F

- **false negative rate** → classification parameters
- **false positive rate** → classification parameters
- **farness** → center of a graph
- **FCFC fingerprints** → substructure descriptors (⊙ fingerprints)
- **FCFP fingerprints** ≡ *Functional Connectivity FingerPrints* → substructure descriptors (⊙ fingerprints)
- **feature maps** ≡ *topological feature maps*
- **feature reduction** ≡ *variable reduction*
- **feature tree** → molecular graph
- **Ferreira–Kiralj hydrophobicity parameters** → lipophilicity descriptors
- ***FHACA* index** → charged partial surface area descriptors
- ***FHASA* index** ≡ *RSAM index* → charged partial surface area descriptors
- **$FHASA_2$ index** → charged partial surface area descriptors (⊙ *RSAM* index)
- ***FHBCA* index** → charged partial surface area descriptors
- ***FHBSA* index** → charged partial surface area descriptors
- ***FHDCA* index** → charged partial surface area descriptors
- ***FHDSA* index** ≡ *RSHM index* → charged partial surface area descriptors
- **Fibonacci numbers** → symmetry descriptors (⊙ Merrifield–Simmons index)
- **Field characterization for Reaction Analysis and Understanding** ≡ *FRAU Features*
- **field effect** → electronic substituent constants
- **field-fitting alignment** → alignment rules
- **field-inductive constant** → electronic substituent constants (⊙ field/resonance effect separation)
- **field-inductive effect** ≡ *polar effect* → electronic substituent constants
- **field/resonance effect separation** → electronic substituent constants
- **^{19}F inductive constant** → electronic substituent constants (⊙ inductive electronic constants)
- **finite probability scheme** → equivalence classes
- **first eigenvector algorithm** → canonical numbering
- **fingerprints** → substructure descriptors
- **first-order sparse matrix** → algebraic operators (⊙ sparse matrices)
- **first Zagreb index** → Zagreb indices
- **FLAP fingerprints** → substructure descriptors (⊙ pharmacophore-based descriptors)
- **flash point** → physico-chemical properties
- **flexibility index based on path length** → flexibility indices

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

ISBN: 978-3-527-31852-0

flexibility indices

These are molecular descriptors proposed for the quantification of **molecular flexibility** (or, dually, **molecular rigidity**) and **bond flexibility** (or, dually, **bond rigidity**) [von der Lieth, Stumpf-Nothof *et al.*, 1996].

The concept of molecular flexibility is of primary importance in chemistry since molecular flexibility influences the chemical and biological properties of compounds as well as their interactions with other molecules.

Only a few attempts to quantify this concept can be found in the literature. Molecular dynamics and → *grid-based QSAR techniques* have been developed to account for conformational flexibility of compounds [Clark, Willett *et al.*, 1992, 1993; Hahn, 1997; Godha, Mori *et al.*, 2000].

In most cases, a local description of the flexibility is also required to distinguish between flexible and rigid parts of molecules. Therefore, both bond flexibility indices and molecular flexibility quantification were proposed.

Usually, the flexibility of a bond within a molecule is related to its → *bond order*, the nature of the atoms incident to the bond, bond participation in one or more cyclic structures, and the branching of adjacent atoms. A single acyclic bond formed by two C_{sp^3} atoms is regarded as freely rotatable, while bonds in polyaromatic ring systems, double and triple bonds are usually regarded as rigid. Analogously, the structural features decreasing molecular flexibility (or increasing molecular rigidity) are few atoms and the presence of rings and branching. It is usually assumed that a completely flexible molecule has an endless chain of C_{sp^3} atoms.

Most of the flexibility indices are derived from the → *H-depleted molecular graph.*

The most popular flexibility indices are listed below.

- **Rotatable Bond Number** (*RBN*)

This is the number of bonds that allow free rotation around themselves; these bonds are any single bond, not in a ring, bound to a nonterminal heavy atom. In other words, the rotatable bond number is the count of $C_{sp^3}-C_{sp^3}$ and $C_{sp^3}-C_{sp^2}$ bonds in the molecule, often excluding potentially rotatable bonds such as $-OH$ and $-CH_3$ [Bath, Poirrette *et al.*, 1995; Godha, Mori *et al.*, 2000]. It has also been suggested to exclude from the count amide C−N bonds because of their high rotational energy barrier [Veber, Johnson *et al.*, 2002].

As a general systematic rule, all single bonds that satisfy the following criteria are identified as rotatable bonds: (a) the heavy atoms (i.e., nonhydrogen atoms) A–B, connected by a single bond, must be connected to a second atom (C, D) as the following C–A–B–D. This second atom may be a hydrogen atom. (b) The external bond C–A or B–D must not be a triple bond unless the triple bonded atom is connected to another atom. (c) The bond A–B must not be part of a ring [Munk, Jørgensen and Pedersen, 2001].

Moreover, a general formula for the calculation of the number of rotatable bonds is [Head, Smythe *et al.*, 1996]

$$RBN = \mathrm{N}_{nt} + \sum_r (n_r - 4)$$

where N_{nt} is the number of nonterminal freely rotatable bonds, the summation goes over the rings in a molecule, and n_r is the number of single bonds in any nonaromatic ring.

Another formula for the rotatable bond number was proposed by [Oprea, 2000], taking explicitly into account the role of rings in a molecule:

$$RBN = \mathrm{N}_{nt} + \sum\nolimits_r (n_r - 4 - RGB_r - R_\mathrm{B})$$

where N_{nt} is the number of nonterminal freely rotatable bonds (but single bonds observed in groups such as sulfonamides (N−S) or esters (C−O) are excluded); the summation goes over the rings in a molecule, where n_r is the number of single bonds in the rth nonaromatic ring with six or more bonds, RGB_r is the number of rigid bonds in the rth ring, R_B is the number of bonds shared by the rth ring with any other ring, that is, the number of → *ring bridges.*

The complementary quantity to the rotatable bond number is the **rigid bond number**, denoted by RGB, which is defined as the difference between the total number of bonds in a molecule and the total number of rotatable bonds (including terminal single bonds) [Oprea, 2000; Zheng, Luo *et al.*, 2005].

The **rotatable bond fraction**, denoted by RBF, is the fraction of rotatable bonds over the total number of bonds:

$$RBF = \frac{RBN}{B}$$

where B is the total number of bonds in a molecule.

- **flexibility index based on path length** (F_K)

This is a topological index encoding information about molecular flexibility defined as a function of the length L of the longest chain in a molecule and the count ${}^3P$ of paths of length three [Kier and Hall, 1983c]:

$$F_K = \frac{L}{1 - 1/{}^3P}$$

The F_K index increases with increasing chaining and decreases with increasing branching.

F_K is set at zero by definition if no three-bond path is present, that is, for all compounds with ${}^3P = 0$.

- **Kier molecular flexibility index** (Φ)

This is a measure of molecular flexibility derived from the → *Kier alpha-modified shape descriptors* ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$:

$$\Phi = \frac{{}^1\kappa_\alpha \cdot {}^2\kappa_\alpha}{A}$$

where A is the total number of atoms in a molecule. The Kier shape indices calculated from the → *H-depleted molecular graph* depend on the heteroatoms by the parameter α [Kier, 1989; Kier and Hall, 1999d]; ${}^1\kappa_\alpha$ encodes information about the count of atoms and relative cyclicity of molecules, whereas ${}^2\kappa_\alpha$ encodes information about branching or relative spatial density of molecules. The atom count A allows comparisons among isomers.

- **global flexibility index** (GS)

This is a measure of molecular flexibility derived from additive contributions of path flexibilities as [von der Lieth, Stumpf-Nothof *et al.*, 1996]

$$GS = \frac{2}{A \cdot (A-1)} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} LS_{ij}$$

where A is the number of vertices in the H-depleted molecular graph. LS_{ij} is the **local simple flexibility index** relative to the path connecting vertices i and j, defined as

$$LS_{ij} = (d_{ij}+1) - \left(\frac{NRB + 0.75 \cdot {}^4F + 0.50 \cdot {}^3F}{2}\right)_{ij}$$

where d_{ij} is the → *topological distance* between vertices i and j, which is the length of the shortest path between the two vertices; NRB is the number of nonrotatable bonds in the path p_{ij}; ${}^4F$ and ${}^3F$ are the number of branching atoms with → *vertex degree* equal to four and three, that is, with four and three adjacent vertices, respectively, in the path p_{ij}.

An acyclic nonbranched chain of C_{sp^3} atoms is regarded to as completely flexible, and LS simply equals the number of atoms involved in the considered path.

- **bond flexibility index** (Φ_{BD})

This is an index encoding information about the flexibility of a bond and calculated as the mean of the atom flexibility indices Φ^a of the two atoms forming the bond [von der Lieth, Stumpf-Nothof *et al.*, 1996]:

$$\Phi_{BD} = \frac{\Phi_i^a + \Phi_j^a}{2}$$

where i and j are the atoms incident to the considered bond. Φ_{BD} is between 0 (not flexible) and 10 (completely flexible).

Flexibility indices Φ^a are defined for each molecule atom, provided it belongs to one of the defined molecular substructures: aromatic rings, multiple bonds, conjugated systems, simple rings, chains, and bridges. For all the atoms belonging to double bonds, triple bonds, or aromatic rings, $\Phi^a = 0.5$. For all atoms belonging to simple rings, Φ^a is equal to the → *Kier molecular flexibility index* Φ; for atoms in condensed ring systems, Kier molecular flexibility index Φ is calculated for both the simple rings and the condensed system, the smallest value is taken as the Φ^a value for each atom in the ring. Moreover, for each substitution group in the ring or condensed ring, a value of 0.2 is subtracted from the Φ^a value. The C_{sp^3} atoms in the chains and bridges are assigned a $\Phi^a = 10$ and this value is corrected by subtracting the sum of the atom masses of all the atoms in the next four adjacent shells, divided by 100.

The **bond rigidity index** ρ_{BD}, obtained from the bond flexibility index Φ_{BD} in the same range but with opposite meaning, is defined as

$$\rho_{BD} = 10 - \Phi_{BD}$$

- **Kier bond rigidity index** (ρ_{KB})

This is a measure of the rigidity/flexibility of a bond within a molecule, derived from the → *Kier molecular flexibility index* Φ. This index is defined as [von der Lieth, Stumpf-Nothof *et al.*, 1996]:

$$\rho_{KB} = \left(\sum\nolimits_k \Phi_k^f\right) - \Phi + 1$$

where Φ^f is the Kier flexibility index calculated for a single fragment. The summation goes over all molecule fragments obtained by breaking the bond of interest. It is a measure of bond rigidity because it represents the increase in flexibility of the fragments with respect to the parent molecule; this difference increases as the rigidity of the broken bond increases.

- **molecular flexibility number** (ϕ)

This is an index of molecule flexibility defined as [Dannenfelser and Yalkowsky, 1996; Jain, Yang *et al.*, 2004a, 2004b]

$$\phi = 2.85^{\tau}$$

where the value 2.85 is proportional to the difference in energy (kJ/mol) between trans and gauche conformations and τ is the number of torsional angles. The most common expression to derive the number of torsional angles τ is

$$\tau = \mathrm{N}_{\mathrm{sp}^3} + 0.5 \cdot \mathrm{N}_{\mathrm{sp}^2} + 0.5 \cdot NRG - 1$$

where $\mathrm{N}_{\mathrm{sp}^3}$, $\mathrm{N}_{\mathrm{sp}^2}$, and NRG are the number of sp^3-hybridized chain atoms, sp^2-hybridized chain atoms, and the number of fused ring systems, respectively.

🕮 [Luisi, 1977; Fisanick, Cross *et al.*, 1993; Bradbury, Mekenyan *et al.*, 1996; Bayada, Hemersma *et al.*, 1999; Oprea and Gottfries, 2001a; Martinek, Ötvös *et al.*, 2005]

➢ **flexible descriptors** ≡ *variable descriptors*
➢ **Flexsim-R fingerprints** → affinity fingerprints
➢ **Flexsim-S fingerprints** → affinity fingerprints
➢ **Flexsim-X fingerprints** → affinity fingerprints
➢ **F matrix** → layer matrices (⊙ cardinality layer matrix)
➢ ***F*-measure** → classification parameters
➢ **FM method** ≡ *Dewar–Grisdale approach* → electronic substituent constants (⊙ field/resonance effect separation)
➢ **FMMF method** ≡ *Dewar–Golden–Harris approach* → electronic substituent constants (⊙ field/resonance effect separation)
➢ **folding degree index** → spectral indices
➢ **folding profile** → spectral indices (⊙ folding degree index)
➢ **Forbes–Mozley similarity coefficient** → similarity/diversity (⊙ Table S9)
➢ **forest** → graph
➢ **formal degree** → weighted matrices (⊙ weighted adjacency matrices)
➢ **formal oxidation number** → multiple bond descriptors
➢ **forward Fukui function** → quantum-chemical descriptors (⊙ Fukui functions)
➢ **Fossum similarity coefficient** → similarity/diversity (⊙ Table S9)
➢ **Fourier analysis** → spectra descriptors

■ fractals

Fractals are geometric structures of fractional dimension; their theoretical fundamentals and physical applications were studied by Mandelbrot [Mandelbrot, 1982]. By definition, any structure possessing a self-similarity or a repeating motif invariant under a transformation of scale is called *fractal* and may be represented by a fractal dimension. Mathematically, the fractal dimension D_f of a set is defined through the relation

$$N(\varepsilon) \propto \varepsilon^{-D_f}$$

where N(ε) denotes the number of spheres of radius ε needed to cover the whole set. The fractal dimension of a set can be interpreted as the amount of information needed to fully specify a point of the set.

The concept of fractal dimension has been applied, for example, to represent tertiary structure of protein surface [Isogai and Itoh, 1984; Wagner, Colvin *et al.*, 1985; Åqvist and Tapia, 1987; Wang, Shi *et al.*, 1990; Poirrette, Artymiuk *et al.*, 1997; Tominaga and Fujiwara, 1997a; Tominaga, 1998a; Torrens, 2002], flexibility of alkanes [Rouvray and Pandey, 1986; Rouvray and Kumazaki, 1991], molecular shape [Mayer, Farin *et al.*, 1986], chromatographic profiles [Yiyu, Minjun *et al.*, 2003]; moreover fractals have been used in data structure comparison [Tominaga and Fujiwara, 1997a; Tominaga, 1998a] and in → *cell-based methods* [Agrafiotis and Rassokhin, 2002].

- **fractional bond order** → bond order indices
- **fractional charged partial negative surface areas** → charged partial surface area descriptors
- **fractional charged partial positive surface areas** → charged partial surface area descriptors
- **fragmental adjacency matrix** → adjacency matrix
- **fragmental connectivity index** → adjacency matrix
- **fragmental constants** ≡ *group contributions* → group contribution methods
- **fragmental degree** → adjacency matrix
- **fragment-based descriptors** ≡ *substructure descriptors*
- **fragment count** → count descriptors
- **fragment ID numbers** → ID numbers
- **Fragment Molecular Connectivity indices** → connectivity indices

■ **fragment topological indices** (FTI)

Derived from → *topological indices* calculated by graph dissections, fragment topological indices were proposed to reflect the interactions between the excised fragment and the remainder of the molecule [Mekenyan, Bonchev *et al.*, 1988a].

Let $\mathcal{G}$ be a → *H-depleted molecular graph* with A vertices and $\mathcal{G}'$ a subgraph (i.e., a fragment $\mathcal{F}$) of $\mathcal{G}$ with A' vertices, with $A' < A$ by definition. Topological indices (TI) that consider only vertices and edges belonging to the subgraph are called **internal fragment topological indices** and denoted by IFTI. The two following requirements were proposed for IFTI:

$$0 \leq \mathrm{IFTI}(\mathcal{F}) < \mathrm{TI}(\mathcal{G}) \quad \text{and} \quad 0 \leq \mathrm{IFTI}(\mathcal{G}-\mathcal{F}) < \mathrm{TI}(\mathcal{G})$$

where TI denotes one of the common defined topological indices, restricted to those increasing with the increase in the number of graph vertices. $\mathrm{IFTI}(\mathcal{G}-\mathcal{F})$ is the topological index calculated on the complementary subgraph.

Indices that describe a fragment in connection with the remainder of the graph are called **external fragment topological indices** and are denoted by EFTI. They were proposed as the difference in value between the topological index $\mathrm{TI}(\mathcal{G})$ for the whole graph and the internal fragment indices for both the fragment $\mathrm{IFTI}(\mathcal{F})$ and the remainder of the molecule $\mathrm{IFTI}(\mathcal{G}-\mathcal{F})$:

$$\mathrm{EFTI}(\mathcal{F}) = \mathrm{TI}(\mathcal{G}) - \left[\mathrm{IFTI}(\mathcal{F}) - \sum\nolimits_k \mathrm{IFTI}(\mathcal{G}-\mathcal{F})_k\right]$$

where the summation goes over all the $\mathcal{G}-\mathcal{F}$ disconnected components; there will be only one component if the subgraph $\mathcal{G}-\mathcal{F}$ is a connected graph. The following requirement was proposed for EFTI:

$$\mathrm{EFTI}(\mathcal{F}) < \mathrm{TI}(\mathcal{G})$$

The **normalized fragment topological indices** (NIFTI and NEFTI) may be obtained by dividing IFTI and EFTI by the topological index for the whole graph as

$$\mathrm{NIFTI}(\mathcal{F}) = \frac{\mathrm{IFTI}(\mathcal{F})}{\mathrm{TI}(\mathcal{G})} \quad \text{and} \quad \mathrm{NIFTI}(\mathcal{F}) = \frac{\mathrm{EFTI}(\mathcal{F})}{\mathrm{TI}(\mathcal{G})}$$

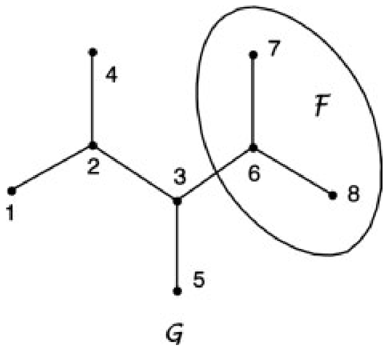
A special case of normalized fragment topological indices NIFTI is the → *graphical bond order.*

It has to be noted that IFTI($\mathcal{F}$) is a constant for a given fragment of any molecule, whereas NIFTI($\mathcal{F}$), EFTI($\mathcal{F}$) and NEFTI($\mathcal{F}$) depend upon the molecule as a whole. Moreover, the following general relation holds:

$$\mathrm{TI}(\mathcal{G}_1) > \mathrm{TI}(\mathcal{G}_2) \rightarrow \mathrm{EFTI}(\mathcal{F} \subset \mathcal{G}_1) > \mathrm{EFTI}(\mathcal{F} \subset \mathcal{G}_2)$$

Example F1

Calculation of internal and external first Zagreb index. The molecular fragment consists of vertices 6, 7, and 8, and **A** is the adjacency matrix.



$$\mathbf{A} = \begin{array}{c|ccccc|ccc}
\text{Atom} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
3 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
4 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hline
6 & 0 & 0 & 1 & 0 & 0 & \mathbf{0} & \mathbf{1} & \mathbf{1} \\
7 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{0} \\
8 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{0}
\end{array}$$

vertex degrees δ

Atom	$\mathcal{G}$	$\mathcal{F}$	$\mathcal{G}-\mathcal{F}$
1	1	–	1
2	3	–	3
3	3	–	2
4	1	–	1
5	1	–	1
6	3	2	–
7	1	1	–
8	1	1	–

$$\text{first Zagreb index :} \quad M_1 = \sum_{i=1}^{A} \delta_i^2$$

$$M_1(\mathcal{G}) = 5 \times 1^2 + 3 \times 3^2 = 32$$

$$\mathrm{IFM}_1(\mathcal{F}) = 2^2 + 2 \times 1^2 = 6$$

$$\mathrm{IFM}_1(\mathcal{G}-\mathcal{F}) = 3 \times 1^2 + 2^2 + 3^2 = 16$$

$$\begin{aligned}\mathrm{EFM}_1(\mathcal{F}) &= M_1(\mathcal{G}) - \mathrm{IFM}_1(\mathcal{F}) - \mathrm{IFM}_1(\mathcal{G}-\mathcal{F}) \\ &= 32 - 6 - 16 = 10\end{aligned}$$

➢ **F-ratio test** → regression parameters

FRAU Features (FF)

FRAU (Field characterization for Reaction Analysis and Understanding) features encode information about the reaction field of the atoms in a molecule in the three-dimensional space [Satoh, Itono *et al.*, 1999; Satoh, 2007]. These descriptors are based on steric and electrostatic interactions determined by a pseudoreactant.

To calculate FRAU descriptors, first the frontier surface of each atom is determined by drawing a sphere around each atom with radius equal to or larger than the atomic van der Waals radius. The surface of this sphere is taken as the atomic surface. The part of the atomic surface overlapping with the sphere of another atom is called *interior surface*, and the atomic surface minus the interior surface is called *frontier surface*. Then, a number of points are evenly distributed on the atomic surface.

Only points on the frontier surface are used to evaluate features of the molecule by means of a probe. There are three kinds of FRAU descriptors: the *extent of reaction field* (FF_i^{field}), the *electrostatic feature* ($\mathrm{FF}_i^{electro}$), and the *steric feature* (FF_i^{steric}).

For each *i*th atom in the molecule, the *extent of reaction field* FF_i^{field} is defined as the number of points on the surface frontier of the atom.

To evaluate the *electrostatic feature*, a unit charge is taken as the probe placed at every point on the atomic frontier surface. Then, $\mathrm{FF}_i^{electro}$ (in kcal/mol) is calculated as

$$\mathrm{FF}_i^{electro} = \frac{1}{\mathrm{FF}_i^{field}} \cdot \sum_{k=1}^{\mathrm{FF}_i^{field}} \sum_{j=1}^{A} \frac{331.8417 \cdot q_j}{r_{kj}}$$

where A is the number of atoms in the molecule, q_j the net charge of the *j*th atom, and r_{kj} the geometric distance between the *k*th surface point and the *j*th atom.

Estimation of the *steric feature* FF_i^{steric} is based on the van der Waals interaction between an atom as the probe (e.g., sp^3 carbon atom) placed at the frontier surface points and every atom in the molecule. The FF_i^{steric} (in kcal/mol) of the *i*th atom is calculated by the MM3 force-field equation as

$$\mathrm{FF}_i^{steric} = \frac{1}{\mathrm{FF}_i^{field}} \cdot \sum_{k=1}^{\mathrm{FF}_i^{field}} \sum_{j=1}^{A} \left(\sqrt{\eta_i \cdot \eta_j}\right)$$

$$\times \left[1.84 \times 10^5 \cdot \exp\left(-12.0 \cdot \left(\frac{R_i^{vdw} + R_j^{vdw}}{r_{kj}}\right)^{-1}\right) - 2.25 \cdot \left(\frac{R_i^{vdw} + R_j^{vdw}}{r_{kj}}\right)^6\right]$$

where A is the number of atoms in the molecule, the first summation goes over all the points on the surface frontier of the atom and the second summation on all the atoms of the molecule; η indicates the atomic → *hardness*, R^{vdw} the atomic van der Waals radius, and r_{kj} the geometric distance between the probe atom place at the *k*th surface point and any *j*th atom of the molecule.

Finally, to evaluate similarities and differences among the FRAU features of atoms, these are projected on a 2D map by means of the → *Self-Organizing Map* approach.

Unlike the common → *grid-based QSAR techniques*, FRAU expresses features of molecules having different size and substructures since they do not require alignment of molecules.

FRAU features were applied to classify and predict reagent functions and to distinguish 3D stereochemical environments of atoms.

- **free energy of hydration density tensor** → hydration free energy density
- **free molecular volume** → volume descriptors (⊙ molar volume)
- **free valence index** → quantum-chemical descriptors

■ **Free–Wilson analysis** (≡ *FW Analysis*)

Free–Wilson analysis is a QSAR approach searching for a relationship between a biological response and the presence/absence of substituent groups on a common molecular skeleton [Free and Wilson, 1964; Kubinyi, 1990, 1993b]. The approach, called *de novo approach* when first presented in 1964, is based on the assumption that each substituent gives an additive and constant effect to the biological activity regardless of the other substituents in the rest of the molecule, that is, the substituent effects are considered to be independent of each other. Compounds → *congenericity* is also another basic requirement.

Once a common skeleton for the chemical analogues is defined, regression analysis is performed, considering a number S of substitution sites, denoted as R_s ($s = 1, S$), and, for each site, a number N_s of different substituents. Hydrogen atoms are also considered as substituents if present in a substitution site of some compounds.

The **Free–Wilson descriptors** of the *i*th compound are → *indicator variables*, denoted by $\mathrm{I}_{i,ks}$, where $\mathrm{I}_{i,ks} = 1$ if the *k*th substituent is present in the *s*th site of the *i*th molecule, and $\mathrm{I}_{i,ks} = 0$ otherwise. These descriptors are usually collected in a table called the **Free–Wilson matrix**, denoted as **FW**, where the rows represent the data set molecules and each column represents a substituent in a specific site (Example F2).

The total number of indicator descriptors (i.e., the Free–Wilson matrix columns) is

$$p = \sum_{s=1}^{S} \mathrm{N}_s$$

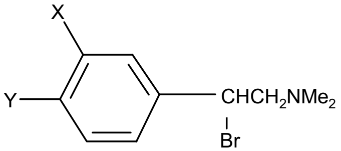
where S is the number of substitution sites and N_s the number of substituents per site.

Given the number of substitution sites S and the number of substituents for each site N_s, the maximum number of structurally diverse compounds that can be studied by a Free–Wilson model is

$$n^{\max} = \prod_{s=1}^{S} \mathrm{N}_s$$

Example F2

Parent molecule and Free–Wilson matrix for 22 derivatives of *N*,*N*-dimethyl-α-bromophenetylamines; X and Y indicate two substitution sites.



In both X and Y sites, hydrogen, fluorine, chlorine, bromine, iodine, and methyl substituents are allowed ($\mathrm{N}_\mathrm{X} = \mathrm{N}_\mathrm{Y} = 6$). The 22 studied derivatives are coded as in the Free–Wilson matrix below. The first row of the Free–Wilson matrix is the H-substituted phenetylamine, used as the reference compound in the Fujita–Ban model and usually excluded in the original Free–Wilson approach.

ID	X	Y	H	m-F	m-Cl	m-Br	m-I	m-Me	H	p-F	p-Cl	p-Br	p-I	p-Me
1	H	H	1	0	0	0	0	0	1	0	0	0	0	0
2	H	F	1	0	0	0	0	0	0	1	0	0	0	0
3	H	Cl	1	0	0	0	0	0	0	0	1	0	0	0
4	H	Br	1	0	0	0	0	0	0	0	0	1	0	0
5	H	I	1	0	0	0	0	0	0	0	0	0	1	0
6	H	Me	1	0	0	0	0	0	0	0	0	0	0	1
7	F	H	0	0	0	0	0	0	1	0	0	0	0	0
8	Cl	H	0	0	0	0	0	0	1	0	0	0	0	0
9	Br	H	0	0	0	1	0	0	1	0	0	0	0	0
10	I	H	0	0	0	0	1	0	1	0	0	0	0	0
11	Me	H	0	0	0	0	0	1	1	0	0	0	0	0
12	Cl	F	0	0	1	0	0	0	0	1	0	0	0	0
13	Br	F	0	0	0	1	0	0	0	1	0	0	0	0
14	Me	F	0	0	0	0	0	1	0	1	0	0	0	0
15	Cl	Cl	0	0	0	0	0	0	0	0	1	0	0	0
16	Br	Cl	0	0	0	1	0	0	0	0	1	0	0	0
17	Me	Cl	0	0	0	0	0	1	0	0	1	0	0	0
18	Cl	Br	0	0	1	0	0	0	0	0	0	1	0	0
19	Br	Br	0	0	0	1	0	0	0	0	0	1	0	0
20	Me	Br	0	0	0	0	0	1	0	0	0	1	0	0
21	Me	Me	0	0	0	0	0	1	0	0	0	0	0	1
22	Br	Me	0	0		1	0	0	0	0	0	0	0	1

$$p = \sum_{s=1}^{2} \mathrm{N}_s = 6 + 6 = 12 \qquad n^{\max} = \prod_{s=1}^{2} \mathrm{N}_s = 6 \times 6 = 36$$

The **Free–Wilson model** (also called **additivity model**) is defined as

$$\hat{y}_i = b_0 + \sum_{s=1}^{S} \sum_{k=1}^{\mathrm{N}_s} b_{ks} \cdot \mathrm{I}_{i,ks}$$

where b_0 is the intercept of the model corresponding to the theoretical biological activity of a compound without any substituent and b_{ks} are the regression coefficients. The biological response y is usually used in the form log(1/C), where C is the concentration achieving a fixed effect. The regression coefficients b_{ks} of the Free– Wilson model give the importance of each *k*th substituent in each *s*th site in increasing/decreasing the response with respect to the unsubstituted compound.

Fujita–Ban analysis is a **modified Free–Wilson analysis** that accounts for the activity contribution of each substituent relatively to the activity of a → *reference compound* [Fujita and Ban, 1971]. Any compound can be chosen as the reference, but usually the H-substituted compound is adopted: the row vector corresponding to the reference compound is characterized by all the descriptor binary values equal to zero. The Free–Wilson matrix in the Fujita–Ban analysis does not contain the descriptors corresponding to the substituents of the reference compound.
The **Fujita–Ban model** is defined as

$$\hat{y}_i = b_0 + \sum_{s=1}^{S} \sum_{k=1}^{\mathrm{N}_s} b_{ks} \cdot \mathrm{I}_{i,ks}$$

which differs from the Free–Wilson model because the intercept b_0 corresponds to the estimated biological activity of the reference compound, that is, $b_0 = \hat{y}_{REF}$, whereas in the Free–Wilson model it corresponds to the theoretical biological activity of a "naked" compound, that is, without any substituent.

The Fujita–Ban model is a linear transformation of the classical Free–Wilson model: indeed, group contributions of the Free–Wilson model can be transformed into Fujita–Ban group contributions by subtracting the group contributions of the corresponding substituents of the reference compound.

The **Cammarata–Yau analysis** is similar to the Fujita–Ban analysis, the only difference being that the intercept is maintained constant and equal to the response of the reference compound [Cammarata and Yau, 1970]; this means that the observed responses are first transformed as

$$y_i' = y_{i-}y_{REF}$$

where y_{REF} is the response for the reference compound.

Accordingly, the **Cammarata–Yau model** is a regression through the origin defined as

$$\hat{y}_i = \sum_{s=1}^{S} \sum_{k=1}^{\mathrm{N}_s} b_{ks} \cdot \mathrm{I}_{i,ks}$$

The **Bocek–Kopecky analysis** is another modified Free–Wilson approach proposed to take into account interaction terms, that is, nonlinear effects [Bocek, Kopecky *et al.*, 1964; Kopecky, Bocek *et al.*, 1965]. The **Bocek–Kopecky model** is defined as

$$\hat{y}_i = b_0 + \sum_{s=1}^{S} \sum_{k=1}^{\mathrm{N}_s} b_{ks} \cdot \mathrm{I}_{i,ks} + \sum_{s=1}^{S-1} \sum_{s'=s+1}^{S} \sum_{k=1}^{\mathrm{N}_s} \sum_{k'=1}^{\mathrm{N}_{s'}} b_{kk',ss'} \cdot \mathrm{I}_{i,ks} \cdot \mathrm{I}_{i,k's'}$$

The Fujita–Ban model having the hydrogen substituted compound as the reference compound is related to the → *Hansch linear model* by the following relationship:

$$b_{ks} \approx \sum_{j=1}^{J} b_j \cdot \phi_{ks,j}$$

where b_j are the Hansch regression coefficients, J the number of considered substituent properties (e.g., lipophilic, electronic, and steric properties), and $\phi_{ks,j}$ the *j*th substituent group constant for the *k*th substituent in the *s*th site. This relationship means that the group contribution b_{ks} in the Fujita–Ban model of the *k*th substituent in the *s*th site is numerically equivalent to the weighted sum of all the → *physico-chemical properties* of that substituent [Singer and Purcell, 1967; Kubinyi and Kehrhahn, 1976; Kubinyi, 1988b].

A great advantage of these approaches is the possibility of a complete → *reversible decoding*, that is, the possibility to interpret by the model *how* and *where* the response is increased/decreased.

The main shortcomings of the Free–Wilson related approaches are that (1) structural variation is necessary in at least two different sites; (2) a relatively large number of variables is necessary to describe a relatively small number of compounds; (3) the models can be used to predict a maximum number of compounds equal to $n^{\max} - n$, where n is the number of compounds effectively used in the model; and (4) predictions of substituents not included in the analysis are usually not reliable.

Related to the Free–Wilson analysis is the → *DARC/PELCO analysis*, which is an extension of the former to the → *hyperstructure* concept [Duperray, Chastrette *et al.*, 1976a].

Moreover, molecular descriptors different from Free–Wilson descriptors were calculated by transformation of the Free–Wilson matrix through → *Fourier analysis* [Holik and Halamek, 2002]. In this case, Fourier analysis is used to change site- and substituent-oriented binary variables into a few real numbers [Holik and Halamek, 2002].

To calculate Fourier coefficients, each row of the Free–Wilson matrix, denoted by $\mathbf{FW}(n, p)$, where n is the number of molecules and p the number of site/substituent indicator variables, is transformed into cosine and sine terms according to the following equation:

$$F_{ix}(k) = \sum_{j=1}^{p} [\mathbf{FW}]_{ij} \cdot \cos(\phi_{kj}) \quad \text{and} \quad F_{iy}(k) = \sum_{j=1}^{p} [\mathbf{FW}]_{ij} \cdot \sin(\phi_{kj})$$

where F_{ix} and F_{iy} are the two Fourier coefficients of the ith molecule, $[\mathbf{FW}]_{ij}$ indicates the elements of the ith row of the Free–Wilson matrix, and ϕ_{kj} is defined as

$$\phi_{kj} = \frac{2 \cdot \pi \cdot k \cdot (j{-}1)}{p} \qquad k = 1, 2, \ldots, L$$

where p is the number of matrix columns and L a user-defined integer parameter. Depending on the size p of the matrix $\mathbf{FW}(n, p)$, a different number of linearly independent Fourier coefficients are obtained: if p is odd, then this number is $(p-1)/2$, if p is even, then there are $(p-2)/2$ independent Fourier coefficients. A data compression is performed if L is chosen to be smaller than the number of linearly independent Fourier coefficients.

Example F3

Calculation of the Fourier coefficients for the second row (compound with $R_1 = H$ and $R_2 = F$) of the data set comprised of 22 *N*,*N*-dimethyl-α-bromo-phenetylamines (Appendix C). The corresponding Free–Wilson matrix ($n = 22$ and $p = 12$) is reported in Example F2.

$$\mathbf{x} = [\,1\;\;0\;\;0\;\;0\;\;0\;\;0\;\;0\;\;1\;\;0\;\;0\;\;0\;\;0\,]$$

For $k = 1$ and $j = 1, 2, \ldots, 12$, the ϕ terms are

$$[0\;\;0.524\;\;1.047\;\;1.571\;\;2.094\;\;2.618\;\;3.142\;\;3.665\;\;4.189\;\;4.712\;\;5.236\;\;5.760]$$

and the first coefficient F_{2x} for $k = 1$ is

$$F_{2x}(1) = 1 \times \cos(0) + 0 \times \cos(0.524) + \cdots + 0 \\ \times \cos(2.618) + 1 \times \cos(3.143) + \cdots + 0 \times \cos(5.760) = 0.134$$

Then k is increased (until $k = 5$) and the other cosine Fourier coefficients are calculated. Finally, the procedure is repeated with sine function. As a result, the 12 original variables are transformed into a vector $\mathbf{f}$ of 10 real-valued variables:

$$\mathbf{f} = [0.134\;1.500\;1.000\;0.500\;1.866\;-0.500\;0.866\;-1.000\;0.866\;-0.500]$$

🕮 [Craig, 1972; Cammarata, 1972; Cammarata and Bustard, 1974; Thomas, Berkoff *et al.*, 1975; Kubinyi, 1976a; Hall and Kier, 1978a; Schaad, Hess Jr. *et al.*, 1981; Duewer, 1990; Liwo, Tarnowska *et al.*, 1992; Franke and Buschauer, 1992; Simmons, Dixson *et al.*, 1992; Franke and Buschauer, 1993; Henrie II, Plummer *et al.*, 1993; Norinder, 1993; Singh, Ojha *et al.*, 1993; De Castro and Reissmann, 1995; Hasegawa, Shigyou *et al.*, 1995; Timofei, Kurunczi *et al.*, 1995; Hatrìk and Zahradnìk, 1996; Hasegawa, Yokoo *et al.*, 1996; Fleischer, Frohberg *et al.*, 2000; Tomić, Nilsson *et al.*, 2000; Shi, Qian *et al.*, 2001; Waisser, 2001; Holik and Halamek, 2002; Pimple, Kelkar *et al.*, 2004; Prabhakar, Gupta *et al.*, 2005; Saxty, Woodhead *et al.*, 2007]

- **Free–Wilson descriptors** → Free–Wilson analysis
- **Free–Wilson matrix** → Free–Wilson analysis
- **Free–Wilson model** → Free–Wilson analysis
- **freezing point** → physico-chemical properties (⊙ melting point)
- **Friedman's lack-of-fit function** → variable selection (⊙ Genetic Function Approximation)
- **frontier orbitals** → quantum-chemical descriptors
- **frontier orbital electron densities** → quantum-chemical descriptors
- **F strain** ≡ *substituent front strain* → steric descriptors
- **fugacity** → physico-chemical properties
- **Fujita–Ban analysis** → Free–Wilson analysis
- **Fujita–Ban model** → Free–Wilson analysis
- **Fujita steric constant** → steric descriptors (⊙ Taft steric constant)
- **Fukui functions** → quantum-chemical descriptors
- **Functional Connectivity FingerPrints** → substructure descriptors (⊙ fingerprints)
- **functional group count** → count descriptors
- **functional group filters** → property filters

■ **functional coordination index** (I_C)

Based on the idea to model the morphology–functionality relationship in organisms, the functional coordination index is defined in terms of the → *molecular surface* area *SA*, the molecular weight MW, and the standard enthalpy of formation H_f as [Torrens, 2003a]

$$I_C = \frac{\mathrm{H}_f}{I_m} = \frac{\mathrm{MW} \cdot \mathrm{H}_f}{SA}$$

where I_m is the **morphologic index** defined as

$$I_m = \frac{SA}{\mathrm{MW}}$$

- **functionality index** → ETA indices
- **FW analysis** ≡ *Free–Wilson analysis*