y_0 + \sum_{s=1}^{N_S} b_s \cdot I_{is}$$

where y_0 is the response of the parent molecule, that is, the focus of the hyperstructure, and b_s are the regression coefficients called *perturbations*. In spite of the formal analogy with the \rightarrow *Free-Wilson model*, where the contribution to the biological activity of each group in a substitution site is considered additive and independent of the structural variation in the rest of the molecule, the DARC/PELCO model considers the contribution of a substituent group to the biological activity as the sum of ordered perturbations given by all the vertices starting from the focus and characterizing that group (Figure D1).

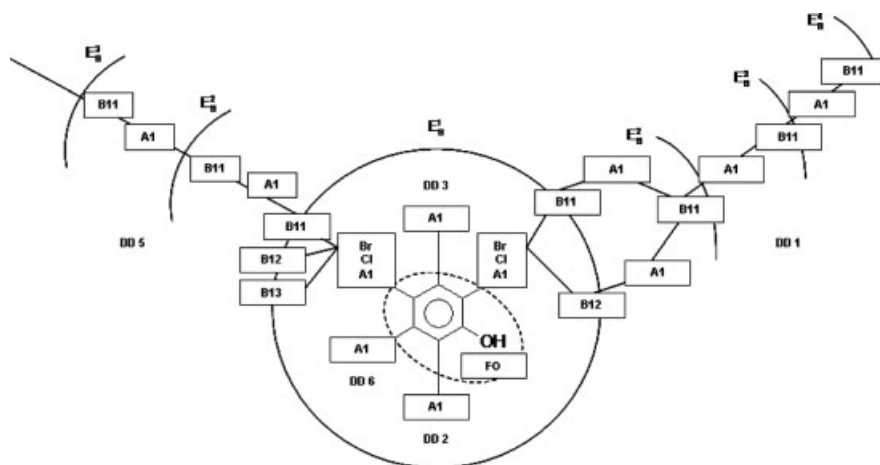


Figure D1 Example of DARC/PELCO hyperstructure. The focus is constituted by the phenol.

A particular advantage of this method is the determination of the structural area for reliable predictions. This area, called **proference**, consists of all the structures generated by the hyperstructure that do not belong to the population of data set compounds (**population trace**). In other words, predictable structures are localized in the hyperstructure and their activity is predicted by interpolation, using the corresponding topochromatic vector where each vertex present ($I_{is} = 1$) refers to a site existing in at least one data set compound. Different levels of reliability are determined depending on the extent to which the structure is surrounded by data set compounds.

📖 [Dubois, Laurent *et al.*, 1967, 1975a, 1975b; Duperray, Chastrette *et al.*, 1976a, 1976b; Dubois, Mercier *et al.*, 1979, 1986; Mercier and Dubois, 1979; Dubois, Chrétien *et al.*, 1980; Panaye, MacPhee *et al.*, 1980; Bawden, 1983; Doucet, Panaye *et al.*, 1983; Dubois, Sicouri *et al.*, 1984; Dubois and Sobel, 1985; Dubois, Panaye *et al.*, 1987; De La Guardia, Carrión *et al.*, 1988; Bonchev, 1989; Attias and Dubois, 1990; Mercier, Troullier *et al.*, 1990; Mercier, Mekenyan *et al.*, 1991; Dubois and Loukianoff, 1993; Mekenyan, Mercier *et al.*, 1993; Panaye, Doucet *et al.*, 1993; Carabédian and Dubois, 1998; Dubois, Doucet *et al.*, 1999]

- **DARC/PELCO descriptors** → DARC/PELCO analysis
- **DARC/PELCO matrix** → DARC/PELCO analysis
- **DARC/PELCO model** → DARC/PELCO analysis
- **Daren fitness function** → regression parameters
- **Dash–Behera steric density parameter** \equiv *steric density parameter* → steric descriptors
- **data** → data set
- **data distance matrix** → similarity/diversity
- **data matrix** → data set

■ data set

This is a collection of *objects* described by one or more *variables*. An **object** is a basic unit in data analysis; for example, an individual, a molecule, an experiment, and a sample. Each object is described by one or more measurements, called **data**. A **variable** represents a characteristic of the objects that may take any value from a specified set, for example, a physico-chemical property, a molecular descriptor.

A data set is often considered as a sample from a population and the sample parameters calculated from the data set as estimates of the population parameters (→ *statistical indices*). Moreover, it is used to calculate statistical models such as quantitative → *structure/response correlations*. In this case, the data set is organized into a **data matrix X** with n rows and p columns where each row corresponds to an object of the data set and each column to a variable; therefore, each matrix element x_{ij} represents the value of the j th variable for the i th object ($i = 1, \dots, n$; $j = 1, \dots, p$).

Data set variables can be distinguished by their role in the models as independent and dependent variables. **Independent variables** (or **explanatory variables**, **predictor variables**) are those variables assumed to be capable of taking part of a function suitable to model the response variable. **Dependent variables** (or **response variables**) are variables (often obtained from experimental measures) for which the interest is to find a statistical dependence on one or more independent variables. Independent variables constitute the data matrix **X**, whereas dependent variables are collected into a matrix **Y** with n rows and r columns ($r = 1$ when only one response variable is defined) (Figure D2). Moreover, additional information about the belonging of objects to different classes can be stored in a class vector **c**, which consists of integers from 1 to G ; each integer indicates a class and G is the total number of classes.

→ *Regression analysis* is the methodology searching for mathematical models describing relationships between a set of independent variables and a response variable **y** (or a set of response variables **Y**), whereas → *classification* is the methodology searching for mathematical models describing relationships between a set of independent variables and the classification vector **c**, able to assign each object to its proper class.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & & \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & & x_{np} \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1r} \\ y_{21} & y_{22} & & \\ \dots & \dots & \dots & \dots \\ y_{n1} & y_{n2} & & y_{nr} \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ c_n \end{pmatrix}$$

Figure D2 n objects described by an \mathbf{X} data matrix of p independent variables, a \mathbf{Y} matrix of r responses, and a \mathbf{c} vector of G class assignments of the objects. c_1, \dots, c_n are G integer numbers or labels representing the class assignment of the n objects.

In several cases, before applying classification methods, the class vector \mathbf{c} is transformed into a set of G binary vectors by a procedure called **class unfolding**. This procedure consists in assigning each object a binary vector that is comprised of $G - 1$ values equal to 0 and one value equal to 1 corresponding to the class the object belongs to (Figure D3). In other words, class unfolding transforms the n -dimensional vector \mathbf{c} into a binary matrix \mathbf{C} with n rows (the objects) and G columns (the classes).

$$\mathbf{c} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \end{pmatrix} \Rightarrow \mathbf{C} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Figure D3 Example of class unfolding of 10 objects assigned to three different classes.

To estimate the predictive capabilities of a model by \rightarrow *validation techniques*, the data set can be split into different parts: the **training set** (or **learning set**), that is, the set of objects used for model building, the **test set**, that is, the set of objects used to optimize the goodness of prediction of a model obtained from the training set, and the **external evaluation set** (or **evaluation set**), that is, a new data set used to perform further external validation of the model obtained from the training set.

The use of several variables in describing objects increases the complexity of the data and therefore the \rightarrow *model complexity*: noise, variable correlation, redundancy of information provided by the variables, and unbalanced information and not useful information give the data an intrinsic complexity that must be resolved. This happens in the case of spectra, each constituted, for example, by 800–1000 digitalized signals, which are highly correlated variables. Usually, \rightarrow *variable reduction* and \rightarrow *variable selection* improve the quality of models (in particular, their predictive power) and information extracted from models. \rightarrow *Chemometrics* provides several useful tools able to check the different kinds of information contained in the data [Frank and Todeschini, 1994].

Data sets can be analyzed by \rightarrow *exploratory data analysis*, usually based on multivariate techniques, such as \rightarrow *principal component analysis*; \rightarrow *cluster analysis* allows the evaluation of similarity/diversity among the objects or, by transposing the **X** data matrix, among the variables. Similarity and diversity among the objects of a data set are encoded in the \rightarrow *similarity matrix* and in the \rightarrow *data distance matrix*, respectively.

- **Daylight fingerprints** \rightarrow substructure descriptors (\odot fingerprints)
- **Daylight-FingerPrint drug-like Score** \rightarrow scoring functions (\odot Property and Pharmacophore Features Score)
- **dBx descriptors** \rightarrow shape descriptors
- **D/D index** \rightarrow molecular geometry
- **decimal adjacency vector** \rightarrow adjacency matrix
- **decomposition** \rightarrow equivalence classes
- **degeneracy of molecular descriptors** \rightarrow molecular descriptors
- **degradability** \rightarrow environmental indices (\odot persistence)
- **degree-adjacency matrix** $\equiv \chi$ *matrix* \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **degree centrality** \rightarrow center of a graph
- **degree complexity** \equiv *mean information content on the vertex degree magnitude* \rightarrow topological information indices
- **degree distance of the graph** \rightarrow Schultz molecular topological index
- **degree of unsaturation** \rightarrow multiple bond descriptors
- **delocalization** \rightarrow delocalization degree indices

■ delocalization degree indices

Delocalization degree indices are molecular descriptors accounting for the π -electron mobility in a molecule.

To indicate this peculiar behavior of the π -electrons in a molecule, several different terms were historically used. The term **conjugation** (or π -conjugation) referred to unsaturated hydrocarbons was introduced to indicate a molecule with alternating saturated and unsaturated bonds. Then, the term conjugation can be preferably referred to topological aspects of a molecule. The term **resonance** was derived in the framework of the *Valence Bond* (VB) theory, accordingly to which the reference structure corresponds to the most stable structure and resonance energies are obtained by taking into account the contributions of all other (less stable) resonance forms. The term **delocalization** was derived in the framework of the *Molecular Orbital* (MO) theory, according to which the wave function is composed by molecular orbitals already delocalized over the entire molecule and delocalization energies are calculated with respect to a reference system with completely localized orbitals, for example, the atomic orbitals [Bruschi, 2005].

The concept of delocalization is also closely related to the concept of **aromaticity**. In effect, the concept of aromaticity is one of the most important general concepts for an understanding of organic chemistry and physico-chemical properties [Lloyd, 1996]. The term aromaticity has a long history in the chemistry development and dates back to the first use by Kekulé, Erlenmeyer, and Körner in the 1860s [Kekulé, 1865, 1866a, 1866b; Erlenmeyer, 1866; Körner, 1869, 1874].

Aromaticity, as well as the terms of aromatic character and resonance, is associated with the ground-state properties of cyclic π -electron compounds which (a) are more stable than the chain

analogues or classical localized structures due to an energy called resonance energy, (b) have bond lengths between those typical of single and double bonds, (c) have a π -electron ring current that is induced when the molecule is exposed to external magnetic fields, leading to specific values of ^1H -NMR chemical shifts and increased values of the \rightarrow *magnetic susceptibility*. Moreover, (d) from the chemical reactivity point of view, due to the tendency to maintain the π -electron structure, substitution is preferred to addition [Krygowski and Cyranski, 2001].

Several definitions of the molecule aromaticity were given, each based on one or more of the aforementioned properties of the so-called aromatic compounds.

However, for a general definition, *aromaticity* may be assumed to be a phenomenon that occurs, even though to different extents, when molecules show all the aforementioned properties.

In fact, aromaticity is an ‘excess property’, which means a deviation from a property additive scheme, and consequently quantitative measures of aromaticity need the assumption of some reference state.

Delocalization degree indices, commonly also called **resonance indices** and **aromaticity indices**, are molecular descriptors giving measures, in general, of the electron delocalization or, more specifically, of the aromatic character of compounds; these indices are of different kind, depending on different theoretical, physico-chemical, experimental, and geometrical aspects of the aromatic behavior. These can be distinguished into energy-based, geometry-based, and magnetic property-based indices.

The first theoretical explanation – based on quantum mechanical approach – of the aromatic character of a molecule was given by physical chemist E. Hückel in 1931. The **Hückel’s rule** estimates whether a planar ring molecule will have aromatic properties and was expressed as the $4n + 2$ rule formulated by von Doering in 1951. A cyclic ring molecule satisfies Hückel’s rule when the number of its π electrons equals $4n + 2$, n being zero or any positive integer.

Hückel’s rule is not valid for many compounds containing more than three fused aromatic nuclei in a cyclic fashion like in pyrene or coronene.

Strictly related to the concept of aromaticity is the resonance energy of a molecule.

The **resonance energy** RE is a theoretical quantity introduced to explain the stability of benzene and is used for predicting the electron delocalization degree of conjugated systems [Wheland, 1955; Salem, 1966; Randić, 1989; Trinajstić, 1991, 1992]. The general definition of resonance energy is

$$\text{RE} = E_{\pi}(\text{conjugated molecule}) - E_{\pi}(\text{reference structure})$$

where E_{π} is the π -electron energy.

The *resonance energy per electron* (REPE) is a size-independent quantity obtained by dividing the total resonance energy for the number of π -electrons.

Aromaticity indices defined in terms of resonance energy are commonly called **resonance indices**.

A number of resonance measures were proposed based on different theoretical quantum-chemistry approaches [Krygowski, Cyranski *et al.*, 2000; Randić, 2003a]. However, it was recognized that the main difference among the proposed approaches lies in the definition of the specific measure and the nonconjugated reference structure and not in the use of different MO theories.

The most common resonance indices are reported below.

- **Hückel resonance energy**

This is the classical definition of resonance energy obtained from a reference structure containing carbon–carbon isolated double bonds with π -electron energy of ethylene [Streitweiser, 1961]:

$$\text{HRE} = E_{\pi}(\text{conjugated molecule}) - 2 \cdot N_{\text{C}=\text{C}}$$

where $N_{\text{C}=\text{C}}$ is the number of double bonds in a Kekulé structure of the molecule.

This criterion to define the resonance energy fails in many cases, overestimating aromaticity of rather unstable compounds.

📖 [Gutman and Trinajstić, 1972]

- **Dewar resonance energy**

This is the resonance energy defined as the difference between the π -energy E_{π} of the compound and the reference energy estimated by the bond contributions of the corresponding nonconjugated structure, in the framework of SCF π -MO approximation:

$$\text{DRE} = E_{\pi}(\text{conjugated molecule}) - \sum_b n_b \cdot E_b$$

where n_b is the number of bonds having bond energy E_b [Dewar and Longuet-Higgins, 1952; Dewar and Gleicher, 1965; Dewar, 1969; Dewar and de Llano, 1969; Dewar, Kohn *et al.*, 1971].

This definition of resonance energy makes a clear distinction between aromatic (positive DRE), antiaromatic (negative DRE) and nonaromatic (near-zero DRE) conjugated molecules. Extensive tables of resonance energies were also obtained in the framework of the HMO approximation by Hess and Schaad [Hess Jr. and Schaad, 1971a, 1971b, 1973; Hess Jr., Schaad *et al.*, 1972]. Moreover, extensions and modifications of the calculation of the reference structure energy were proposed by other authors [Baird, 1969, 1971].

📖 [Dewar, Harget *et al.*, 1969; Dewar and Harget, 1970a; Dewar and Trinajstić, 1970; Schaad and Hess Jr., 1972; Hess Jr., Schaad *et al.*, 1975]

Other resonance energy indices (RE) are derived from the molecular structure and do not involve direct energy measures. Some of them were defined by fitting resonance energy values of a number of compounds, while others were derived from molecular topology.

The most popular of them are listed below.

- **Green resonance energy**

This is defined for benzenoid systems as

$$\text{GRE} = \frac{AB}{3} + \frac{AB^*}{10}$$

where AB is the total number of aromatic bonds and AB^* is the number of bonds contained in one benzene ring and linking two others [Green, 1956].

- **Bartell resonance energy**

This is a resonance energy index that relates π -energy to the Pauling \rightarrow bond order P by the following expression:

$$\text{BRE} = \frac{4}{3} \times \beta \times \left(N_{\text{C}=\text{C}} - \sum_b P_b^2 \right)$$

where the summation runs over the π bonds, β is the resonance integral, and $N_{\text{C}=\text{C}}$ is the number of formal double bonds [Bartell, 1963, 1964]. The Pauling bond orders are obtained from the analysis of the Kekulé resonance structures.

- **Carter resonance energy**

This is defined for benzenoid systems as

$$\text{CRE} = 0.6 \times N_{\text{C}=\text{C}} + 1.5 \times \ln K - 1$$

where 0.6 and 1.5 are empirical parameters, $N_{\text{C}=\text{C}}$ the number of double bonds in one Kekulé structure, and K is the \rightarrow *Kekulé number* of the molecule which is the number of Kekulé structures in a molecule [Carter, 1949].

- **Herndon resonance energy**

This is defined for benzenoid systems as

$$\text{HRE} = 1.185 \times \ln K$$

where 1.185 was obtained by fitting the Dewar-deLlano SCF π -MO resonance energy values and K is the \rightarrow *Kekulé number* of the molecule [Herndon, 1973b, 1974b; Herndon and Ellzey Jr, 1974].

- **Wilcox resonance energy**

This is defined for general aromatic systems including alternant four-membered rings as

$$\text{WRE} = 0.445 \times \ln \text{CSC} - 0.17 \times N_4$$

where CSC is the \rightarrow *corrected structure count* and N_4 is the number of four-membered rings [Wilcox Jr, 1968, 1969]. Parameter values were obtained by fitting the Hess-Schaad resonance energy values [Hess Jr., Schaad *et al.*, 1975].

- **McClelland resonance energy**

This is a measure of resonance energy defined in terms of the number of atoms A_π and bonds B_π involved in a π -system as [McClelland, 1971]:

$$\text{MCRE} = 0.92 \times \sqrt{2 \cdot A_\pi \cdot B_\pi}$$

- **Hosoya resonance energy**

This is a measure of aromatic stability of conjugated systems defined as [Hosoya, Hosoi *et al.*, 1975]

$$\text{HoRE} = \tilde{Z} - Z$$

where Z is the \rightarrow *Hosoya Z index* and \tilde{Z} is the \rightarrow *stability index* defined as the sum of the absolute values of the coefficients c_{2i} appearing alternatively in the \rightarrow *characteristic polynomial* of the adjacency matrix:

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

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

In acyclic graphs, \tilde{Z} and Z are equal, the \rightarrow *Z-counting polynomial* being in this case coincident with the characteristic polynomial.

- **Aihara resonance energy**

This is a measure of aromatic stability of conjugated systems defined as

$$\text{ARE} = 6.0846 \cdot \log\left(\frac{Z^*}{Z}\right)$$

where Z is the \rightarrow *Hosoya Z index* and Z^* is defined as

$$Z^* = \prod_{i=1}^A (1 + \lambda_i^2)^{1/2}$$

where λ_i denotes the eigenvalues of the \rightarrow *characteristic polynomial* of the adjacency matrix of the molecular graph and A is the number of atoms [Aihara, 1976, 1977b, 1977a, 1978].

- **topological resonance energy**

This is a resonance energy index defined as [Gutman, Milun *et al.*, 1977; Trinajstić, 1992]:

$$\text{TRE} = E_\pi - E_{\text{REF}} = \sum_{i=1}^A g_i \cdot (\lambda_i - \lambda_i^{\text{REF}})$$

where E_π is the Hückel π -electron energy of the molecule and E_{REF} is the reference energy obtained from the \rightarrow *matching polynomial* of the corresponding molecular graph; λ_i denotes the eigenvalues of the \rightarrow *characteristic polynomial* of the adjacency matrix of the molecular graph and λ_i^{REF} the eigenvalues of the matching polynomial, g_i is the occupation number on the i th molecular orbital that can take values 0, 1, or 2, and A is the number of atoms.

The average TRE index is defined as TRE/N_π where N_π is the number of π electrons in the molecule.

📖 [Babic, Brinkmann *et al.*, 1997; Jurić, Nikolić *et al.*, 1997]

- **Krygowski bond energy**

Based on the concept of **Pauling's bond number** n [Pauling, 1947], n being the number of shared electron pairs in the bond, the Krygowski bond energy is defined as the sum over all the π bonds of a function of the differences of bond lengths between the reference and the actual bonds [Krygowski, Ciesielski *et al.*, 1995]:

$$\text{KBE} = \sum_{b=1}^{AB} E(n)_b = \sum_{b=1}^{AB} 87.99 \cdot \exp[2.255 \cdot (1.533 - R(n)_b)]$$

where AB is the number of π bonds, $E(n)$ is the energy of a bond with a Pauling's bond number n , 87.99 is the bond energy of a single C–C bond ($E(1)$), and $R(n)$ is the length of the bond with bond number n . This equation was derived from the empirical relationship that combines bond energy and bond number:

$$E(n) = E(1) \cdot n^k$$

where k is an empirical constant.

Table D1 Resonance indices for some benzenoid compounds.

Compound	K	Dewar	Green	Aihara	Herndon	TRE
Benzene	2	0.869	2.00	0.273	0.821	0.276
Naphthalene	3	1.323	3.67	0.389	1.302	0.390
Anthracene	4	1.600	5.33	0.475	1.643	0.476
Phenanthrene	5	1.933	5.43	0.546	1.907	0.546
Pyrene	6	2.098	6.43	0.562	2.123	0.592
Naphthacene	5	1.822	7.00	0.553	1.907	0.558
3,4-Benzophenanthrene	8	2.478	—	0.687	2.464	—
1,2-Benzanthracene	7	2.291	7.10	0.643	2.306	—
Chrysene	8	2.483	7.20	0.688	2.464	0.684
Triphenylene	9	2.654	7.30	0.739	2.604	—
Perylene	9	2.619	8.20	0.598	2.604	—

Data from [Swinborne-Sheldrake, Herndon *et al.*, 1975; Aihara 1977b; Trinajstić 1992]. K is the Kekulé number.

Several aromaticity indices were defined in terms of bond lengths r_b and bond orders π_b , exploiting the typical structural features of aromatic compounds [Krygowski, Ciesielski *et al.*, 1995]. The most known aromaticity indices of this kind are listed below and some values provided in Table D3.

• Julg–François index (A_j)

Based on the idea that bond alternation causes the aromatic character to decrease, the first aromaticity index based on the \rightarrow *molecular geometry* was proposed by Julg and François [Julg and François, 1967] as

$$A_j = 1 - \frac{225}{AB} \times \sum_{b=1}^{AB} \left(1 - \frac{r_b}{\bar{r}_\pi} \right)^2$$

where AB is the number of π peripheral bonds involved in the aromatic system, r_b the geometric distance of the considered π bond, and \bar{r}_π is the average π bond length. The constant 225 results from the normalization conditions to obtain a value of 0 for the Kekulé structure of benzene and 1 for any system with all bonds of equal length.

• bond alternation coefficient (BAC)

This is a purely geometric aromaticity index defined as

$$BAC = \sum_b (r_{b+1} - r_b)^2$$

where r_{b+1} and r_b are consecutive bond lengths in the rings; the summation runs over all π bonds of the molecule (or fragment); note that the sequence of bonds on which the sum runs is obviously well defined for monocyclic systems but less defined for polycyclic systems [Binsch and Heilbronner, 1968].

• **Bird aromaticity indices** ($\equiv I_R$ aromaticity indices)

These are general aromaticity indices based on the statistical degree of uniformity of the bond orders of the ring periphery and distinguished in I_5 , I_6 , and $I_{5,6}$ for five-, six-membered rings, and five-, six-fused rings, respectively [Bird, 1985, 1986]. These indices are defined as

$$I_R = 100 \cdot \left(1 - \frac{V}{V_R}\right)$$

where V_R is a constant depending on the considered ring (e.g., $V_R = 35$ for a five-membered heterocycle and $V_R = 33.3$ for a six-membered heterocycle; for systems consisting of a five-membered and a six-membered ring fused together, $V_R = 35$).

The term V is defined in terms of the bond order variance:

$$V = \frac{100}{\bar{\pi}} \cdot \sqrt{\frac{\sum_b (\pi_b - \bar{\pi})^2}{AB}}$$

where the sum runs on the π bonds, AB is the number of π bonds, $\bar{\pi}$ is the average bond order, and π_b is the \rightarrow Gordy's bond order of the b bond defined as

$$\pi_b \equiv \pi_{ij} = \frac{a}{r_b^2} - b$$

where a and b are constants depending on the π bond type and r_b is the bond length.

• **RC index**

This is an aromaticity index based on the idea of *ring current* whose magnitude is determined by its weakest link in the ring [Jug, 1983, 1984]. The weakest link is considered as the bond with the minimum total bond order:

$$RC = \min_b (\pi_b)$$

where b runs over all the π bond system and π_b denotes the bond orders.

By analogy, a complementary aromatic index, called **maximum bond length** (LB) was defined as the longest bond length in the π -electron system under consideration [Krygowski and Ciesielski, 1995; Krygowski, Ciesielski *et al.*, 1995]:

$$LB = \max_b (r_b)$$

where b runs over all the π bonds and r_b denotes the π bond lengths.

• **HOMA index** (\equiv Harmonic Oscillator Model of Aromaticity index)

This index is based on the degree of alternation of single/double bonds, measuring the bond length deviation from the optimal length attributed to the typical aromatic state [Kruszewski and Krygowski, 1972; Krygowski, 1993; Krygowski and Ciesielski, 1995; Krygowski, Ciesielski *et al.*, 1995; Krygowski, Cyranski *et al.*, 1996]. The *HOMA* index is defined as:

$$HOMA = \frac{1}{n_k} \cdot \sum_k \left[1 - \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (r_k^{opt} - r_b)^2 \right]$$

where the first summation runs over n_k aromatic bond types, AB_k is the number of π bonds of the k -th aromatic bond type, r_b is the actual bond length, α_k and r_k^{opt} denote a numerical constant and the typical aromatic bond length referring to the k -th aromatic bond type (Table D2), respectively.

The *HOMA* index for the k th aromatic bond type can be decomposed in two terms describing two different contributions to a decrease in aromaticity; one contribution, due to the bond elongation, is called *EN* and the other one, due to the bond length alternation, is called *GEO*:

$$\begin{aligned} HOMA_k &= 1 - \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (r_k^{opt} - r_b)^2 = 1 - \left[\alpha_k \cdot (r_k^{opt} - \bar{r}_k)^2 + \frac{\alpha_k}{AB_k} \cdot \sum_{b=1}^{AB_k} (\bar{r}_k - r_b)^2 \right] = \\ &= 1 - EN_k - GEO_k \end{aligned}$$

Table D2 Values of parameter α and optimal bond lengths r^{opt} for different aromatic bond types.

Bond	α	r^{opt}	Bond	α	r^{opt}
C \approx C ^a	257.7	1.388	C \approx P	118.91	1.698
C \approx C	98.89	1.397	C \approx S	94.09	1.677
C \approx N	93.52	1.334	N \approx N	130.33	1.309
C \approx O	157.38	1.265	N \approx O	57.21	1.248

^a1,3-butadiene.

• **HOSE index** (\equiv *Harmonic Oscillator Stabilization Energy index*)

This is an aromaticity index based on the energy deformation derived from a simple harmonic oscillator potential and defined as [Krygowski and Wiecowski, 1981; Krygowski, Anulewicz *et al.*, 1983, 1995; Bird, 1997]:

$$HOSE = 301.15 \cdot \left[\sum_{b=1}^{n_1} k'_b \cdot (r'_b - r_0^s)^2 + \sum_{b=1}^{n_2} k''_b \cdot (r''_b - r_0^d)^2 \right]$$

where the first summation runs on single bonds and the second one on double bonds; r_0^s and r_0^d denote the reference bond lengths for single and double bonds, respectively; k'_b and k''_b are the corresponding force constants of the b bond; and r'_b and r''_b stand for the lengths of π bonds in the molecule. The force constants are calculated accordingly to the relation:

$$k_b = a + b \cdot r_b$$

where a and b are two constants and r_b is the actual π bond length.

HOSE index allows estimation of aromaticity also in cases of very small changes in aromatic character of the molecule in question.

• **Dewar index** (D_p)

This is a local aromaticity index [Dewar and Longuet-Higgins, 1952; Dewar, 1969], defined in the framework of the perturbation molecular orbital (PMO) theory as the energy difference

between a conjugated radical (R) and a conjugated hydrocarbon (H) obtained by joining a new carbon atom to the a, b, c, \dots atoms of the radical:

$$D_P = E(H) - E(R) = 2 \cdot (C_a + C_b + \dots)$$

where P indicates the particular site of the molecule and C_a, C_b, \dots , are the coefficients of the normalized nonbonding molecular orbital of the radical, corresponding to the atom a, b, \dots

Approximate relationships between the Dewar index and topological descriptors were derived by Gutman [Gutman, 1977].

- **benzene-likeliness index (B_L)**

This is an aromaticity index calculated from molecular topology. It is defined in terms of the first-order \rightarrow *valence connectivity index* ${}^1\chi^v$ divided by the number B of bonds of the molecule (hydrogen bonds excluded), and normalized on the benzene molecule [Kier and Hall, 1986]:

$$B_L = \frac{{}^1\chi^v/B}{2/6} = 3 \times \frac{{}^1\chi^v}{B}$$

where ${}^1\chi^v = 2$ and $B = 6$ for benzene. Some typical values of this index are $B_L(\text{benzene}) = 1$, $B_L(\text{thiophene}) = 0.97$, $B_L(\text{pyrrole}) = 0.95$, $B_L(\text{pyridine}) = 0.93$, and $B_L(\text{imidazole}) = 0.86$.

Other aromaticity indices were derived from the magnetic characteristics of aromatic compounds, such as exaltation of diamagnetic susceptibility and ${}^1\text{H}$ -NMR shielding.

The most important are discussed below.

- **diamagnetic susceptibility exaltation (Λ)**

It is a parameter that exploits the enhanced diamagnetic anisotropy (due to the “ring current”) of benzenoid aromatic systems with respect to the noncyclic delocalized counterpart and can be calculated both experimentally and theoretically. It is defined as the difference between the \rightarrow *magnetic susceptibility* of the aromatic system M and that one of the reference structure R [Dauben, Wilson *et al.*, 1968]:

$$\Lambda_M = \chi_M - \chi_R$$

Positive values reveal the presence of ring currents and therefore of benzenoid aromatic systems. The diamagnetic susceptibility exaltation highly depends on the ring size.

Significant relationships between aromaticity measured by diamagnetic susceptibility exaltation and derivatives of the total molecular valence for molecule and HOMO orbital, calculated *ab initio* in the framework the \rightarrow *Density Function Theory*, were also found [Balawender, Komorowski *et al.*, 1998].

- **NICS index (\equiv Nucleus-Independent Chemical Shift index)**

The NICS index was defined as the negative value of the absolute shielding obtained by ${}^1\text{H}$ -NMR spectroscopy and computed at a ring center or at some other interesting point of the system [Schleyer, Maerker *et al.*, 1996; Krygowski and Cyranski, 2001]. Rings with negative NICS values are defined as aromatic, and the more negative NICS more aromatic the rings are. Consequently, antiaromatic systems show positive NICS values. NICS values are sensitive to the basis set used for calculation; it has been suggested to use NICS values obtained from the (6-31 + G^*) basis set, whenever it is possible.

Some correlation between NICS index and *HOMA* index was found out [Szatyłowicz, Krygowski *et al.*, 2007]. However, criticism about the use of NICS as aromaticity descriptor is made by Lazzeretti [Lazzeretti, 2004], concluding that “... *a quantitative theory of aromaticity based on NICS is epistemologically inconsistent.*”

• Delocalization Index (DI)

The degree of π -delocalization between atoms A and B was quantified in the framework of the \rightarrow *AIM theory*, by the delocalization index DI_{AB} , which is defined as [Poater, Fradera *et al.*, 2003a, 2003b; Poater, Garcia-Cruz *et al.*, 2004; Krygowski, Ejsmont *et al.*, 2004]

$$DI_{AB} = -2 \cdot \iint_{A \ B} \Gamma_{\text{exc}}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

where Γ_{exc} is the exchange-correlation density over the basins of atoms A and B, which are defined from a condition of zero-flux gradient.

Moreover, the mean of all DI of *para*-related carbons in a given six-membered ring, called **Para-Delocalization Index (PDI)**, has been defined as an aromaticity criterion based on electron delocalization. This index is closely related to NICS and *HOMA* indices: the higher the PDI indices the higher the absolute value of NICS and the higher the *HOMA* values, thus reflecting greater aromaticity.

Table D3 Values of aromaticity indices for some compounds.

Compound	<i>HOMA</i>	BAC	A_j	I_6	<i>LB</i>	Δ	NICS ^a
Benzene	1.000	0.000	1.000	100.0	1.397	13.7	−9.7
Naphthalene	0.802	0.088	0.932	81.3	1.424	30.5	−9.9
Anthracene	0.696	0.098	0.889	79.2	1.446	48.6	−13.3/−8.2
Phenathrene	0.727	0.101	0.878	77.1	1.465	46.2	−10.2/−6.5
Pyrene	0.728	0.094	0.916	80.1	1.438	57.3	—
3,4-Benzophenanthrene	0.671	0.107	—	73.8	1.460	—	—
Triphenylene	0.722	0.070	0.906	85.1	1.478	—	—

Data from [Dauben, Wilson *et al.* 1968; Krygowski, Ciesielski *et al.* 1995;

Krygowski and Cyranski 2001; and Randić, 2003a].

^aWhen double, NICS data refer to values relative to the central ring and outer rings, respectively.

📖 Additional references are listed in the thematic bibliography (under *chemical compound classes: conjugated systems*).

- **Delocalization Index** \rightarrow delocalization degree indices
- **delocalized effect** \equiv *resonance effect* \rightarrow electronic substituent constants
- **delta matrix** \rightarrow distance-path matrix
- **delta number** \rightarrow distance-path matrix
- **Dennis similarity coefficient** \rightarrow similarity/diversity (☉ Table S9)

- **dense matrices** → algebraic operators (\odot sparse matrices)
- **dense Wiener matrix** \equiv *path-Wiener matrix* → Wiener matrix
- **density** → physico-chemical properties
- **Density Functional Theory** → quantum-chemical descriptors
- **density index** → adjacency matrix
- **Density Of States** → quantum-chemical descriptors (\odot EIM descriptors)
- **Depczynski fitness function** → regression parameters
- **dependent variables** → data set
- **Description, Acquisition, Retrieval and Computer-aided design** → DARC/PELCO analysis
- **desolvation energy fields** → molecular interaction fields
- **$\det|A + D|$ index** → determinant-based descriptors
- **$\det|A|$ index** → determinant-based descriptors
- **$\det|D|$ index** → determinant-based descriptors
- **determinant** → algebraic operators

■ determinant-based descriptors

These are molecular descriptors defined in terms of the → *determinant* of a matrix representing a → *molecular graph*. Molecular descriptors similar to the determinant-based descriptors are also calculated by using → *permanent* and → *hafnian* of any matrix representing a molecular graph, such as → *per(D) index*, → *shaf(D) index*, and → *lhaf(D) index* [Schultz, Schultz *et al.*, 1992; Schultz and Schultz, 1992]. Moreover, still based on the determinant is the → *characteristic polynomial* of a molecular matrix, which plays a very important role in the calculation of molecular descriptors.

The most popular determinant-based descriptors are discussed below.

• $\det|A|$ index

The determinant of the → *adjacency matrix* **A**. It was observed that this determinant often equals zero and this is a necessary and sufficient condition for the presence of nonbonding molecular orbitals in Hückel theory. The actual numerical value of $\det|A|$ is correlated to the thermodynamic stability of the molecule [Graovac and Gutman, 1978, 1979; Trinajstić, 1992; Gutman and Vidović, 2002a].

• $\det|D|$ index

The determinant of the → *distance matrix* **D**. In an isomeric series, this index takes same values for all compounds, alternating from negative values for isomers with an even number of nonhydrogen atoms to positive values for compounds with an odd number of nonhydrogen atoms [Schultz, Schultz *et al.*, 1990, 1993]. For acyclic alkanes with a number *A* of atoms, the following relation holds [von Knop, Müller *et al.*, 1991]:

$$\det|D| = -(-2)^{A-2} \cdot (A-1)$$

• $\det|A + D|$ index

It is the determinant of the → *adjacency-plus-distance matrix*, which is the matrix resulting from the sum of the → *adjacency matrix* **A** and the → *distance matrix* **D** of a → *H-depleted molecular graph*; this matrix is also used to calculate the → *Schultz molecular topological index* [Schultz, Schultz *et al.*, 1990, 1993; von Knop, Müller *et al.*, 1991]. Demonstrated to be more discriminant

than previously described determinant-based descriptors, the absolute value of this index increases with the size of the molecules, negative for molecules with an even number of nonhydrogen atoms and positive for those with an odd number. Moreover, it decreases with increasing branching in an isomeric series of compounds, that is, the degree of substitution increases.

The logarithm of this index was used to model different \rightarrow *physico-chemical properties* [Cash, 1995c], showing that it can suffer from the drawback that the determinant values are often equal to zero.

The determinant of the 3D adjacency-plus-geometry matrix (${}^b\mathbf{A} + \mathbf{G}$) was also proposed as a molecular topographic descriptor; ${}^b\mathbf{A}$ is the 3D adjacency matrix or \rightarrow *bond length-weighted adjacency matrix* whose entries corresponding to bonded atoms are \rightarrow *bond distances* instead of 1 and \mathbf{G} the \rightarrow *geometry matrix* [Mihalić, Nikolić *et al.*, 1992].

• **general a_N -index (GAI)**

A topological index defined as the absolute value of the determinant of the **orbital interaction matrix of linked atoms (OIMLA)** [Xu, Wang *et al.*, 1992a, 1992b; Xu, 1992]:

$$GAI = |\det(\mathbf{OIMLA})|$$

OIMLA is a symmetric \rightarrow *weighted adjacency matrix* of dimension $2B \times 2B$ whose diagonal elements are the relative energies of the atomic hybrid orbitals (setting at zero the energy of the C_{sp3} orbital) and the off-diagonal elements represent the interaction type of hybrid orbitals, assumed to be proportional to the corresponding overlap integrals. This matrix is derived from a \rightarrow *H-depleted molecular graph* called **orbital interaction graph of linked atoms (OIGLA)**, which is a \rightarrow *directed graph* where arcs (ordered pairs of vertices) are used to describe interactions between hybrid orbitals. Entries equal to zero indicate that no interaction between hybrid orbitals is considered. Both atomic hybrid orbital energies and overlap integrals are obtained by methods of \rightarrow *computational chemistry*.

GAI was found to be a useful index for the discrimination of *cis/trans* isomerism (\rightarrow *cis/trans descriptors*) and to model the chromatographic behavior of phosphorus derivatives.

GAI is an extension to molecules containing heteroatoms and/or multiple bonds of the a_N -index, previously defined only for alkane derivatives. The a_N -index was calculated as the absolute value of the constant term of the characteristic polynomial of **OIMLA** where diagonal entries are zero and off-diagonal entries are calculated in a similar way as for GAI [Yang and Kiang, 1983].

Note that \rightarrow *graph of atomic orbitals* (GAO) is another representation of molecules that accounts for atom orbitals.

📖 [McClelland, 1974; Kiang, 1980, 2008; Graovac, Juvan *et al.*, 1999; Morón, Campillo *et al.*, 2000]

- **detour complement index** \rightarrow detour matrix
- **detour complement matrix** \rightarrow detour matrix
- **detour-delta matrix** \rightarrow detour matrix
- **detour distance** \rightarrow detour matrix
- **detour-distance combined matrix** \rightarrow detour matrix

- **detour distance–topological distance combined matrix** \equiv *detour-distance combined matrix* \rightarrow detour matrix
- **detour/distance quotient matrix** \rightarrow detour matrix
- **detour distance–geometric distance combined matrix** \rightarrow matrices of molecules (\odot Table M3)
- **detour distance/geometric distance quotient matrix** \rightarrow matrices of molecules (\odot Table M2)
- **detour distance–resistance distance combined matrix** \rightarrow matrices of molecules (\odot Table M3)
- **detour distance/resistance distance quotient matrix** \rightarrow matrices of molecules (\odot Table M2)
- **detour distance–topographic distance combined matrix** \rightarrow matrices of molecules (\odot Table M3)
- **detour distance/topographic distance quotient matrix** \rightarrow matrices of molecules (\odot Table M2)
- **detour distance–topological distance combined matrix** \rightarrow matrices of molecules (\odot Table M3)
- **detour distance/topological distance quotient matrix** \equiv *detour/distance quotient matrix* \rightarrow detour matrix
- **detour index** \rightarrow detour matrix

■ detour matrix (Δ)

The detour matrix Δ of a graph G (or **maximum path matrix**) is a square symmetric $A \times A$ matrix, A being the number of graph vertices, whose entry i – j is the length of the longest path from vertex v_i to vertex v_j ($^{\max}p_{ij}$), [Harary, 1969a; Buckley and Harary, 1990; Ivanciuc and Balaban, 1994b; Amić and Trinajstić, 1995; Trinajstić, Nikolić *et al.*, 1997; Randić, DeAlba *et al.*, 1998]:

$$[\Delta]_{ij} = \begin{cases} \Delta_{ij} = |^{\max}p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The length of the longest path between the vertices v_i and v_j is the maximum number of edges that separate the two vertices and is called **detour distance** and denoted as Δ_{ij} .

This definition is exactly the “opposite” of the definition of the \rightarrow *distance matrix* whose off-diagonal elements are the lengths of the shortest paths between the considered vertices. However, the distance and detour matrices coincide for acyclic graphs, there being only one path connecting any pair of vertices.

The maximum value entry in the i th row is called **atom detour eccentricity** $^{\Delta}\eta_i$ (also **vertex path eccentricity** or simply **path eccentricity**):

$$^{\Delta}\eta_i = \max_j([\Delta]_{ij})$$

From the distribution of the element values in the i th row of the detour matrix, the **maximum path degree sequence** of the i th vertex is derived as a local vector-descriptor defined as

$$\{n_{i0}, n_{i1}, \dots, n_{im}, \dots, n_{ik}\}$$

where n_{im} is the number of vertices in the molecular graph located at a detour distance equal to m from the vertex v_i , k is the maximum detour distance in the graph ($^{\max}\Delta$), and n_{i0} is equal to 1 by definition. Analogously, from the distribution of the element values in the upper or lower

triangle of the detour matrix, the **maximum path frequency sequence** is derived as a molecular vector-descriptor defined as

$$\{\Delta F_0, \Delta F_1, \dots, \Delta F_m, \dots, \Delta F_k\}$$

where ΔF_m is the number of detour distances equal to m in the molecular graph; obviously, ΔF_0 equals the number of vertices in the graph.

The **maximum path sum** of the i th vertex, denoted by $MPVS_i$, is a local vertex invariant defined as the sum of the lengths of the longest paths between vertex v_i and any other vertex in the molecular graph, that is,

$$MPVS_i = VS_i(\Delta) = \sum_{j=1}^A [\Delta]_{ij}$$

where VS is the \rightarrow row sum operator.

$A \rightarrow$ *Wiener-type index*, originally called **MPS topological index** [Ivanciuc and Balaban, 1994b] but usually known as **detour index** and denoted by w [Amić and Trinajstić, 1995; Lukovits, 1996b; Lukovits and Razinger, 1997], was proposed as the sum of the detour distances between any two vertices in the molecular graph. It is calculated as

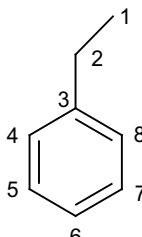
$$w \equiv Wi(\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A MPVS_i$$

where Wi is the \rightarrow *Wiener operator* and $MPVS_i$ is the maximum path sum of the i th vertex.

Other molecular descriptors are derived from the \rightarrow *detour polynomial*.

Example D1

Detour matrix Δ for the H-depleted molecular graph of ethylbenzene. $\Delta\eta_i$ is the atom detour eccentricity, $MPVS_i$ is the maximum path sum of the i th vertex, and w is the detour index.

	Atom	1	2	3	4	5	6	7	8	$MPVS_i$	Atom	$\Delta\eta_i$
	1	0	1	2	7	6	5	6	7	34	1	7
2	1	0	1	6	5	4	5	6		28	2	6
3	2	1	0	5	4	3	4	5		24	3	5
4	7	6	5	0	5	4	3	4		34	4	7
5	6	5	4	5	0	5	4	3		32	5	6
6	5	4	3	4	5	0	5	4		30	6	5
7	6	5	4	3	4	5	0	5		32	7	6
8	7	6	5	4	3	4	5	0		34	8	7

$$w = \frac{1}{2} \times (34 + 28 + 24 + 34 + 32 + 30 + 32 + 34) = 124$$

For edge-weighted graphs, the **weighted detour matrix** (or **edge-weighted detour matrix**), denoted as ${}^w\Delta$, was proposed [Nikolić, Trinajstić *et al.*, 1996a]. The off-diagonal i - j entry is defined as the maximum path weight, that is, the maximum sum of edge weights along the

path between the vertices v_i and v_j , which is not necessarily the longest possible path between them.

A modified detour matrix was proposed by substituting diagonal zero elements with the length of the longest path from each vertex to itself (i.e., the size of the cycle containing the considered vertex). From this modified matrix, the same molecular descriptors defined above can be calculated [Rücker and Rücker, 1998].

The **detour-path matrix**, denoted as Δ_P , is a \rightarrow *combinatorial matrix* analogously defined as the \rightarrow *distance-path matrix* D_P ; it is a square symmetric matrix $A \times A$ whose off-diagonal entry $i-j$ is the count of all paths of any length m ($1 \leq m \leq \Delta_{ij}$) that are included within the longest path from vertex v_i to vertex v_j (Δ_{ij}) [Diudea, 1996a]. The diagonal entries are zero.

Each entry $i-j$ of the detour-path matrix is calculated from the detour matrix Δ as the following:

$$[\Delta_P]_{ij} = \binom{\Delta_{ij} + 1}{2} = \frac{\Delta_{ij}^2 + \Delta_{ij}}{2}$$

that is, as all the possible combinations of two elements taken from $\Delta_{ij} + 1$ elements (binomial coefficient).

The **hyperdetour index** ww can be obtained by applying the \rightarrow *Wiener operator* Wi to the detour-path matrix as

$$ww \equiv Wi(\Delta_P) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta_P]_{ij}$$

or to the symmetric \rightarrow *Cluj-detour matrix* as

$$ww \equiv Wi(\mathbf{SCJ}\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{SCJ}\Delta]_{ij}$$

For acyclic graphs, the hyperdetour index ww is equal to the \rightarrow *hyper-distance-path index* D_P obtained from the distance-path matrix D_P and to the \rightarrow *hyper-Wiener index* WW obtained from the \rightarrow *Wiener matrix*.

The **detour-delta matrix**, denoted as Δ_Δ , is another \rightarrow *combinatorial matrix* derived as the difference between the \rightarrow *detour-path matrix* Δ_P and the \rightarrow *detour matrix* Δ [Janežič, Miličević *et al.*, 2007]:

$$\Delta_\Delta = \Delta_P - \Delta$$

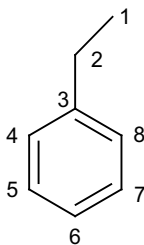
This matrix is a square symmetric matrix of dimension $A \times A$, A is the number of graph vertices, and enumerates the number of all longest paths larger than unity between vertices v_i and v_j in a graph: matrix entries are defined as the following binomial coefficients:

$$[\Delta_\Delta]_{ij} = \begin{cases} \binom{\Delta_{ij}}{2} = \frac{\Delta_{ij} \cdot (\Delta_{ij} - 1)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where Δ_{ij} is the \rightarrow *detour distance* between vertices v_i and v_j .

Example D2

Detour-path matrix and detour-delta matrix for the H-depleted molecular graph of ethylbenzene; ww is the hyperdetour index and $Wi(\Delta_{\Delta})$ is the \rightarrow Wiener-type index derived from the detour-delta matrix.



detour-path matrix									detour-delta matrix								
Atom	1	2	3	4	5	6	7	8	Atom	1	2	3	4	5	6	7	8
1	0	1	3	28	21	15	21	28	1	0	0	1	21	15	10	15	21
2	1	0	1	21	15	10	15	21	2	0	0	0	15	10	6	10	15
3	3	1	0	15	10	6	10	15	3	1	0	0	10	6	3	6	10
4	28	21	15	0	15	10	6	10	4	21	15	10	0	10	6	3	6
5	21	15	10	15	0	15	10	6	5	15	10	6	10	0	10	6	3
6	15	10	6	10	15	0	15	10	6	10	6	3	6	10	0	10	6
7	21	15	10	6	10	15	0	15	7	15	10	6	3	6	10	0	10
8	28	21	15	10	6	10	15	0	8	21	15	10	6	3	6	10	0

$$ww = 2 \times 1 + 3 + 3 \times 6 + 7 \times 10 + 9 \times 15 + 4 \times 21 + 2 \times 28 = 368$$

$$Wi(\Delta_{\Delta}) = 1 + 3 \times 3 + 7 \times 6 + 9 \times 10 + 4 \times 15 + 2 \times 21 = 244$$

From the detour matrix and the distance matrix, a combined matrix, called **detour–distance combined matrix** $\Delta \wedge D$ (or **maximum–minimum path matrix** or **detour distance–topological distance combined matrix**), is defined as [Ivanciuc and Balaban, 1994b]

$$[\Delta \wedge D]_{ij} = \begin{cases} \Delta_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij} & \text{if } i > j \end{cases}$$

This is a square unsymmetrical $A \times A$ matrix, where the upper triangle of the matrix contains the elements of the detour matrix (information about the longest paths) and the lower triangle contains the elements of the topological \rightarrow distance matrix (information about the shortest paths).

The **maximum–minimum path sum** of the i th vertex, denoted by $MmPVS_i$, is a local vertex invariant defined as the sum of the lengths of the longest and shortest paths between vertex v_i and any other vertex in the molecular graph. It is calculated as the sum of the elements in the $\Delta \wedge D$ matrix row and column corresponding to the i th vertex or, alternatively, as the sum of the \rightarrow vertex distance degree σ_i calculated from the distance matrix D and the maximum path

sum $MPVS_i$ of the i th vertex calculated from the detour matrix Δ :

$$MmPVS_i = VS_i(\Delta \wedge D) + CS_{j=i}(\Delta \wedge D) = \sum_{j=1}^A [\Delta \wedge D]_{ij} + \sum_{i=1}^A [\Delta \wedge D]_{ij} = MPVS_i + \sigma_i$$

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

A combined molecular index, called **detour–Wiener combined index** (or **MmPS topological index**) and denoted as $w \wedge W$, is defined as the sum of the lengths of the longest and shortest paths between any two vertices in the molecular graph and is calculated from the detour–distance combined matrix as

$$w \wedge W = \sum_{i=1}^A \sum_{j=1}^A [\Delta \wedge D]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A MmPVS_i = w + W$$

where w is the detour index, that is, the sum of the lengths of the longest paths in the graph, and W is the \rightarrow *Wiener index*, that is, the sum of the lengths of the shortest paths.

It must be noted that for acyclic graphs the following relation holds:

$$w = W = (w \wedge W)/2$$

The transpose of the detour–distance combined matrix is the **distance–detour combined matrix** $D \wedge \Delta$ (or **minimum–maximum path matrix**, or **topological distance–detour distance combined matrix**), defined as [Janežič, Miličević *et al.*, 2007]

$$[D \wedge \Delta]_{ij} = \begin{cases} d_{ij} & \text{if } i < j \\ 0 & \text{if } i = j \\ \Delta_{ij} & \text{if } i > j \end{cases}$$

Note that several molecular descriptors derived from this matrix, such as the \rightarrow *spectral indices* and \rightarrow *Wiener-type indices*, are the same as those from the detour–distance combined matrix, because eigenvalues of detour–distance matrix and distance–detour matrix and total sum of their matrix elements coincide.

The **distance/detour quotient matrix** (or **topological distance/detour distance quotient matrix**), denoted as D/Δ , is also derived from detour and distance matrices but it is a square symmetric matrix $A \times A$ whose off-diagonal entries are the ratio of the lengths of the shortest over the longest path between any pair of vertices [Randić, 1997c]. It is defined as

$$[D/\Delta]_{ij} = \begin{cases} \frac{d_{ij}}{\Delta_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} and Δ_{ij} are the topological and detour distances between vertices v_i and v_j , respectively. Some local and graph invariants can be calculated from this matrix. The row sums were proposed as local invariants showing a high discriminatory ability; branching vertices tend to have smaller row sums than bridging vertices. If the D/Δ matrix row sums of vertices belonging to single rings (or cycles) in the molecule are summed up, the **D/Δ ring indices**, which can be considered special substructure descriptors reflecting local geometrical environments in complex cyclic systems, are obtained. Moreover, the half sum of all row sums, which corresponds to the half sum of all entries of the D/Δ matrix, was proposed as an index

of \rightarrow *molecular cyclicity*, showing regular variation with increase in cyclicity in graphs of the same size. It is called the **D/ Δ index** (or **Wiener sum index**) and is defined as [Randić, 1997c]:

$$D/\Delta \equiv Wi(D/\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [D/\Delta]_{ij}$$

where Wi is the \rightarrow *Wiener operator*.

The D/Δ index decreases as the cyclicity of the molecule increases, so that it reaches the maximum value for the monocyclic graph C_A and the minimum for the \rightarrow *complete graph* K_A , A being the number of vertices of the actual graph G_A . Therefore, a more suitable measure of molecular cyclicity was proposed as a standardized D/Δ index, called **cyclicity index** and denoted by γ [Randić, 1997c]

$$\gamma = \frac{D/\Delta(C_A) - D/\Delta(G_A)}{D/\Delta(C_A) - D/\Delta(K_A)}$$

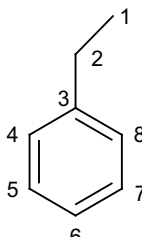
The **average cyclicity index** is calculated simply as

$$\bar{\gamma} = \frac{\gamma}{A}$$

where A is the number of graph vertices. Both the cyclicity and the average cyclicity index allow comparison of cyclic systems of different sizes; they represent the deviation of the cyclicity of the actual molecule from that of the size-corresponding monocyclic molecule C_A . Moreover, the leading eigenvalue of the distance/detour quotient matrix was proposed as another descriptor to account for cyclicity [Randić, 1997c; Pisanski, Plavšić *et al.*, 2000].

Example D3

Detour–distance combined matrix and distance/detour quotient matrix for the H-depleted molecular graph of ethylbenzene. $MmPVS_i$ is the maximum–minimum path sum of the i th vertex; VS_i and CS_j indicate the matrix row and column sums, respectively. w and W are the detour and Wiener index, respectively; $w \wedge W$ and D/Δ are the detour–Wiener combined index and the D/Δ index, respectively.

	Atom	1	2	3	4	5	6	7	8	VS_i
	1	0	1	2	7	6	5	6	7	34
	2	1	0	1	6	5	4	5	6	28
	3	2	1	0	5	4	3	4	5	24
	4	3	2	1	0	5	4	3	4	22
	5	4	3	2	1	0	5	4	3	22
	6	5	4	3	2	1	0	5	4	24
	7	4	3	2	3	2	1	0	5	20
	8	3	2	1	2	3	2	1	0	14
	CS_j	22	16	12	26	26	24	28	34	188

Atom	$MmPVS_i$		Atom	1	2	3	4	5	6	7	8	VS_i
1	56	$\mathbf{D}/\Delta =$	1	0	1	1	0.43	0.67	1	0.67	0.43	5.20
2	44		2	1	0	1	0.33	0.60	1	0.60	0.33	4.86
3	36		3	1	1	0	0.20	0.50	1	0.50	0.20	4.40
4	48		4	0.43	0.33	0.20	0	0.20	0.50	1	0.50	3.16
5	48		5	0.67	0.60	0.50	0.20	0	0.20	0.50	1	3.67
6	48		6	1	1	1	0.50	0.20	0	0.20	0.50	4.40
7	48		7	0.67	0.60	0.50	1	0.50	0.20	0	0.20	3.67
8	48		8	0.43	0.33	0.20	0.50	1	0.50	0.20	0	3.16
	376											

$w + W = 124 + 64 = 188$
 $w \wedge W = \frac{1}{2} \times (56 + 44 + 36 + 48 + 48 + 48 + 48 + 48) = 376/2 = 188$
 $D/\Delta = \frac{1}{2} \times (5.20 + 4.86 + 4.40 + 3.16 + 3.67 + 4.40 + 3.67 + 3.16) = 32.52/2 = 16.26$

A variant of the distance/detour quotient matrix is the **detour/distance quotient matrix** (or **detour distance/topological distance quotient matrix**), denoted by Δ/\mathbf{D} , whose off-diagonal elements are the reciprocal of the distance/detour quotient matrix elements [Plavšić, Trinajstić *et al.*, 1998]:

$$[\Delta/\mathbf{D}]_{ij} = \begin{cases} \frac{\Delta_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

A Wiener type index derived from this matrix is the Δ/\mathbf{D} index defined as

$$\Delta/D \equiv Wi(\Delta/\mathbf{D}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta/\mathbf{D}]_{ij}$$

where Wi is the \rightarrow Wiener operator.

Note that detour/distance and distance/detour quotient matrices do not have much sense for acyclic structures, all the elements being equal to 1, since detour and topological distances are the same.

For detour matrix, distance–detour combined matrix and detour–distance combined matrix there can also be defined the corresponding reciprocal matrices, which are the **reciprocal detour matrix** Δ^{-1} , **reciprocal distance–detour combined matrix** $\mathbf{D} \wedge \Delta^{-1}$, and the **reciprocal detour–distance combined matrix** $\Delta \wedge \mathbf{D}^{-1}$, as

$$[\Delta^{-1}]_{ij} = \begin{cases} \Delta_{ij}^{-1} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \quad [\mathbf{D} \wedge \Delta^{-1}]_{ij} = \begin{cases} d_{ij}^{-1} & \text{if } i < j \\ 0 & \text{if } i = j \\ \Delta_{ij}^{-1} & \text{if } i > j \end{cases} \quad [\Delta \wedge \mathbf{D}^{-1}]_{ij} = \begin{cases} \Delta_{ij}^{-1} & \text{if } i < j \\ 0 & \text{if } i = j \\ d_{ij}^{-1} & \text{if } i > j \end{cases}$$

All elements equal to zero are left unchanged in the reciprocal matrices. Moreover, \rightarrow *Harary detour indices* are derived from the reciprocal detour matrix and \rightarrow *Harary detour-distance indices* from the reciprocal detour-distance or distance-detour combined matrix.

The **detour complement matrix** ΔC for simple graphs is defined as [Janežič, Miličević *et al.*, 2007]

$$[\Delta C]_{ij} = \begin{cases} A - [\Delta]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of atoms. The half sum of the detour complement matrix is the **detour complement index**:

$$Wi(\Delta C) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\Delta C]_{ij}$$

where Wi is the \rightarrow *Wiener operator*.

By analogy with the \rightarrow *reverse Wiener matrix*, the **reverse detour matrix**, denoted as $R\Delta$, was defined as [Janežič, Miličević *et al.*, 2007]

$$[R\Delta]_{ij} = \begin{cases} \max \Delta - [\Delta]_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where $\max \Delta$ is the length of the longest path in the graph. The half sum of the elements of this matrix is the **reverse detour index**:

$$Wi(R\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [R\Delta]_{ij}$$

where Wi is the \rightarrow *Wiener operator*.

📖 [Harary, 1969a; Diudea, Pârvt *et al.*, 1997a; Linert and Lukovits, 1997; Trinajstić, Nikolić *et al.*, 1997; Randić, DeAlba *et al.*, 1998]

- **detour-path matrix** \rightarrow detour matrix
- **detour polynomial** \rightarrow characteristic polynomial-based descriptors
- **detour-Wiener combined index** \rightarrow detour matrix
- **Dewar-Grisdale approach** \rightarrow electronic substituent constants (\odot field/resonance effect separation)
- **Dewar-Golden-Harris approach** \rightarrow electronic substituent constants (\odot field/resonance effect separation)
- **Dewar index** \rightarrow delocalization degree indices
- **Dewar resonance energy** \rightarrow delocalization degree indices
- **DFT** \equiv *Density Functional Theory* \rightarrow quantum-chemical descriptors
- **DFT-based descriptors** \rightarrow quantum-chemical descriptors
- **diagonal matrix** \rightarrow algebraic operators
- **diagonal operator** \rightarrow algebraic operators (\odot diagonal matrix)
- **diamagnetic susceptibility exaltation** \rightarrow delocalization degree indices
- **Dice similarity coefficient** \rightarrow similarity/diversity (\odot Table S9)

- **dielectric constant** → physico-chemical parameters
- **dielectric susceptibility** → physico-chemical parameters
- **difference in atomic charge-weighted surface area** → charged partial surface area descriptors
- **difference in charged partial surface area** → charged partial surface area descriptors
- **difference indices** → combined descriptors
- **difference in total charge-weighted surface area** → charged partial surface area descriptors
- **difference matrices** → matrices of molecules
- **differential connectivity indices** → combined descriptors
- **differential descriptors** → combined descriptors
- **differential Shannon's entropy** → information content
- **diffusivity** → grid-based QSAR techniques (⊙ VolSurf descriptors)
- **digraph** → graph
- **DiP descriptors** ≡ *Distance Profile descriptors* → substructure descriptors
- **dipolarity/polarizability term** → Linear Solvation Energy Relationships
- **dipole moment** → electric polarization descriptors
- **dipole moment components** → electric polarization descriptors
- **dipole polarization** → electric polarization descriptors
- **dipole term** ≡ *dipolarity/polarizability term* → Linear Solvation Energy Relationships
- **directed graph** ≡ *digraph* → graph
- **directional WHIM density** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM descriptors** → WHIM descriptors
- **directional WHIM shape** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM size** → WHIM descriptors (⊙ directional WHIM descriptors)
- **directional WHIM symmetry** → WHIM descriptors (⊙ directional WHIM descriptors)
- **disconnected graph** → graph (⊙ connected graph)
- **discrete wavelet transforms** → spectra descriptors
- **disjoint principal properties** → Principal Component Analysis
- **dispersion** → distance matrix
- **dissection of a graph** → graph
- **dissociation constant** → physico-chemical properties (⊙ equilibrium constants)
- **distance-adjacency map matrix** → biodescriptors (⊙ proteomics maps)
- **distance code centric index** → centric indices
- **distance complement/distance quotient matrix** → distance matrix
- **distance complement matrix** → distance matrix
- **distance connectivity index** ≡ *Balaban distance connectivity index*

■ distance-counting descriptors

Proposed by Clerc and Terkovich [Clerc and Terkovich, 1990], also called **start-end vectors** (or **SE-vectors**), are → *vectorial descriptors* collecting → *path counts* of different lengths relative to pairs of atom types in the → *H-depleted molecular graph*.

These descriptors are conceptually the same as the → *topological atom pairs*, the difference is that SE-vectors are based on simple atom types and between any two atom types all the paths are calculated instead of the shortest one.

To generate SE-vectors, first, all nonhydrogen atoms are assigned one or more atom types, which are defined by the chemical element of the atoms to account for heteroatoms. Moreover,

three additional atom types are considered: the generic atom type (T), which is assigned to any atom, the sp^2 -hybridized atom type ("2"), and the sp -hybridized atom type ("3"). Each atom is assigned at least one and at most three types.

Then, for each i th atom, the atomic path counts mP_i of length m ($m = 0, \dots, L$) are calculated; the path counts of the same m th order are summed up over all atoms to give the corresponding m th order molecular path count mP , divided by 2 for lengths $m > 0$. L is the maximum path length considered and is usually set at a reasonable number depending on the \rightarrow data set (typically set at 5).

Different vectors are obtained depending on the considered combination of atom types. For instance, the TT-vector encodes information on the paths between any two atoms in the graph, independent of their chemical type and hybridization state. This vector descriptor encodes information about branching, size, and the cyclicity of molecules. Moreover, the NT-vector consists of the number of paths of different lengths, starting from the nitrogen atoms and ending at all of the remaining atoms. Analogously, the NN-vector consists of the number of paths of different lengths, starting from a nitrogen atom and ending at the other nitrogen atoms; N2-vector and O2-vector represent the mutual position of double bonds and nitrogen or oxygen atoms in the molecule, respectively.

The first element of each SE-vector, that is, the zero-order molecular path count, corresponds to the number of occurrences of the considered graph elements: for example, the first entry in the TT-vector is the number of heavy atoms in the molecule, in the OO-vector it is the number of oxygen atoms, and in the 2T-vector the number of sp^2 -hybridized atoms. To calculate path counts of higher order, atomic path counts are summed up; however, if the atom types are the same, each path of nonzero length is counted twice and therefore the sum has to be divided by two.

The final distance-counting descriptor is obtained by chaining in an arbitrary but fixed way all of the calculated SE-vectors, such as

$$\{TT_0, TT_1, \dots, TT_L; NT_0, NT_1, \dots, NT_L; NN_0, NN_1, \dots, NN_L; \dots; 22_0, 22_1, \dots, 22_L; \dots\}$$

where each bin represent the occurrence number of paths of a given length for each combination of two atom types.

The number of bins in the final vector is

$$\text{number of bins} = \frac{n(n+1)}{2} \cdot (L+1)$$

where n is the number of different atom types and L the maximum path length. It follows that the dimension of SE-vectors increases linearly with the maximum path length and with the square of the number of atom types.

SE-vectors describe the global topology of a molecule, also taking into account the presence of heteroatoms and multiple bonds as well as their numbers and relative position. However, they do not encode information about stereochemistry of molecules.

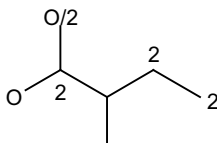
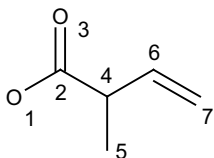
To obtain SE-vectors not depending on molecular size, a normalization scheme is introduced. The TT-vector is normalized with reference to the chain graph, by first subtracting the ${}^{CG}TT$ -vector of the chain graph and then dividing by it:

$$\left\{ \frac{TT_0 - {}^{CG}TT_0}{{}^{CG}TT_0}, \frac{TT_1 - {}^{CG}TT_1}{{}^{CG}TT_1}, \dots, \frac{TT_L - {}^{CG}TT_L}{{}^{CG}TT_L} \right\}$$

The SE-vectors other than TT-vector are normalized by dividing each entry by the total number of the corresponding atom types in the molecule, that is, the value of the first entry in the vector or the total sum of the vector entries.

Example D4

Some SE-vectors for the molecule shown below.



Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	1	2	2	1	0
2	1	3	2	1	0	0
3	1	1	2	2	1	0
4	1	3	3	0	0	0
5	1	1	2	3	0	0
6	1	2	2	2	0	0
7	1	1	1	2	2	0
sum	7	12	14	12	4	0
SE(TT)	7	6	7	6	2	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	1	2	2	1	0
3	1	1	2	2	1	0
sum	2	2	4	4	2	0
SE(OT)	2	2	4	4	2	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	1	0	1	0	0	0
3	1	0	1	0	0	0
sum	2	0	2	0	0	0
SE(OO)	2	0	1	0	0	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
2	1	3	2	1	0	0
3	1	1	2	2	1	0
6	1	2	2	2	0	0
7	1	1	1	2	2	0
sum	4	7	7	7	3	0
SE(2T)	4	7	7	7	3	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
2	1	1	1	1	0	0
3	1	1	0	1	1	0
6	1	1	1	1	0	0
7	1	1	0	1	1	0
sum	4	4	2	4	2	0
SE(22)	4	2	1	2	1	0

Atom	⁰ P	¹ P	² P	³ P	⁴ P	⁵ P
1	0	1	1	1	1	0
3	1	1	0	1	1	0
sum	1	2	1	2	2	0
SE(O2)	1	2	1	2	2	0

$$\{\text{SE(TT)}; \text{SE(2T)}; \text{SE(22)}; \dots\} \equiv \{7, 6, 7, 6, 2, 0; 4, 7, 7, 7, 3, 0; 4, 2, 1, 2, 1, 0; \dots\}$$

Modified SE-vectors were proposed simply using the shortest path between any pair of atom types in place of all the existing paths. These descriptors were called **SESP-Top vectors**, where “SP” stands for *shortest path* [Baumann, 2002a]. Moreover, to take into account not only the topological and hybridization aspects of the molecule but also the stereochemical and con-

formational information, a further development of SE-vectors was proposed exploiting the 3D spatial coordinates (x, y, z) of a molecule. For each pair of atom types, the shortest paths up to a maximum length are counted but rather than incrementing by one the corresponding bin value, each path gives a contribution equal to the ratio of the geometric distance to the topological distance. Of course, the algorithm is not applied to the first entry of each vector, the distance being equal to zero. Calculations are performed by using the \rightarrow *geometric distance/topological distance quotient matrix* **G/D** instead of the \rightarrow *distance matrix*. These 3D vectorial descriptors are called **SESP-Geo vectors** [Baumann, 2002a]; a different implementation is \rightarrow *Distance Profile descriptors*, calculated by binning the geometric distances between pairs of atom types.

📖 [Baumann and Clerc, 1997; Baumann, Affolter *et al.*, 1997; Affolter, Baumann *et al.*, 1997; Baumann, 1999]

- **distance degree** \rightarrow distance matrix
- **distance degree centric index** \rightarrow centric indices

■ distance-degree matrices

Distance-degree matrices are a class of graph-theoretical matrices, which can be either vertex matrices, whose entries refer to atoms, or edge matrices, whose entries refer to bonds. In both cases, the matrix entries are defined by weighting topological distances between two graph elements (i.e., vertices or edges) by their connectivities [Ivanciuc, 1989, 1999c, 2000c].

Distance–valency matrices, denoted by **Dval**, are square ($A \times A$) matrices, A being the number of atoms, defined for vertex- and edge-weighted graphs as [Ivanciuc, 1999c, 2000c]

$$[\mathbf{Dval}(\alpha, \beta, \gamma; w)]_{ij} = \begin{cases} d_{ij}^{\alpha}(w) \cdot val_i^{\beta}(w) \cdot val_j^{\gamma}(w) & \text{if } i \neq j \\ w_i \cdot val_i^{(\beta+\gamma)}(w) & \text{if } i = j \end{cases}$$

where w is the \rightarrow *weighting scheme* used to calculate vertex w_i and edge parameters w_{ij} , and val is the \rightarrow *valency of the vertex* defined as the sum of the weights w_{ij} of the edges incident to the vertex v_i . $d_{ij}(w)$ is the weighted distance between vertices v_i and v_j (see \rightarrow *weighted distance matrices*); α , β , and γ are exponential parameters; unsymmetrical matrices are obtained for $\beta \neq \gamma$.

For simple molecular graphs, where vertex and edge parameters are equal to 1, vertex valencies coincide with the \rightarrow *vertex degrees* δ and the distance between pairs of vertices is the \rightarrow *topological distance* d_{ij} ; in this case, distance–valency matrices are properly called **vertex–distance–vertex–degree matrices**, denoted by **D δ** [Janežič, Miličević *et al.*, 2007], and defined as

$$[\mathbf{D\delta}(\alpha, \beta, \gamma)]_{ij} = \begin{cases} d_{ij}^{\alpha} \cdot \delta_i^{\beta} \cdot \delta_j^{\gamma} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

Some distance–degree matrices, representing simple molecular graphs and derived from selected combinations of α , β , and γ parameters, result into other well-known graph-theoretical matrices defined in the literature. Examples are the \rightarrow *distance matrix* ($\alpha = 1, \beta = 0, \gamma = 0$), the \rightarrow *Harary matrix* ($\alpha = -1, \beta = 0, \gamma = 0$), and \rightarrow *XI matrix* ($\alpha = 0, \beta = -1/2, \gamma = -1/2$).

The row sums of vertex–distance–vertex–degree matrices for different combinations of the three parameters are \rightarrow *local vertex invariants*, proposed by Ivanciuc with the name \rightarrow *VTI indices* [Ivanciuc, 1989] and extensively studied by Perdih [Perdih and Perdih, 2002a, 2002b, 2002c, 2002d, 2002e, 2003b, 2003e, 2003d].

Note. The vertex–distance–vertex–degree matrices with $\beta = 0$ were called **$v^m d^n$ matrices** by Perdih [Perdih and Perdih, 2002a] and the **general distance-degree matrix** was denoted by $G(a, b, c)$ [Perdih and Perdih, 2004], whose elements are $v_i^a v_j^b d_{ij}^c$ [Perdih and Perdih, 2003c], where v denotes the vertex degree δ and a , b , and c are the parameters corresponding to β , γ , and α , respectively. The diagonal elements of these matrices are equal to zero. Using the notations adopted in this book, the general distance-degree matrix elements are defined as the following:

$$[G(a, b, c)]_{ij} = \delta_i^a \cdot \delta_j^b \cdot d_{ij}^c$$

From $v^m d^n$ matrices, several \rightarrow *branching indices* were proposed [Perdih, 2003].

\rightarrow *Wiener-type indices* [Ivanciuc, 2000i] were calculated from symmetric distance–valency matrices, while \rightarrow *matrix sum indices* were calculated from unsymmetrical distance–valency matrices [Ivanciuc, 1999c]. Moreover, \rightarrow *characteristic polynomial-based descriptors*, \rightarrow *Hosoya-type indices*, \rightarrow *spectral indices*, \rightarrow *hyper-Wiener-type indices*, and \rightarrow *spectral moments* were derived from distance–valency matrices and tested in QSAR/QSPR modeling [Ivanciuc, 1999c, 2000c].

Edge-distance–edge-degree matrices, denoted by ${}^E D\epsilon$, are square ($B \times B$) matrices, B being the number of bonds, defined as [Janežič, Miličević *et al.*, 2007]

$$[{}^E D\epsilon(\alpha, \beta, \gamma)]_{ij} = \begin{cases} {}^e d_{ij}^\alpha \cdot \epsilon_i^\beta \cdot \epsilon_j^\gamma & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where ${}^e d_{ij}$ is the topological distance between any pair of bonds and ϵ_i and ϵ_j are their edge degrees, and α , β , and γ are exponential parameters. Similar to vertex–distance–vertex–degree matrices, unsymmetrical edge matrices are obtained for $\beta \neq \gamma$.

It must be noted that the edge-distance–edge-degree matrix of a molecular graph is the vertex-distance–vertex-degree matrix of the corresponding \rightarrow *line graph*.

- **distance degree sequence** \equiv *vertex distance code* \rightarrow distance matrix
- **distance-delta matrix** \equiv *delta matrix* \rightarrow distance-path matrix
- **distance-detour combined matrix** \rightarrow detour matrix
- **distance/detour quotient matrix** \rightarrow detour matrix
- **distance–distance combined matrices** \rightarrow molecular geometry
- **distance/distance complement quotient matrix** \rightarrow distance matrix
- **distance/distance matrices** \rightarrow molecular geometry
- **distance distribution moments** \rightarrow distance matrix
- **distance-enhanced exponential sum connectivities** \rightarrow exponential sum connectivities
- **distance exponent index** \rightarrow biodescriptors (\odot peptide sequences)
- **distance-extended matrices** \equiv *expanded distance matrices*

■ Distance Geometry (DG)

A QSAR method proposed with the aim of automatically finding the simplest receptor binding site consistent with the binding data is based on the following assumptions: (1) binding is observed to occur on a single receptor site; (2) each ligand molecule has a well-determined chemical 3D structure and its flexibility is also taken into account; (3) no chemical modification of the molecules occurs during the binding, although their conformations may change; (4) the free energy of such a conformational change is small compared to the free energy of the binding; (5) the experimental free energy of binding is modeled by adding the interaction energies for all contact distances between parts of the ligand molecule and receptor site; and (6) the receptor site

is considered relatively rigid with respect to ligand conformational flexibility [Blumenthal, 1970; 1977, 1978, 1979, 1980, 1981, 1991; Ghose and Crippen, 1990].

In the framework of the DG method, each ligand molecule is represented as a collection of points in space, each corresponding to an atom or group of atoms, and the conformation of the molecule is described in terms of Euclidean distances between points. The matrix containing Euclidean distances between all possible pairs of points is the \rightarrow *geometry matrix* of the molecule when each point corresponds to a single atom. To account for molecular flexibility, a matrix of lower bounds on the interpoint distances and a matrix of upper bounds are also defined; fixed spatial distances are represented by equal values in these matrices.

The binding site of the receptor is represented in an analogous way, that is, representing the interesting binding site regions by points and collecting their relative positions in a receptor geometry distance matrix. However, unlike molecule points, the site points may be called either “empty” or “filled”. An empty site point is a vacant place where a ligand point might be lying when binding takes place, whereas a filled site point indicates the position of receptor steric blocking groups, precluding the presence of any ligand point during binding.

The free energy of binding is calculated in a simplified all-or-nothing fashion by adding up contributions from each pair of ligand point and site point “contact.” The individual interaction energy contributions, taken from a reference list, are collected into an energy matrix where each row corresponds to a type of ligand point and each column to a type of receptor site point. If the fit between the calculated and the experimental binding free energy is not satisfactory, the interaction energy contributions or the number and/or geometry of the site points can be changed.

An extension of the distance geometry approach is given by the **Voronoi binding site models**, proposed with the aim of reducing excessive details in site model shape [Crippen, 1987; Srivastava, Richardson *et al.*, 1993]. In this approach, the receptor site is not represented by points but by nonoverlapping regions, called \rightarrow *Voronoi polyhedra*, that cover the whole space. Each atom would always lie in one and only one region, that is, in a Voronoi polyhedron, and a binding mode would consist of a listing of the regions in which each atom is located [Boulu and Crippen, 1989; Boulu, Crippen *et al.*, 1990].

📖 [Gordon, 1980; Ghose and Crippen, 1982, 1983, 1984, 1985b, 1985a; Sheridan, Nilakantan *et al.*, 1986; Ghose, Logan *et al.*, 1995; Grdadolnik and Mierke, 1997; Wildman and Crippen, 2002; Imre, Veress *et al.*, 2003; Raymond and Willett, 2003; Ursu and Diudea, 2005]

➤ **distance index** \equiv *distance degree* \rightarrow distance matrix

■ distance matrix (D)

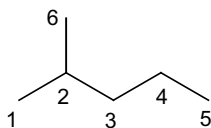
Derived from the \rightarrow *H-depleted molecular graph* G , the distance matrix (viz., **vertex distance matrix** or **minimum path matrix**) summarizes in matrix form the topological distance information between all the pairs of nonhydrogen atoms in a molecule [Harary, 1964; Hakimi and Yau, 1965; Harary, 1969a; Patrinos and Hakimi, 1973; Gutman and Polansky, 1986b; Rouvray, 1986a; Buckley and Harary, 1990; Mihalić, Veljan *et al.*, 1992; Trinajstić, 1992; Hage and Harary, 1995]. The **topological distance** d_{ij} is the number of edges along the shortest \rightarrow *path* $^{\min}p_{ij}$ between the vertices v_i and v_j , that is, the length of the \rightarrow *geodesic* between v_i and v_j :

$$[\mathbf{D}]_{ij} \equiv d_{ij} = \begin{cases} |^{\min}p_{ij}| & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

The off-diagonal entries of the distance matrix are equal to 1 if vertices v_i and v_j are adjacent (i.e., the atoms i and j are bonded and $d_{ij} = a_{ij} = 1$ where a_{ij} are elements of the \rightarrow adjacency matrix A) and are greater than 1 otherwise. The diagonal elements are of course equal to zero. The distance matrix is symmetric with dimension $A \times A$, where A is the number of atoms.

Example D5

Distance matrix, vertex distance degrees σ_i , and atom eccentricities η_i of 2-methylpentane.



$D =$

Atom	1	2	3	4	5	6	σ_i	η_i
1	0	1	2	3	4	2	12	4
2	1	0	1	2	3	1	8	3
3	2	1	0	1	2	2	8	2
4	3	2	1	0	1	3	10	3
5	4	3	2	1	0	4	14	4
6	2	1	2	3	4	0	12	4

For vertex- and edge-weighted graphs, the distance matrix entry $i-j$ could be defined as the minimum sum of edge weights along the path between the vertices v_i and v_j , which is not necessarily the shortest possible path between them or otherwise as the sum of the weights of the edges along the shortest path between the considered vertices. Diagonal entries are the vertex weights. Different \rightarrow weighting schemes were proposed from which a number of \rightarrow weighted distance matrices were derived.

The **distance degree** (or **vertex distance degree**, **distance number**, **distance index**, **distance rank**, **distance sum**, **vertex distance sum**, **distance of a vertex**) is a local vertex invariant, denoted as σ_i , and defined as the distance matrix row sum:

$$\sigma_i \equiv VS_i(D) = \sum_{j=1}^A d_{ij}$$

where VS stands for vertex sum and is the \rightarrow row sum operator, that is, the sum of the matrix elements in a row.

The distance degrees of the vertices of 2-methylpentane are shown in Example D5. For instance, the distance degree of vertex 2 in 2-methylpentane is $\sigma_2 = d_{21} + d_{23} + d_{24} + d_{25} + d_{26} = 1 + 1 + 2 + 3 + 1 = 8$.

Vertex distance degrees are \rightarrow local vertex invariants: high values are observed for \rightarrow terminal vertices (e.g., in 2-methylpentane, $\sigma = 12$ for terminal vertices 1 and 6, and $\sigma = 14$ for terminal vertex 5), while low values for \rightarrow central vertices. Moreover, among the terminal vertices, vertex distance degrees are small if the vertex is next to a branching site (e.g., in 2-methylpentane, vertices 1 and 6 are directly bonded to vertex 2 that represents a branching site) and larger if the terminal vertex is far away (e.g., in 2-methylpentane, terminal vertex 5 is three bonds far away from the branching site 2).

The half-sum of all the elements d_{ij} of the distance matrix, which is equal to the half sum of the distance degrees σ_i of all the vertices [Harary, 1959], is the well-known \rightarrow Wiener index W , which is one of the most popular topological indices used in QSAR modeling [Wiener, 1947c].

The total sum of the entries of the distance matrix is another topological index called **Rouvray index** (or **rank distance** or **total vertex distance**) and is denoted as I_{ROUV} , which is twice the Wiener index W :

$$I_{\text{ROUV}} = \sum_{i=1}^A \sum_{j=1}^A d_{ij} = \sum_{i=1}^A \sigma_i = 2W$$

For example, in 2-methylpentane, the Rouvray index, derived from distance values, is $I_{\text{ROUV}} = 10 \times 1 + 10 \times 2 + 6 \times 3 + 4 \times 4 = 64$ or, alternatively, derived from distance degrees, is $I_{\text{ROUV}} = 12 + 8 + 8 + 10 + 14 + 12 = 64$.

The average row sum of the distance matrix is a molecular invariant called **average graph distance degree**, which coincides with the average Rouvray index, defined as [Skorobogatov and Dobrynin, 1988]:

$$\bar{\sigma} = \frac{1}{A} \sum_{i=1}^A \sigma_i = \frac{2 \cdot W}{A} = \frac{I_{\text{ROUV}}}{A}$$

For example, in 2-methylpentane, the average distance degree is $\bar{\sigma} = 64/6 = 10.667$.

The **compactness**, denoted as $Comp$, was defined as another function of the Wiener index W [Doyle and Garver, 1977]:

$$\frac{1}{Comp} = \frac{4 \cdot W}{A(A-1)} = 2 \cdot \bar{W}$$

Note that the compactness is the reciprocal of twice the \rightarrow mean Wiener index \bar{W} . The smaller the Wiener index the larger the compactness of the molecule. For example, in 2-methylpentane,

$$Comp = \frac{6 \cdot (6-1)}{4 \cdot 32} = 0.234$$

The **mean distance degree deviation**, denoted as $\Delta\sigma$, is derived from the average distance degree as [Skorobogatov and Dobrynin, 1988]:

$$\Delta\sigma = \frac{1}{A} \cdot \sum_{i=1}^A |\sigma_i - \bar{\sigma}|$$

For example, in 2-methylpentane

$$\Delta\sigma = \frac{|12-10.667| + |8-10.667| + |8-10.667| + |10-10.667| + |14-10.667| + |12-10.667|}{6} = 2$$

The minimum value of the distance degrees of the molecule atoms is another molecular invariant called **unipolarity** [Skorobogatov and Dobrynin, 1988]:

$$\sigma^* = \min_i(\sigma_i)$$

For example, in 2-methylpentane, $\sigma^* = \min\{12, 8, 8, 10, 14, 12\} = 8$.

Other molecular invariants immediately derived from distance degrees σ_i are **centralization** $\Delta\sigma^*$ [Skorobogatov and Dobrynin, 1988], **variation** $\Delta\sigma^+$, and **dispersion** σ_2^* , defined respectively as [Konstantinova and Skorobogatov, 1995]

$$\Delta\sigma^* = 2 \cdot W - A \cdot \sigma^* \quad \Delta\sigma^+ = \max_i(\sigma_i - \sigma^*) \quad \sigma_2^* = \min_i \left(\frac{1}{A} \cdot \sum_{j=1}^A d_{ij}^2 \right)$$

where W is the Wiener index, A is the number of graph vertices, σ^* is the unipolarity, and d_{ij} denotes the topological distances.

For example, in 2-methylpentane

$$\Delta\sigma^* = 64 - 6 \times 8 = 16 \quad \sigma_2^* = \min(5.667, 2.667, 2.333, 4, 7.667, 5.667) = 2.333$$

$$\Delta\sigma^+ = \max(12-8, 8-8, 8-8, 10-8, 14-8, 12-8) = \max(4, 0, 0, 2, 6, 4) = 6$$

The **PRS index** (or **Product of Row Sums index**) is defined as the product of the vertex distance degrees σ_i [Schultz, Schultz *et al.*, 1992]:

$$\text{PRS} = \prod_{i=1}^A \sigma_i \quad \text{or} \quad \log(\text{PRS}) = \log\left(\prod_{i=1}^A \sigma_i\right) = \sum_{i=1}^A \log(\sigma_i)$$

where the second expression is suggested in QSAR/QSPR modeling due to the large values that can be reached by the PRS index. This index is related to the \rightarrow *permanent* of the distance matrix. For example, in 2-methylpentane

$$\begin{aligned} \log(\text{PRS}) &= 2 \times \log(12) + 2 \times \log(8) + \log(10) + \log(14) = \\ &= 2 \times 1.079 + 2 \times 0.903 + 1 + 1.146 = 6.1107 \end{aligned}$$

Still based on the vertex distance degree σ_i , **R* indices** and **R+ index** were defined as [Randić, Balaban *et al.*, 2001]:

$$R^* = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (\sigma_i \cdot \sigma_j)^\lambda \quad \lambda = \pm \frac{1}{2}; \pm 1 \quad R^+ = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot \frac{(\sigma_i + \sigma_j)}{2}$$

where a_{ij} denotes the elements of the \rightarrow *adjacency matrix*, assuming value equal to 1 for pairs of adjacent vertices.

Table D4 Some molecular descriptors derived from the distance matrix for C8 data set (Appendix C – Set 1).

C8	I_{ROUV}	$\bar{\sigma}$	$\Delta\sigma$	σ^*	$\Delta\sigma^*$	$\Delta\sigma^+$	$\log(\text{PRS})$
n-octane	168	21.000	4.000	16	40	12	24.172
2M	158	19.750	3.750	15	38	12	23.694
3M	152	19.000	3.500	14	40	12	23.369
4M	150	18.750	3.313	13	46	12	23.252
3E	144	18.000	3.500	12	48	12	22.920
22MM	142	17.750	3.063	13	38	12	22.835
23MM	140	17.500	3.125	12	44	12	22.713
24MM	142	17.750	3.250	13	38	10	22.840
25MM	148	18.500	3.500	14	36	8	23.187
33MM	134	16.750	2.813	11	46	12	22.351
34MM	136	17.000	3.000	12	40	10	22.478
2M3E	134	16.750	3.250	11	46	10	22.357
3M3E	128	16.000	3.000	10	48	10	21.980
223MMM	126	15.750	2.563	11	38	10	21.881
224MMM	132	16.500	2.875	12	36	8	22.271
233MMM	124	15.500	2.625	10	44	10	21.748
234MMM	130	16.250	2.938	11	42	8	22.139
2233MMMM	116	14.500	2.250	10	36	6	21.241

I_{ROUV} , $\bar{\sigma}$, $\Delta\sigma$, σ^* , $\Delta\sigma^*$, and $\Delta\sigma^+$ are Rouvray index, average distance degree, mean distance degree deviation, unipolarity, centralization, and variation, respectively.

The maximum value entry in the i th row of the distance matrix **D** is called **atom eccentricity** (or **vertex eccentricity**) and denoted as η_i :

$$\eta_i = \max_j(d_{ij})$$

The atom eccentricity is a local vertex invariant representing the maximum distance from a vertex to any other vertex in the graph.

For example, the eccentricity of vertex 2 in 2-methylpentane is 3, which is the topological distance between vertices 2 and 5. Eccentricities of the nonhydrogen atoms of 2-methylpentane are listed in Example D5.

From the vertex eccentricity definition, a graph can be immediately characterized by two molecular descriptors known as **topological radius** R and **topological diameter** D . The topological radius of a molecule is defined as the minimum vertex eccentricity and the topological diameter is defined as the maximum vertex eccentricity, according to the following [Harary, 1969a]:

$$R = \min_i(\eta_i) \quad \text{and} \quad D = \max_i(\eta_i)$$

For example, the radius of 2-methylpentane is 2, while the diameter is 4.

Based on the combined use of topological radius and diameter is the \rightarrow *graph-theoretical shape coefficient*. Moreover, simple molecular descriptors are calculated as some functions of vertex eccentricities [Konstantinova, 1996]. These are the **eccentricity** η , **average atom eccentricity** $\bar{\eta}$, and **eccentric** $\Delta\eta$ defined, respectively, as the following [Skorobogatov and Dobrynin, 1988]:

$$\eta = \sum_{i=1}^A \eta_i \quad \bar{\eta} = \frac{1}{A} \cdot \sum_{i=1}^A \eta_i \quad \Delta\eta = \frac{1}{A} \cdot \sum_{i=1}^A |\eta_i - \bar{\eta}|$$

where A is the number of graph vertices and η_i the eccentricity of the i th vertex.

Example D6

Eccentricity η , average atom eccentricity $\bar{\eta}$, and eccentric $\Delta\eta$ for 2-methylpentane.

$$\eta = 4 + 3 + 2 + 3 + 4 + 4 = 20 \quad \bar{\eta} = \frac{20}{6} = 3.333$$

$$\Delta\eta = \frac{|4-3.333| + |3-3.333| + |2-3.333| + |3-3.333| + |4-3.333| + |4-3.333|}{6} = 0.667$$

Other molecular descriptors based on atom eccentricity values combined with other \rightarrow *local vertex invariants* are the \rightarrow *eccentric connectivity index*, the \rightarrow *eccentric distance sum*, the \rightarrow *connective eccentricity index*, the \rightarrow *eccentric adjacency topochemical indices*, the \rightarrow *superadjacency index*, and the \rightarrow *eccentricity-based Madan indices*.

From the frequencies of the row entries of the distance matrix, the **vertex distance code** (or **distance degree sequence** of a vertex, **DDS_i**) is defined as the ordered sequence of the numbers of occurrence of the different distance values for each i th vertex [Ivanciuc and

Balaban, 1999c]:

$$\mathbf{DDS}_i \equiv \{^1f_i, ^2f_i, ^3f_i, \dots, \eta_i f_i\}$$

where $^1f_i, ^2f_i, ^3f_i, \dots$, called **vertex distance counts**, indicate the frequencies of distances equal to 1, 2, 3, ..., respectively, from vertex v_i to any other vertex and η_i is the i th atom eccentricity. The vertex distance count of first-order 1f_i coincides with the \rightarrow *vertex degree* δ_i , that is, the number of first neighbors, while 2f_i and 3f_i correspond to the \rightarrow *connection number* (i.e., number of second neighbors), and *polarity number* (i.e., number of third neighbors) for the i th vertex, respectively. Moreover, the distance degree can also be expressed in terms of vertex distance counts as

$$\sigma_i = \sum_{k=1}^{\eta_i} {}^k f_i \cdot k$$

where the sum runs over the different distance values k and η_i is the maximum distance from the i th vertex.

The total number of distances in the graph equal to k is called **graph distance count** of k th order ${}^k f$; it is obtained as

$${}^k f = \frac{1}{2} \cdot \sum_{i=1}^A {}^k f_i$$

where A is the number of graph vertices and ${}^k f_i$ is the number of distances equal to k .

The **graph distance code** is the ordered sequence of graph distance counts:

$$\mathbf{GDC} \equiv \{^1f, ^2f, ^3f, \dots, {}^D f\}$$

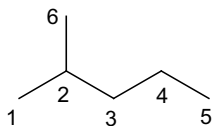
where D is the topological diameter.

The **graph distance index** is defined as the square sum of all graph distance counts [Rouvray, 1983]:

$$\mathbf{GDI} = \sum_{k=1}^D ({}^k f)^2$$

Example D7

Distance degree sequences of the atoms of 2-methylpentane.



Atom	1f_i	2f_i	3f_i	4f_i
1	1	2	1	1
2	3	1	1	—
3	2	3	—	—
4	2	1	2	—
5	1	1	1	2
6	1	2	1	1

For instance, the distance degree sequence of vertex 2 is $\mathbf{DDS}_2 = \{3, 1, 1\}$, which means that there are three vertices one bond away from v_2 , one vertex located at distance two from v_2 , and one vertex at distance three. The graph distance code is $\mathbf{GDC} = \{5, 5, 3, 2\}$ and the graph distance index is $GDI = 25 + 25 + 9 + 4 = 63$.

The **polarity number** p (or **Wiener polarity number**) was defined by Wiener in 1947 [Wiener, 1947c] as the number of pairs of graph vertices, which are separated by three edges. It is usually assumed that the polarity number accounts for the flexibility of acyclic structures, p being equal to the number of bonds around which free rotations can take place. Moreover, it relates to the steric properties of molecules.

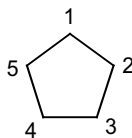
The polarity number is usually calculated from the distance matrix as the number of pairs of vertices at a topological distance equal to 3, that is,

$$p_2 = {}^3f$$

where 3f is the graph distance count of third order, that is, count of entries equal to 3 in the upper or lower triangular submatrix of \mathbf{D} [Platt, 1947, 1952]. In this case, the polarity number is denoted by p_2 to distinguish it from the original p . For acyclic graphs $p = p_2$, while in cycle-containing graphs the two numbers are generally different.

Example D8

H-depleted molecular graph, distance matrix, and polarity numbers p and p_2 for cyclopentane.



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

The polarity number p , calculated from the molecular graph, is equal to 5, which corresponds to the count of the following paths: $p_{14} = \{1, 2, 3, 4\}$, $p_{25} = \{2, 3, 4, 5\}$, $p_{31} = \{3, 4, 5, 1\}$, $p_{42} = \{4, 5, 1, 2\}$, and $p_{53} = \{5, 1, 2, 3\}$. The polarity number p_2 , calculated on the distance matrix is zero, because there are no entries equal to 3 in the distance matrix of cyclopentane.

Two other definitions of polarity number, based on the same original concept, were also introduced to have a greater discriminating ability among cyclic structures suitable for QSAR modeling purposes [Lukovits and Linert, 1998]. The polarity number p_3 was defined as the number of ways a path of length three can be laid upon the hydrogen-depleted graph. Moreover, the polarity number p_4 was defined as the number of ways a path of length 3 can be laid upon the acyclic edges of the graph (including those cases in which the second edge of the path considered coincides with a cyclic edge) $+ 1.8 \times N$, where N is the number of ways the path of length 3 can be laid upon the cyclic part of the graph; all edges not belonging to a cycle are acyclic edges and the product $1.8 \times N$ has to be rounded to yield an integer. For acyclic structures, $p = p_2 = p_3 = p_4$ (Table D5).

Table D5 Values of polarity numbers p , p_2 , p_3 , p_4 , and Wiener index W for some molecules.

Compound	p	p_2	p_3	p_4	W
Cyclopropane	0	0	3	5	3
Cyclobutane	4	0	4	7	8
Methyl-cyclopropane	2	0	3	5	8
<i>n</i> -Pentane	2	2	2	2	28
Cyclopentane	5	0	5	9	15
<i>i</i> -Propyl-cyclopentane	11	6	11	9	62
<i>n</i> -Propyl-cyclopentane	11	5	10	10	67

For example, in cyclopropane, the polarity number p is equal to zero because there are no pairs of vertices that are separated by three edges, while the polarity number p_3 is equal to 3 because there are three paths of length 3: $p_{11} = \{1, 2, 3, 1\}$, $p_{22} = \{2, 3, 1, 2\}$, $p_{33} = \{3, 1, 2, 3\}$. The polarity number p_4 is $1.8 \times 3 = 5.4$, which is rounded to 5, because the contribution of the acyclic part is zero and there are three ways the path of length 3 can be laid upon the cyclic part of the graph (in this case this number coincides with p_3).

The \rightarrow distance polynomial is the characteristic polynomial of the distance matrix used to derive the \rightarrow Hosoya Z' index; moreover, \rightarrow eigenvalues of the distance matrix, \rightarrow spectral moments, \rightarrow determinant, and \rightarrow permanent were proposed as molecular descriptors, even if their use in QSAR modeling is limited since they tend to reach very high values.

Topological distances, distance degrees, eccentricities, distance frequencies, topological radius, and diameter are used to calculate several \rightarrow topological information indices and other molecular descriptors such as \rightarrow Balaban distance connectivity index, \rightarrow hyper-Wiener index, \rightarrow expanded Wiener number, \rightarrow expanded distance indices, \rightarrow Schultz molecular topological index, \rightarrow eccentricity-based Madan indices, \rightarrow Petitjean shape indices, \rightarrow Molecular Distance-Edge vector, \rightarrow Sh indices, \rightarrow Xu index, \rightarrow global flexibility index, several \rightarrow ID numbers, \rightarrow delta number, \rightarrow hyperdistance path index, some among the \rightarrow triplet topological indices, \rightarrow second-grade structural parameters, \rightarrow steric vertex topological index, \rightarrow superpendentic index, and \rightarrow topological charge indices.

Moreover, the most common \rightarrow autocorrelation descriptors are calculated from a molecular graph by using the topological distance as the lag.

Table D6 Some molecular descriptors derived from the distance matrix for C8 data set (Appendix C – Set 1).

C8	R	D	η	$\bar{\eta}$	$\Delta\eta$	p_2	MSD
<i>n</i> -Octane	4	7	44	5.500	1.000	5	0.463
2M	3	6	39	4.875	0.906	5	0.431
3M	3	6	38	4.750	0.813	6	0.410
4M	3	6	37	4.625	0.875	6	0.403
3E	3	5	33	4.125	0.656	7	0.381
22MM	3	5	34	4.250	0.750	5	0.380
23MM	3	5	33	4.125	0.656	7	0.371
24MM	3	5	33	4.125	0.656	6	0.377
25MM	3	5	34	4.250	0.750	5	0.398

(Continued)

Table D6 (Continued)

C8	R	D	η	$\bar{\eta}$	$\Delta\eta$	p_2	MSD
33MM	3	5	32	4.000	0.500	7	0.353
34MM	3	5	32	4.000	0.500	8	0.357
2M3E	2	4	27	3.375	0.625	8	0.349
3M3E	2	4	26	3.250	0.563	9	0.331
223MMM	2	4	27	3.375	0.625	8	0.326
224MMM	2	4	28	3.500	0.625	5	0.346
233MMM	2	4	26	3.250	0.563	9	0.319
234MMM	2	4	27	3.375	0.625	8	0.338
2233MMMM	2	3	22	2.750	0.375	9	0.295

R is the topological radius, D the topological diameter, η the eccentricity, $\bar{\eta}$ the average atom eccentricity, $\Delta\eta$ the eccentric, p_2 the polarity number, and MSD the mean square distance index.

Some graph-theoretical matrices derived from the distance matrix are reported below.

Generalized distance matrices, denoted as \mathbf{D}^λ , are derived from the distance matrix by raising its elements to an exponent λ :

$$[\mathbf{D}^\lambda]_{ij} = \begin{cases} d_{ij}^\lambda & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where λ is any real exponent.

The **distance distribution moments**, denoted as D_λ , are the moments of the distribution of topological distances d_{ij} in a molecular graph, derived from generalized distance matrices defined for positive integer λ values [Klein and Gutman, 1999]:

$$D_\lambda \equiv Wi(\mathbf{D}^\lambda) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij}^\lambda \quad \lambda = 1, 2, 3, \dots$$

where λ is a positive integer and denotes the power of the distance matrix elements and Wi is the \rightarrow *Wiener operator*. Note that distance distribution moments are a subclass of the generalized \rightarrow W_λ indices proposed by Gutman [Gutman, 1997; Gutman, Vidović *et al.*, 1998] and D_1 is the \rightarrow *Wiener index* W .

Normalized distance distribution moments were used to define \rightarrow *molecular profiles* and the second moment D_2 takes part in defining the \rightarrow *hyper-Wiener index*; moreover, the index D_2 was demonstrated to be equal to half the trace of the distance matrix \mathbf{D} raised to the second power [Diudea, 1996a; Diudea and Gutman, 1998].

The **mean square distance index**, denoted as MSD , is calculated from the second-order distance distribution moment as [Balaban, 1983a]:

$$MSD = \left(\frac{\sum_{i=1}^A \sum_{j=1}^A d_{ij}^2}{A \cdot (A-1)} \right)^{1/2} = \left(\frac{\sum_{k=1}^D {}^k f \cdot k^2}{\sum_{k=1}^D {}^k f} \right)^{1/2}$$

where ${}^k f$ is the graph distance count of order k , D is the topological diameter, and A is the number of vertices. The same index restricted to the \rightarrow *terminal vertices* (vertices of degree

one) is called **end point mean square distance index** $D1$ and is applicable only to acyclic graphs.

Both the MSD and the $D1$ indices decrease with increasing \rightarrow *molecular branching* in an isomeric series.

Example D9

The square distance matrix \mathbf{D}^2 , second-order distance distribution moment D_2 , and the mean square distance (MSD) index for 2-methylpentane.

	Atom	1	2	3	4	5	6	
$\mathbf{D}^2 =$	1	0	1	4	9	16	4	$D_2 = 5 \times 1 + 5 \times 4 + 3 \times 9 + 2 \times 16 = 84$
	2	1	0	1	4	9	1	
	3	4	1	0	1	4	4	From square distances:
	4	9	4	1	0	1	9	
	5	16	9	4	1	0	16	$MSD = \left(\frac{10 \times 1 + 10 \times 4 + 6 \times 9 + 4 \times 16}{6 \times (6-1)} \right)^{1/2}$
	6	4	1	4	9	16	0	
								$= \left(\frac{168}{30} \right)^{1/2} = 2.366$
								From distance counts:
								$MSD = \left(\frac{5 \times 1^2 + 5 \times 2^2 + 3 \times 3^2 + 2 \times 4^2}{5 + 5 + 3 + 2} \right)^{1/2}$
								$= \left(\frac{84}{15} \right)^{1/2} = 2.366$

The **reciprocal distance matrix**, denoted as \mathbf{D}^{-1} (or **Harary matrix**, \mathbf{H} , or **vertex Harary matrix**) is a square symmetric $A \times A$ matrix derived from the distance matrix as

$$[\mathbf{D}^{-1}]_{ij} \equiv [\mathbf{H}]_{ij} = \begin{cases} 1/d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where each off-diagonal element is the reciprocal of the topological distance d between the vertices considered [Ivanciuc, Balaban *et al.*, 1993b; Plavšić, Nikolić *et al.*, 1993b; Lučić, Miličević *et al.*, 2002]; the diagonal elements are zero by definition for simple molecular graphs. The original \rightarrow *Harary index* H is calculated from this matrix as an analogue of the \rightarrow *Wiener index* (see below).

For vertex- and edge-weighted molecular graphs, the reciprocal distance matrix was defined as [Ivanciuc, 2000i]:

$$[\mathbf{D}^{-1}(w)]_{ij} = \begin{cases} 1/[\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ [\mathbf{D}(w)]_{ii} & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a \rightarrow *weighted distance matrix* and w denotes a \rightarrow *weighting scheme*.

The **reciprocal distance sum** RDS_i of the i th vertex is a local invariant defined as the sum of the reciprocal distance matrix elements in the i th row [Ivanciuc, Balaban *et al.*, 1993b]:

$$RDS_i \equiv VS_i(\mathbf{D}^{-1}) = \sum_{j=1}^A d_{ij}^{-1} \quad j \neq i$$

where the symbol VS stands for the \rightarrow row sum operator. From this local vertex invariant, the **Harary index** or **RDSUM index** is derived as the half sum of RDS_i over all the vertices [Plavšić, Nikolić *et al.*, 1993b]:

$$H \equiv RDSUM \equiv Wi(D^{-1}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [D^{-1}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A RDS_i$$

where the symbol Wi refers to \rightarrow Wiener operator.

Moreover, two other molecular descriptors, called **RDSQ index** and **RDCHI index**, respectively, were defined, based on a Randić-like formula [Ivanciuc, Balaban *et al.*, 1993b]:

$$RDSQ = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{1/2} \quad RDCHI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2}$$

where A is the number of vertices and a_{ij} is equal to 1 only for pairs of adjacent vertices and zero otherwise. While the **RDSQ** index increases with both molecular size and \rightarrow molecular branching, the **RDCHI** index increases with molecular size and decreases with molecular branching.

By analogy with Kier–Hall connectivity indices, **Topological Distance Connectivity Indices** (TDCIs), also called **Topological Distance Measure Connectivity Indices** (TDMCIs), were also defined by using the reciprocal distance sum RDS in place of the \rightarrow vertex degree δ [Balaban, Ciubotariu *et al.*, 1990; Ciubotariu, Medeleanu *et al.*, 2004]:

$$\begin{aligned} {}^1\tau &= \sum_{i=1}^A (RDS_i)^{-1/2} \\ {}^2\tau &\equiv RDCHI = \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2} \\ {}^3\tau &= \sum_{k=1}^{2P} (RDS_i \cdot RDS_l \cdot RDS_j)^{-1/2}_k \end{aligned}$$

where a_{ij} is equal to 1 only for pairs of adjacent vertices and zero otherwise. The third index ${}^3\tau$ is derived from paths of length 2 in the graph by weighting each path by the product of the reciprocal distance sums of the vertices involved in the path; 2P is the total number of paths of length 2.

A local vertex invariant, called **generalized reciprocal distance sum**, was proposed as

$${}^\lambda RDS_i = \sum_{j=1}^A d_{ij}^{-\lambda} \quad j \neq i$$

where λ is a positive integer; the **Generalized Topological Distance Indices** (GTDIs), denoted as ${}^1\tau^\lambda$, ${}^2\tau^\lambda$, and ${}^3\tau^\lambda$, were derived from the generalized reciprocal distance sum and defined by using the same formulas as for the TDCIs [Ciubotariu, Medeleanu *et al.*, 2004]. An additional descriptor, the generalized topological distance index of order zero, was also defined as

$${}^0\tau^\lambda = \sum_{i=1}^A {}^\lambda RDS_i = \sum_{i=1}^A \sum_{j=1}^A d_{ij}^{-\lambda} \quad \lambda = 1, 2, 3, 4, \dots \quad i \neq j$$

For $\lambda = 1$, this index coincides with twice the Harary index and for $\lambda = 2$ with twice the Harary number (see below).

From the Harary matrix, a \rightarrow *Balaban-like index*, called **Harary–Balaban index** $^R J$ [Nikolić, Plavšić *et al.*, 2001] or **Harary-connectivity index** [Randić and Pompe, 2001b] was also proposed as

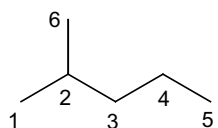
$$^R J \equiv J(\mathbf{D}^{-1}) = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot (RDS_i \cdot RDS_j)^{-1/2}$$

where RDS_i and RDS_j are the reciprocal distance sum of vertex v_i and vertex v_j , respectively; a_{ij} are elements of the adjacency matrix equal to 1 only for pairs of bonded vertices, and zero otherwise, B is the number of edges, C the \rightarrow *cyclomatic number*, and $J(\mathbf{D}^{-1})$ stands for \rightarrow *Balaban-like indices*. Note that in the formula proposed by Randić–Pompe, the coefficient $B/(C+1)$ is not considered.

Moreover, the reciprocal distance sums of the graph vertices were proposed as an alternative to vertex distance degrees to detect the \rightarrow *graph center* [Plavšić, Nikolić *et al.*, 1993b].

Example D10

Reciprocal distance matrix \mathbf{D}^{-1} and related molecular descriptors of 2-methylpentane. RDS_i is the reciprocal distance sum of the i th vertex.

	Atom	1	2	3	4	5	6	RDS_i
	1	0	1	0.50	0.33	0.25	0.50	2.58
	2	1	0	1	0.50	0.33	1	3.83
	3	0.50	1	0	1	0.50	0.50	3.50
	4	0.33	0.50	1	0	1	0.33	3.16
	5	0.25	0.33	0.50	1	0	0.25	2.33
	6	0.50	1	0.50	0.33	0.25	0	2.58

$$H \equiv RDSUM \equiv W(\mathbf{D}^{-1}) = (2.58 + 3.83 + 3.50 + 3.16 + 2.33 + 2.58)/2 = 17.98/2 = 8.99$$

$$RDSQ = (2.58 \times 3.83)^{1/2} + (3.83 \times 3.50)^{1/2} + (3.83 \times 2.58)^{1/2} + (3.50 \times 3.16)^{1/2} + (3.16 \times 2.33)^{1/2} + \sqrt{9.88} + \sqrt{13.40} + \sqrt{9.88} + \sqrt{11.06} + \sqrt{7.36} = 15.99$$

$$RDCHI = (2.58 \times 3.83)^{-1/2} + (3.83 \times 3.50)^{-1/2} + (3.83 \times 2.58)^{-1/2} + (3.50 \times 3.16)^{-1/2} + (3.16 \times 2.33)^{-1/2} = 1.58$$

$$^R J \equiv J(\mathbf{D}^{-1}) = \frac{5}{0+1} \times [(2.58 \times 3.83)^{-1/2} + (3.83 \times 3.50)^{-1/2} + (3.83 \times 2.58)^{-1/2} + (3.50 \times 3.16)^{-1/2} + (3.16 \times 2.33)^{-1/2}] = 7.89$$

Analogously, the **reciprocal square distance matrix** \mathbf{D}^{-2} is derived from the distance matrix by the following:

$$[\mathbf{D}^{-2}]_{ij} = \begin{cases} 1/d_{ij}^2 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where each off-diagonal element is the square reciprocal of the corresponding element of the distance matrix. This matrix can be useful to take into account the fact that atom interactions

decrease as the square distance between atoms increases, such as in the \rightarrow *topological charge indices*. A variant of the Harary index H (called **Harary number** and denoted as H') was derived from this matrix and proposed as molecular descriptor [Mihalić and Trinajstić, 1992]:

$$H' \equiv Wi(\mathbf{D}^{-2}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}^{-2}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{D}^{-2})$$

where Wi refers to the \rightarrow *Wiener operator* and VS_i denotes the i th row sum of the reciprocal square distance matrix:

$$VS_i(\mathbf{D}^{-2}) = \sum_{j=1}^A d_{ij}^{-2} \quad j \neq i$$

Example D11

Reciprocal square distance matrix \mathbf{D}^{-2} of 2-methylpentane. VS_i indicates the matrix row sum.

Atom	1	2	3	4	5	6	VS_i
1	0	1	0.25	0.111	0.063	0.25	1.674
2	1	0	1	0.25	0.111	1	3.361
3	0.25	1	0	1	0.25	0.25	2.750
4	0.111	0.25	1	0	1	0.111	2.472
5	0.063	0.111	0.25	1	0	0.063	1.486
6	0.25	1	0.25	0.111	0.063	0	1.674

$$H' \equiv Wi(\mathbf{D}^{-2}) = \frac{1}{2} \times (1.674 + 3.361 + 2.750 + 2.472 + 1.486 + 1.674) = \frac{13.416}{2} = 6.708$$

The **distance complement matrix** (also called **reversed distance matrix** or **vertex distance complement matrix**), denoted as **DC**, for simple graphs is defined as [Randić, 1997a; Randić and Pompe, 2001b]:

$$[\mathbf{DC}]_{ij} = \begin{cases} A - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of vertices and d_{ij} the distance between vertices v_i and v_j . The half-sum of the matrix elements is the **complement Wiener index** (or **reversed Wiener index**).

For vertex- and edge-weighted molecular graphs, the distance complement matrix was defined as [Ivanciuc, 2000i; Ivanciuc, 2002c]

$$[\mathbf{DC}(w)]_{ij} = \begin{cases} A - [\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a \rightarrow *weighted distance matrix* and w denotes a \rightarrow *weighting scheme*.

Reciprocal and complement distance matrices were defined because, unlike the distance matrix, the value of the matrix elements corresponding to pairs of vertices decreases when the distance between vertices increases; therefore, molecular descriptors calculated from reciprocal or complement matrices have numerical behavior and meaning opposite to molecular

Table D7 Some molecular descriptors derived from distance, reciprocal distance, and reciprocal square distance matrices for C8 data set (Appendix C – Set 1).

C8	W	\bar{W}	H	H'	RDCHI	RDSQ
n-Octane	84	3.000	9.502	13.743	1.997	24.823
2M	79	2.821	9.731	14.100	1.909	25.922
3M	76	2.714	9.814	14.267	1.885	26.379
4M	75	2.679	9.837	14.317	1.879	26.510
3E	72	2.571	9.920	14.483	1.851	26.966
22MM	71	2.536	10.176	14.767	1.774	28.000
23MM	70	2.500	10.108	14.733	1.788	27.791
24MM	71	2.536	10.059	14.650	1.798	27.552
25MM	74	2.643	9.966	14.467	1.823	27.047
33MM	67	2.393	10.318	15.033	1.737	28.744
34MM	68	2.429	10.179	14.867	1.768	28.164
2M3E	67	2.393	10.201	14.917	1.760	28.294
3M3E	64	2.286	10.438	15.250	1.703	29.355
223MMM	63	2.250	10.576	15.417	1.658	29.940
224MMM	66	2.357	10.431	15.167	1.689	29.222
233MMM	62	2.214	10.625	15.500	1.646	30.180
234MMM	65	2.321	10.389	15.167	1.700	29.120
2233MMMM	58	2.071	11.000	16.000	1.549	31.825

W is the Wiener index, \bar{W} the average Wiener index, H the Harary index, H' the Harary number, RDCHI the RDCHI index, and RDSQ the RDSQ index.

descriptors derived from the distance matrix. The row sums of the distance matrix **D** (i.e., distance degrees) are greater for outer vertices than for the core vertices; thus, for instance, the value of the \rightarrow *Balaban distance connectivity index*, which is based on the inverse of the distance degrees, is much more determined by the core vertices than the outer ones. On the contrary, the row sums of the reciprocal or complement distance matrix are greater for the core vertices and, therefore, \rightarrow *Balaban-like indices* derived from these matrices are much more determined by the outer vertices.

From the weighted distance complement matrix **DC**, the *complement Wiener indices* were derived as \rightarrow *Wiener-type indices* [Ivanciuc, 2000i]:

$$Wi(\mathbf{DC}, w) = \sum_{i=1}^A \sum_{j=i}^A [\mathbf{DC}(w)]_{ij}$$

where w is a weighting scheme for molecular graphs. Different complement Wiener indices are obtained depending on the weighting scheme w .

Moreover, the **complement Balaban index** $^C J$ [Nikolić, Plavšić *et al.*, 2001] or **reversed Balaban index** $1/J$ [Randić and Pompe, 2001b] was derived as a Balaban-like index as

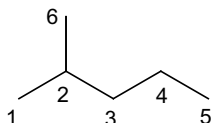
$$\begin{aligned}
 ^C J \equiv 1/J \equiv J(\mathbf{DC}) &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot ({}^C \sigma_i \cdot {}^C \sigma_j)^{-1/2} \\
 &= \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^A a_{ij} \cdot [(A \cdot (A-1) - \sigma_i) \cdot (A \cdot (A-1) - \sigma_j)]^{-1/2}
 \end{aligned}$$

where ${}^c\sigma$ is the row sum of the distance complement matrix, A the number of vertices, B and C are the number of edges and the \rightarrow *cyclomatic number*, respectively, and a_{ij} denotes the elements of the adjacency matrix equal to 1 for pairs of adjacent vertices and zero otherwise.

Example D12

Distance complement matrix and related molecular descriptors for 2-methylpentane.

${}^c\sigma$ indicates the matrix row sums.



DC =

Atom	1	2	3	4	5	6	${}^c\sigma$
1	0	5	4	3	2	4	18
2	5	0	5	4	3	5	22
3	4	5	0	5	4	4	22
4	3	4	5	0	5	3	20
5	2	3	4	5	0	2	16
6	4	5	4	3	2	0	18

$$Wi(\mathbf{DC}) = \frac{1}{2} \times (18 \times 2 + 22 \times 2 + 20 + 16) = 58$$

$${}^cJ \equiv 1/J \equiv J(\mathbf{DC}) =$$

$$= 5 \times \left(\frac{1}{\sqrt{18 \times 22}} + \frac{1}{\sqrt{22 \times 22}} + \frac{1}{\sqrt{22 \times 18}} + \frac{1}{\sqrt{22 \times 20}} + \frac{1}{\sqrt{20 \times 16}} \right) = 1.250$$

A variant of the distance complement matrix is the **complementary distance matrix CD**, which for simple graphs was defined as [Balaban, Mills *et al.*, 2000; Ivanciuc, Ivanciuc *et al.*, 2000a; Ivanciuc, 2000i]:

$$[\mathbf{CD}]_{ij} = \begin{cases} 1 + D - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where D is the molecule diameter, which is the maximum distance in the graph, and 1 represents the minimum distance in the graph.

For vertex- and edge-weighted molecular graphs, the complementary distance matrix was defined as [Ivanciuc, 2000i]

$$[\mathbf{CD}(w)]_{ij} = \begin{cases} d_{\min}(w) + d_{\max}(w) - [\mathbf{D}(w)]_{ij} & \text{if } i \neq j \\ w_i & \text{if } i = j \end{cases}$$

where $\mathbf{D}(w)$ is a \rightarrow *weighted distance matrix*, w denotes a \rightarrow *weighting scheme*, d_{\min} and d_{\max} are the minimum and maximum distances in the weighted molecular graph, respectively. Obviously, for simple graphs, $d_{\min} = 1$ and $d_{\max} = D$.

A number of topological indices were derived from the complementary distance matrix **CD** and tested in QSPR models [Ivanciuc, Ivanciuc *et al.*, 2000a]. Among these are the **complementary Wiener indices** derived from the complementary distance matrix **CD** as \rightarrow *Wiener-type indices* [Ivanciuc, 2000i]:

$$Wi(\mathbf{CD}, w) = \sum_{i=1}^A \sum_{j=i}^A [\mathbf{CD}(w)]_{ij}$$

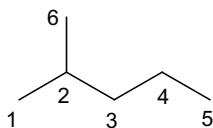
where w is a \rightarrow *weighting scheme*. Different complementary Wiener indices are obtained depending on the weighting scheme w for molecular graphs.

A simple relationship between the complementary Wiener index and the \rightarrow *Wiener index* W holds [Ivanciuc, Ivanciuc *et al.*, 2002b]:

$$Wi(\mathbf{CD}) = \frac{1}{2} \cdot (d_{\max} + d_{\min}) \cdot A \cdot (A-1) - W$$

Example D13

Complementary distance matrix \mathbf{CD} and related Wiener-type index $Wi(\mathbf{CD})$ for 2-methylpentane. VS_i indicates the matrix row sums.



$$\mathbf{CD} =$$

Atom	1	2	3	4	5	6	VS_i
1	0	4	3	2	1	3	13
2	4	0	4	3	2	4	17
3	3	4	0	4	3	3	17
4	2	3	4	0	4	2	15
5	1	2	3	4	0	1	11
6	3	4	3	2	1	0	13

$$Wi(\mathbf{CD}) = \frac{1}{2} \cdot (13 + 17 + 17 + 15 + 11 + 13) = 43$$

Very similar to the complementary distance matrix \mathbf{CD} , the **reverse Wiener matrix**, denoted by \mathbf{RW} , was defined as [Balaban, Mills *et al.*, 2000]

$$[\mathbf{RW}]_{ij} = \begin{cases} D - d_{ij} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where D is the molecule diameter, which is the maximum distance in the graph. Note that all the entries in the reverse Wiener matrix are lower by one than those in the complementary distance matrix for simple graphs. Another property of the reversed Wiener matrix is that matrix elements corresponding to the diameter in the distance matrix have zero values in the reverse Wiener matrix.

The i th row sum of the reverse Wiener matrix was called **reverse-distance sum** and defined as

$$VS_i(\mathbf{RW}) = \sum_{j=1}^A [\mathbf{RW}]_{ij}$$

where VS_i is the row sum operator.

The half sum of the matrix elements was proposed as molecular descriptor with the name of **reverse Wiener index**, denoted by Λ [Balaban, Mills *et al.*, 2000]:

$$\Lambda \equiv Wi(\mathbf{RW}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{RW}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A VS_i(\mathbf{RW})$$

where the symbol Wi denotes the \rightarrow *Wiener operator*. The following relationship between the reverse Wiener index and the \rightarrow *Wiener index* W holds [Ivanciuc, Ivanciuc *et al.*, 2002b]:

$$Wi(\mathbf{RW}) = \frac{1}{2} \cdot D \cdot A \cdot (A-1) - W$$

Moreover, the \rightarrow *spectral indices* $MinSp(\mathbf{RW})$ and $MaxSp(\mathbf{RW})$, the \rightarrow *Hosoya-type index* $Ho(\mathbf{RW})$, the \rightarrow *hyper-Wiener-type index* $HyWi(\mathbf{RW})$, the \rightarrow *Balaban-like index* $J(\mathbf{RW})$, and \rightarrow *Balaban-like information indices* $U(\mathbf{RW})$, $V(\mathbf{RW})$, $X(\mathbf{RW})$, and $Y(\mathbf{RW})$ were tested in property modeling of some alkanes [Ivanciuc, Ivanciuc *et al.*, 2002b].

A **standardized complementary distance matrix** was also proposed to obtain **constant interval reciprocal indices** (CIR indices); the idea was that standardized distances should be uniformly spaced as topological distances [Schultz and Schultz, 1998, 2000]. The elements of the standardized complementary distance matrix are defined as

$$[\mathbf{CIR}]_{ij} = \begin{cases} \frac{1 + d_{\max} - d_{ij}}{d_{\max}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where d_{ij} denotes the elements of the distance matrix, d_{\max} is the largest integer in the distance matrix, that is, the graph diameter D , or in the i th row, that is, the \rightarrow *atom eccentricity* η_i , or arbitrarily user-defined. If d_{\max} is chosen as the largest value in each row, the resulting matrix is unsymmetric. Moreover, d_{\max} can also be equal to the number of graph vertices A .

Using these modified standardized distance matrices instead of the classical distance matrix, **CIRD indices** were defined as \rightarrow *Wiener-type indices* by analogy with the \rightarrow *Wiener index*:

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

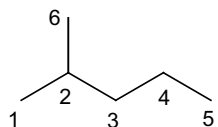
Moreover, **CIRS indices** and **CIRS' indices** were defined by analogy with the \rightarrow *reciprocal Schultz indices*:

$$CIRS = \sum_{i=1}^A [(\mathbf{A} + \mathbf{CIR}) \cdot \mathbf{v}]_i \quad CIRS' = \sum_{i=1}^A \sum_{j=1}^A [\mathbf{CIR} \cdot \mathbf{v}]_{ij} = \sum_{i=1}^A \delta_i \cdot {}^{CIR}\sigma_i$$

where \mathbf{A} is the \rightarrow *adjacency matrix*, \mathbf{v} is the A -dimensional vector of \rightarrow *vertex degrees* δ and ${}^{CIR}\sigma_i$ is the i th row sum of the **CIR** matrix. Different $CIRD$, $CIRS$, and $CIRS'$ indices can be calculated depending on the value chosen for the d_{\max} parameter.

Example D14

Standardized complementary distance matrix **CIR** and related molecular descriptors for 2-methylpentane. Calculation is based on $d_{\max} = D = 4$.

	Atom	1	2	3	4	5	6	$CIR\sigma_i$
	1	0	1	0.75	0.50	0.25	0.75	3.25
	2	1	0	1	0.75	0.50	1	4.25
	3	0.75	1	0	1	0.75	0.75	4.25
	4	0.25	0.75	1	0	1	0.50	3.50
	5	0.25	0.50	0.75	1	0	0.25	2.75
	6	0.75	1	0.75	0.50	0.25	0	3.25

$$CIRD = 5 \times 1 + 2 \times 0.25 + 3 \times 0.50 + 5 \times 0.75 = 10.75$$

$$CIRS' = 1 \times 3.25 + 3 \times 4.25 + 2 \times 4.25 + 2 \times 3.50 + 1 \times 2.75 + 1 \times 3.25 = 37.50$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0.75 & 0.50 & 0.25 & 0.75 \\ 1 & 0 & 1 & 0.75 & 0.50 & 1 \\ 0.75 & 1 & 0 & 1 & 0.75 & 0.75 \\ 0.25 & 0.75 & 1 & 0 & 1 & 0.50 \\ 0.25 & 0.50 & 0.75 & 1 & 0 & 0.25 \\ 0.75 & 1 & 0.75 & 0.50 & 0.25 & 0 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \\ 2 \\ 2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9.50 \\ 10.00 \\ 12.25 \\ 9.00 \\ 7.50 \\ 9.50 \end{bmatrix}$$

$$CIRS = 9.50 + 10.00 + 12.25 + 9.00 + 7.50 + 9.50 = 57.75$$

Two quotient matrices were derived from the distance complement matrix **DC**.

The **distance/distance complement quotient matrix**, denoted by **D/DC**, was defined as [Nikolić, Plavšić *et al.*, 2001]

$$[D/DC]_{ij} = \begin{cases} \frac{d_{ij}}{A-d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

and its reciprocal, the **distance complement/distance quotient matrix**, denoted by **DC/D**, as [Randić and Pompe, 2001b; Nikolić, Plavšić *et al.*, 2001]

$$[DC/D]_{ij} = \begin{cases} \frac{A-d_{ij}}{d_{ij}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

where A is the number of graph vertices and d_{ij} the topological distance between vertices v_i and v_j .

From the matrix \mathbf{D}/\mathbf{DC} and the matrix \mathbf{DC}/\mathbf{D} , two \rightarrow *Balaban-like indices* were derived and called **quotient Balaban index of the first kind** and **quotient Balaban index of the second kind**, respectively [Nikolić, Plavšić *et al.*, 2001].

From distance complement matrix \mathbf{DC} , complementary distance matrix \mathbf{CD} , and reverse Wiener matrix \mathbf{RW} , the corresponding reciprocal matrices, called **reciprocal distance complement matrix** \mathbf{DC}^{-1} [Ivanciuc, 2000i], **reciprocal complementary distance matrix** \mathbf{CD}^{-1} [Ivanciuc, 2000i], and **reciprocal reverse Wiener matrix** \mathbf{RRW} [Balaban, Mills *et al.*, 2000], were also derived by substituting the off-diagonal matrix elements by the corresponding reciprocal and leaving the diagonal elements unchanged. Moreover, \rightarrow *Wiener-type indices* were derived from these reciprocal matrices to be used in QSAR modeling.

Opposite to the distance matrix is the \rightarrow *detour matrix*, where the entries correspond to the length of the longest path between two vertices. Other matrices related to the distance matrix are \rightarrow *geometric distance/topological distance quotient matrix*, \rightarrow *detour–distance combined matrix*, \rightarrow *distance/detour quotient matrix*, \rightarrow *distance–degree matrices*, \rightarrow *expanded distance matrices*, \rightarrow *distance–path matrix*, \rightarrow *delta matrix*, \rightarrow *distance sum layer matrix*, \rightarrow *distance–sequence matrix*.

Note. The distance matrix derived from a graph must not be confused with the distance matrix derived from a \rightarrow *data set*, called here \rightarrow *data distance matrix*.

📖 [Entiger, Jackson *et al.*, 1976; Doyle and Garver, 1977; Bersohn, 1983; Plesnik, 1984; Rouvray, 1986a; Müller, Szymanski *et al.*, 1987; Senn, 1988; Mihalić, Nikolić *et al.*, 1992; Kunz, 1993, 1994; Thangavel and Venuvanalingam, 1993; Dobrynin and Gutman, 1994; Dobrynin, 1995; Chepoi, 1996; Ivanciuc, Laidboeur *et al.*, 1997; Ivanciuc and Ivanciuc, 1999]

- **distance measures** \rightarrow similarity/diversity
- **distance measure connectivity indices** \rightarrow combined descriptors
- **distance-normalized exponential sum connectivities** \rightarrow exponential sum connectivities
- **distance number** \equiv *distance degree* \rightarrow distance matrix

■ distance-path matrix ($\mathbf{D_P}$)

The distance-path matrix or **vertex distance-path matrix**, denoted as $\mathbf{D_P}$, is a \rightarrow *combinatorial matrix* derived from the \rightarrow *distance matrix* \mathbf{D} . It is a square symmetric matrix $A \times A$ whose off-diagonal entries are the number of all paths of any length m ($1 \leq m \leq d_{ij}$) that are included in the shortest path from vertex v_i to vertex v_j (whose \rightarrow *topological distance* is d_{ij}) [Diudea, 1996a]. The diagonal entries are zero.

Each entry i – j of the matrix $\mathbf{D_P}$ is calculated as the following:

$$[\mathbf{D_P}]_{ij} = \binom{d_{ij} + 1}{2} = \frac{d_{ij}^2 + d_{ij}}{2}$$

that is, as all the possible combinations of two elements taken from $d_{ij} + 1$ elements (binomial coefficient).

For vertex- and edge-weighted molecular graphs, the distance-path matrix was defined as [Ivanciuc and Ivanciuc, 1999; Ivanciuc, 2000i]

$$[\mathbf{D_P}(w)]_{ij} = \binom{[\mathbf{D}(w)]_{ij} + 1}{2} = \frac{[\mathbf{D}(w)]_{ij}^2 + [\mathbf{D}(w)]_{ij}}{2}$$

where $\mathbf{D}(w)$ is a \rightarrow *weighted distance matrix* calculated on the basis of the \rightarrow *weighting scheme* w . Unlike the distance-path matrix of simple graphs, whose diagonal elements are equal to zero, the distance-path matrix of weighted graphs has diagonal elements different from zero since depending on the vertex weights.

The **hyper-distance-path index** D_P can be obtained by applying the \rightarrow *Wiener operator* Wi to the distance-path matrix as

$$D_P \equiv Wi(\mathbf{D_P}) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D_P}]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A d_{ij} + \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \binom{d_{ij}}{2} = W + D_\Delta$$

where in the right expression the first term is just the \rightarrow *Wiener index* W and the second the \rightarrow *delta number* D_Δ , which can be considered the “non-Wiener part” of the index [Diudea, Katona *et al.*, 1997].

For acyclic graphs, the hyperdistance-path index D_P coincides with the \rightarrow *hyper-Wiener index* WW derived from the \rightarrow *Wiener matrix*, and with the \rightarrow *hyper-detour index* derived from the \rightarrow *detour-path matrix*. Moreover, it was proposed as extension of the hyper-Wiener index for any graph [Klein, Lukovits *et al.*, 1995].

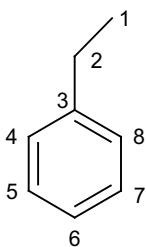
The **reciprocal distance-path matrix** $\mathbf{D_P}^{-1}$ is a matrix whose elements are the reciprocal of the corresponding distance-path matrix elements. The general definition for weighted molecular graphs is [Ivanciuc, 2000i]

$$[\mathbf{D_P}^{-1}(w)]_{ij} = \begin{cases} 1/[\mathbf{D_P}(w)]_{ij} & \text{if } i \neq j \\ [\mathbf{D_P}(w)]_{ii} & \text{if } i = j \end{cases}$$

The \rightarrow *hyper-Harary distance index* is the \rightarrow *Wiener-type index* derived from this matrix.

Example D15

The distance-path matrix $\mathbf{D_P}$ for ethylbenzene. VS_i indicates the matrix row sums. D_P is the hyper-distance-path index.



$\mathbf{D_P} =$

Atom	1	2	3	4	5	6	7	8	VS_i
1	0	1	3	6	10	15	10	6	51
2	1	0	1	3	6	10	6	3	30
3	3	1	0	1	3	6	3	1	18
4	6	3	1	0	1	3	6	3	23
5	10	6	3	1	0	1	3	6	30
6	15	10	6	3	1	0	1	3	39
7	10	6	3	6	3	1	0	1	30
8	6	3	1	3	6	3	1	0	23

$$D_P \equiv Wi(\mathbf{D_P}) = \frac{1}{2} \cdot (51 + 3 \times 30 + 18 + 2 \times 23 + 39) = 122$$

\rightarrow *Characteristic polynomial* of combinatorial matrices as well as \rightarrow *spectral indices* were proposed as molecular descriptors for QSAR problems [Ivanciuc, Diudea *et al.*, 1998].

The **delta matrix** or **distance–delta matrix** or **vertex distance–delta matrix**, denoted as \mathbf{D}_Δ , is another combinatorial matrix defined as the difference between the \rightarrow distance-path matrix \mathbf{D}_P and the \rightarrow distance matrix \mathbf{D} [Diudea, 1996a; Diudea, Katona *et al.*, 1997; Ivanciuc, Diudea *et al.*, 1998]:

$$\mathbf{D}_\Delta = \mathbf{D}_P - \mathbf{D}$$

The delta matrix is a square symmetric matrix of dimension $A \times A$, A being the number of graph vertices, whose entries represent the number of all paths larger than unity included in the shortest path between the considered vertices and are formally defined as

$$[\mathbf{D}_\Delta]_{ij} \equiv \begin{cases} \binom{d_{ij}}{2} = \frac{d_{ij} \cdot (d_{ij} - 1)}{2} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$

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

Applying the \rightarrow Wiener operator Wi to the delta matrix a molecular descriptor called **delta number** D_Δ is obtained as

$$D_\Delta \equiv Wi(\mathbf{D}_\Delta) = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A [\mathbf{D}_\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^A \sum_{j=1}^A \binom{d_{ij}}{2}$$

The delta number can also be derived by

$$D_\Delta = D_P - W$$

where W is the \rightarrow Wiener index and D_P the \rightarrow hyper-distance-path index which coincides with the \rightarrow hyper-Wiener index WW for acyclic graphs. It follows that the delta number is the “non-Wiener” part of the hyper-Wiener index.

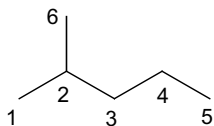
Moreover, the delta number D_Δ can be related to the distance matrix and the Wiener index W by

$$D_\Delta = \frac{\text{tr}(\mathbf{D}^2) - 2 \cdot W}{4} = \frac{D_2 - W}{2}$$

where \mathbf{D}^2 is the distance matrix raised to the second power, and D_2 is the second-order \rightarrow distance distribution moment, that is, the sum of the square distances in the graph.

Example D16

Delta matrix \mathbf{D}_Δ and delta number for the 2-methylpentane.



$\mathbf{D}_\Delta =$

Atoms	1	2	3	4	5	6
1	0	0	1	3	6	1
2	0	0	0	1	3	0
3	1	0	0	0	1	1
4	3	1	0	0	0	3
5	6	3	1	0	0	6
6	1	0	1	3	6	0

$$D_\Delta = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 [\mathbf{D}_\Delta]_{ij} = \frac{1}{2} \cdot \sum_{i=1}^6 \sum_{j=1}^6 \frac{d_{ij}^2 - d_{ij}}{2} = 26$$

- **distance of a vertex** \equiv *distance degree* \rightarrow distance matrix
- **distance polynomial** \rightarrow characteristic polynomial-based descriptors
- **Distance Profile descriptors** \rightarrow substructure descriptors
- **distance rank** \equiv *distance degree* \rightarrow distance matrix
- **distance/resistance quotient matrix** \rightarrow resistance matrix
- **distance-sequence matrix** \rightarrow sequence matrices
- **distance sum** \equiv *distance degree* \rightarrow distance matrix
- **distance-sum-connectivity matrix** \rightarrow weighted matrices (\odot weighted adjacency matrices)
- **distance sum layer matrix** \rightarrow layer matrices
- **distance-valency matrices** \rightarrow distance-degree matrices
- **Diverse Property-Derived method** \rightarrow cell-based methods
- **diversity** \rightarrow similarity/diversity
- **diversity matrix** \equiv *data distance matrix* \rightarrow similarity/diversity
- **DLI score** \rightarrow scoring functions
- **DNA sequences** \rightarrow biodescriptors
- **docking** \rightarrow drug design
- **donor superdelocalizability** \equiv *nucleophilic superdelocalizability* \rightarrow quantum-chemical descriptors
- **Dosmoro complexity index** \rightarrow molecular complexity
- **double-bond count** \rightarrow multiple bond descriptors
- **double bond equivalenst** \equiv *index of hydrogen deficiency* \rightarrow multiple bond descriptors
- **DOS** \equiv *Density Of States* \rightarrow quantum-chemical descriptors (\odot EIM descriptors)

■ double invariants

Double invariants include various molecular descriptors calculated by a general approach for the derivation of graph-invariants, which generates a graph-theoretical matrix whose elements are molecular subgraphs rather than numbers [Randić, Plavšić *et al.*, 1997]. Matrices of this kind are called **graphical matrices** denoted as **GG**; they are square symmetric matrices of dimension $A \times A$, A being the number of graph vertices, and the element $i-j$ is defined as a certain subgraph in some way related to vertices v_i and v_j . There are different approaches to derive graphical matrices. In the first graphical matrix, called **path matrix**, the element $i-j$ was defined as the subgraph made up of all the shortest paths joining vertices v_i and v_j ; this subgraph may be a path or a cyclic fragment, including the molecule as a whole. The diagonal elements are isolated vertices.

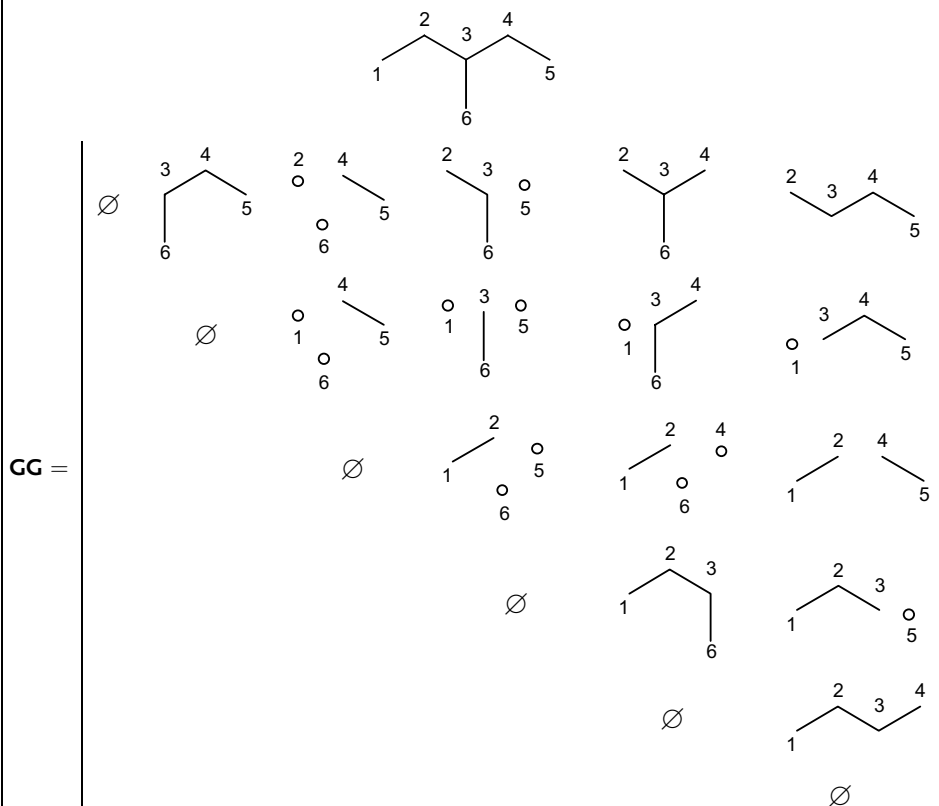
Another graphical matrix was defined so that the off-diagonal element $i-j$ is the subgraph obtained by deleting vertices v_i and v_j and incident edges from the molecular graph [Randić, Basak *et al.*, 2004]. By definition, all the diagonal elements are an empty subgraph. In this vertex-graphical matrix, the entries may be disconnected fragments and all the fragments have the same number of vertices. *Vertex-graphical matrices* can be further distinguished into *dense*, if all the matrix elements except for the diagonal elements are nonzero, and *sparse*, if only off-diagonal elements corresponding to pairs of adjacent vertices are different from the empty subgraph [Nikolić, Miličević *et al.*, 2005].

Another way of constructing graphical matrices is to define their elements $i-j$ as the subgraph obtained by removal of the paths joining vertices v_i and v_j from the molecular graph [Nikolić, Miličević *et al.*, 2005]. These matrices include the *path-graphical matrices*, which are dense since

they are obtained by considering all the pairs of vertices, and the *edge-graphical matrices*, which are sparse since they collect only subgraphs derived by removal of paths connecting pairs of adjacent vertices.

Example D17

Vertex-graphical matrix of 3-methylpentane, obtained by the criterion proposed in [Randić, Basak *et al.*, 2004].



To calculate molecular descriptors, graphical matrices need to be transformed into numerical matrices; this transformation is carried out by means of a graph invariant (e.g., the \rightarrow *Wiener index*). Then, in the resulting numerical matrix the element $i-j$ is defined as the value of the selected graph invariant derived from the subgraph in position $i-j$ in the graphical matrix. By definition, all the elements corresponding to the entries of the graphical matrix being the empty subgraph are equal to zero. The graph invariant of a disconnected subgraph is calculated by adding the values of the graph invariant derived from all the components of the subgraph.

Finally, a second graph invariant, which can be different from the first invariant, is applied to the graphical matrix in the numerical form to obtain the double invariant.

Therefore, double invariants are functions of the type:

$$\mathcal{D}_2\{\mathcal{D}_1(\mathbf{GG})\}$$

It must be noted that the two invariants have to be compatible, that is, the domain of \mathcal{D}_2 is in the range of \mathcal{D}_1 ; moreover, the two invariant operators do not commute, that is,

$$\mathcal{D}_2\{\mathcal{D}_1(\mathbf{GG})\} \neq \mathcal{D}_1\{\mathcal{D}_2(\mathbf{GG})\}$$

Given the large number of available graph-invariants, there are many possible combinations of compatible invariants, resulting in an explosion of new descriptors.

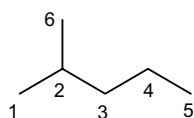
A molecular branching index was proposed for acyclic graphs, based on the double invariant approach and called $\rightarrow \lambda\lambda_1$ *branching index* [Randić, 1997d, 1998a]. A calculation example for 2-methylpentane is given below.

Other proposed double invariants are the **Wiener–Wiener number** [Randić, Basak *et al.*, 2004], denoted as $W(W)$, and the **Hosoya–Wiener index** [Nikolić, Miličević *et al.*, 2005], denoted as ZW .

The double invariant approach can be extended to 3D molecular geometry, defining graphical geometry matrices and the corresponding **3D double invariants**.

Example D18

Calculation of the double invariant $\lambda\lambda_1$ for 2-methylpentane: the graphical matrix \mathbf{GG} consists of the paths joining two vertices v_i and v_j ; the first invariant is the leading eigenvalue of the \rightarrow *adjacency matrix* \mathbf{A} of each subgraph and the double invariant is the leading eigenvalue of the graphical matrix in the numerical form. The symbol ${}^m p_{ij}$ represents a path of length m between vertices v_i and v_j .



$$\lambda\lambda_1 = 6.8313$$

$\mathbf{GG} =$

Atom	1	2	3	4	5	6
1	0	${}^1 p_{12}$	${}^2 p_{13}$	${}^3 p_{14}$	${}^4 p_{15}$	${}^2 p_{16}$
2		0	${}^1 p_{23}$	${}^2 p_{24}$	${}^3 p_{25}$	${}^1 p_{26}$
3			0	${}^1 p_{34}$	${}^2 p_{35}$	${}^2 p_{36}$
4				0	${}^1 p_{45}$	${}^3 p_{46}$
5					0	${}^4 p_{56}$
6						0

Leading eigenvalues of \mathbf{GG} matrix subgraphs

Atom	1	2	3	4	5	6
1	0	1.0000	1.4142	1.6180	1.7321	1.4142
2	1.0000	0	1.0000	1.4142	1.6180	1.0000
3	1.4142	1.0000	0	1.0000	1.4142	1.4142
4	1.6180	1.4142	1.0000	0	1.0000	1.6180
5	1.7321	1.6180	1.4142	1.0000	0	1.7321
6	1.4142	1.0000	1.4142	1.6180	1.7321	0

- **DP descriptor** → charge descriptors (⊙ submolecular polarity parameter)
- **DPD method** ≡ *Diverse Property-Derived method* → cell-based methods
- **DP indices** ≡ *indices of differences of path lengths*

■ DRAGON descriptors

These are various molecular descriptors ranging from → *count descriptors* to more complex → *geometrical descriptors* calculated by the software DRAGON [DRAGON – Talete s.r.l., 2007].

DRAGON software was conceived for the calculation of molecular descriptors by Milano Chemometrics and QSAR Research Group of Todeschini. DRAGON dates back to 1996, with the name of WHIM-3D/QSAR, while the first DRAGON version was released in 2000. Some details about DRAGON are given in Refs [Tetko, 2003; Mauri, Consonni *et al.*, 2006].

The number of DRAGON descriptors is 1664 in the version 5.4; however, this number increased up to 3224 after inclusion of 2D binary atom pairs and 2D frequency atom pairs, both consisting of 780 descriptors in version 5.5.

DRAGON descriptors are distinguished into 22 categories, which are listed in Table D8.

Table D8 List of descriptor categories in DRAGON software.

ID	Description	No.	ID	Description	No.
1	Constitutional descriptors	48	12	Geometrical descriptors	74
2	Topological descriptors	119	13	RDF descriptors	150
3	Walk and path counts	47	14	3D MoRSE descriptors	160
4	Connectivity indices	33	15	WHIM descriptors	99
5	Information indices	47	16	GETAWAY descriptors	197
6	2D autocorrelations	96	17	Functional group counts	154
7	Edge adjacency indices	107	18	Atom-centered fragments	120
8	Burden eigenvalues	64	19	Charge descriptors	14
9	Topological charge indices	21	20	Molecular properties	29
10	Eigenvalue-based indices	44	21	2D binary atom pairs	780
11	Randić molecular profiles	41	22	2D frequency atom pairs	780

Different molecular formats can be used for the descriptor calculation, and principal components can also be calculated, separately for each descriptor category.

In addition to descriptors' calculation, some explorative tools are also available that allow one to calculate descriptor values and their univariate statistics, to project and visualize molecules in the descriptor/response space, to calculate descriptor pair correlations, and to identify the most and the least correlated descriptors with a selected one.

The DRAGON software can be used in combination with the **MobyDigs software** [MobyDigs – Talete s.r.l., 2003; Todeschini, Consonni *et al.*, 2003], which allows QSAR/QSPR analysis by → *variable selection* based on genetic algorithms. To manage the huge number of molecular descriptors provided by DRAGON, several tools are available. Different fitness functions can be selected to evaluate the quality of QSAR models, several populations of models can be contemporarily developed, different validation techniques can be used for model validation, and predictions of the property of new molecules can be performed by means of selected models; → *consensus analysis* is also available.

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

➤ **drug** → drug design

■ drug design

Drug design is a research field where several disciplines are involved, including not only the design but also the pharmacokinetics and the toxicity of drugs. Appropriate chemometric tools, such as experimental design, multivariate analysis, artificial neural networks, are usually used in the planning and evaluation of pharmacokinetics and toxicological experiments, as well as in modeling of the biological activity of drugs. Moreover, → *similarity searching* and → *substructural analysis* are actually very useful for screening a large database of chemical compounds in the search for new drugs and pharmacophores [Purcell, Bass *et al.*, 1973; Martin, 1978; Ariëns, 1979; Franke, 1984; Dean, 1987; Hadzi and Jerman-Blazic, 1987; Ramsden, 1990; Tute, 1990; Kubinyi, 1993a; van de Waterbeemd, 1995; van de Waterbeemd, Testa *et al.*, 1997; Kubinyi, Folkers *et al.*, 1998a, 1998b].

Drug design techniques are all implemented in software packages that run on powerful workstations and can be distinguished as three main groups: **Computer-Aided Drug Design** (CADD) involves all computer-assisted techniques used to discover, design, and optimize biologically active compounds with a possible use as drugs; **Computer-Aided Molecular Design** (CAMD) involves all computer-assisted techniques used to discover, design, and optimize biologically active compounds with desired structure and properties; **Computer-Aided Molecular Modeling** (CAMM, or simply **Molecular Modeling**) is the investigation of molecular structures and properties using → *computational chemistry* and graphical visualization techniques.

Some fundamental terms and concepts in drug design are briefly reviewed below [van de Waterbeemd, Carter *et al.*, 1997; IUPAC Recommendations, 1997, 1998].

A **drug** is any substance for treating, curing, or preventing a disease in human beings or animals. A drug may also be used to make a medical diagnosis or to restore, correct, or modify physiological functions. A **lead compound** is a compound that, because of its biological properties, is taken as a reference structure that is the starting point for the process of identifying new active compounds; moreover, a lead compound should require the presence of at least one marketed drug, derived from that particular lead structure [Oprea, Davis *et al.*, 2001].

A **reference compound** (or **reference structure**) is a compound assumed as the reference for some considered aspect, such as a physico-chemical or biological property, or a molecular skeleton common to a set of compounds (→ *maximum common substructure*).

Lead discovery, generation, and optimization are basic activities in drug design, which are devoted to identifying active new chemical entities, developing new active compounds, and optimizing those able to be transformed into clinically useful drugs.

ADME properties (*Absorption, Distribution, Metabolism, and Elimination properties*) are properties of compounds that are of fundamental pharmaceutical importance; these are all those properties a drug needs to exert its pharmacological activity [van de Waterbeemd, Smith *et al.*, 2001a; Butina, Segall *et al.*, 2002; Ekins and Rose, 2002].

ADME properties have been recognized as a major reason for the failure of drug candidates; thus, several efforts are performed to develop molecular descriptors and models able to predict ADME properties of drug candidates before their synthesis. Bibliographic references to theoretical aspects and QSAR studies on ADME properties are reported in the thematic bibliography.

Bioisosterism is a strategy of medicinal chemistry for rational design of new drugs, applied to a lead compound as a special process of molecular modification consisting in the exchange of one bioisostere for another with the aim of enhancing biological activity without a significant change in the chemical structure [Burger, 1991; Cramer III, Clark *et al.*, 1996; Patani and LaVoie, 1996; Lima and Barreiro, 2005]. Bioisosteres are functional groups or substituents or molecular fragments with similar → *physico-chemical properties* that give similar biological properties to a chemical compound. To apply this approach, not only the chemical structure but also the → *mechanism of action*, at the level of interaction with the receptor, of the lead compound should be completely known, including the knowledge of all of its pharmacophoric centers.

Binding affinity is the tendency of a molecule to associate with another. In particular, the affinity of a drug is its ability to bind to its biological target, and a **ligand** is a compound that can bind to a biological target, thus giving origin to **docking**. Docking studies are computational techniques for the exploration of the possible binding modes of a ligand to a given receptor, enzyme, or other binding site.

A **receptor** is a molecule or a polymeric structure in a cell that acts as the biological target, recognizing and binding a compound. **Receptor mapping** is the technique used to describe geometric, electronic, and other physico-chemical features of a binding site. The active site cavity of a receptor is called the **binding site cavity** (or **receptor cavity**).

A **pharmacophore** is the ensemble of steric, electronic, and other physico-chemical properties that is necessary to ensure optimal supramolecular interactions with a specific biological target structure. In other words, the pharmacophore concept is based on the kinds of interactions observed in molecular recognition: hydrogen bonding, charge–charge, and hydrophobic interactions characterized by defined spatial arrangements. A pharmacophore does not represent a real molecule or a real association of functional groups, it is purely an abstract concept that accounts for the common molecular interaction capacity of a set of compounds with the target structure. A pharmacophore can be considered to be the largest common denominator shared by a set of active molecules.

The concepts of receptor and pharmacophore play a basic role in the alignment of molecules in → *grid-based QSAR techniques*.

A more detailed definition of pharmacophore is given by Bersuker who introduces three additional characteristics that are important in molecule-receptor interaction [Bersuker, 2003].

- 1 The pharmacophore should be defined by not just atoms from the periodic table, but by appropriate atom-in-molecule electronic characteristics.
- 2 Both electronic characteristics and geometry parameters of the pharmacophore vary from one active molecule to another within certain tolerances and the activity may be a function of these variations (i.e., pharmacophore flexibility).
- 3 The pharmacophore is a necessary but not sufficient condition of activity. Even if the pharmacophore is present, the activity of the molecule may be reduced by groups that hinder its proper docking with the receptor or may be enhanced by other groups responsible for properties useful to activity such as hydrophobicity.

The term pharmacophore is used in drug design while the term **toxiphore** is used in toxicology to refer to the set of unique molecular features in a given spatial

arrangement common to all the toxic molecules and deemed to represent the toxicity under consideration.

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

- **drug-like indices** → property filters
- **drug-like scores** \equiv *scoring functions*
- **DSI index** → electronegativity-based connectivity indices
- **dual degree** → vertex degree
- **dual electronic constants σ^+ and σ^-** → electronic substituent constants (⊙ resonance electronic constants)
- **Dubois steric constant** → steric descriptors (⊙ Taft steric constant)
- **Duchowitz–Castro $\log P$** → lipophilicity descriptors
- **dummy variables** \equiv *indicator variables*
- **Dunn model based on surface area** → lipophilicity descriptors
- **DV index** → multiple bond descriptors
- **d-WDEN indices** \equiv *directional WHIM density* → WHIM descriptors (⊙ directional WHIM descriptors)
- **d-WSHA indices** \equiv *directional WHIM shape* → WHIM descriptors (⊙ directional WHIM descriptors)
- **d-WSIZ indices** \equiv *directional WHIM size* → WHIM descriptors (⊙ directional WHIM descriptors)
- **d-WSYM indices** \equiv *directional WHIM symmetry* → WHIM descriptors (⊙ directional WHIM descriptors)
- **dynamic QSAR** → Structure/Response Correlations
- **dynamic reactivity indices** → reactivity indices
- **DZ^K descriptors** → autocorrelation descriptors
- **D/ Δ index** → detour matrix
- **D/ Δ ring indices** → detour matrix
- **D/ Ω index** → resistance matrix