Comprehensive Descriptors

For Structural and

Statistical Analysis

CODESSA PRO

Reference manual

by
Alan R. Katritzky, Ruslan Petrukhin and Hongfang Yang
(University of Florida)

Mati Karelson
(University of Tartu, Estonia

Table of Contents

Chapter 1 History of the CODESSA project

The roots of the CODESSA project go back to a GROUND/GROUNDSTAT project developed at the Center for Heterocyclic Compounds of the University of Florida in 1992 by under the direction by Dr. Alan R. Katritzky (main developer Dr. Ekaterina Gordeeva). The GROUND was a FORTRAN program which calculates the molecular descriptors and the GROUNDSTAT was a statistical analyser.

The "CODESSA" project got its name in 1993 when Victor Lobanov began to write the Borland C++ software for calculating descriptors and a Windows interface for the various segments. About the same time, Professor Mati Karelson, University of Tartu, Estonia wrote the statistical treatment segment of the software. The work was carried out at the Center for Heterocyclic Compounds, under the direction by Professor Alan R. Katritzky.

The latest version of the CODESSA (Version 2.21) was released in 1996.

In 1999, Ruslan Petrukhin initiated and began to work on project CODESSA PRO at the Center for Heterocyclic Compounds under the direction of Professor Alan R. Katritzky. The first stage has been finished on 2000 and resulted in scalable modules for calculation of descriptors (MDC, the only partial intersection with CODESSA V2.21 code) and for model development (MDA). At that stage the concept of the file storage was developed. The second stage has been finished by development of visual interface module (CVI) and procedure of self-registration for calculation modules. At the end of year of 2001 the size of code exceed 100,000 lines of C++, FORTRAN and assembler code.

The major part of the work on project CODESSA PRO was done by Ruslan Petrukhin, but since beginning many people kindly help to realize the project. Andre Lomaka helped to develop DANN and PLS modules. Prof. Mati Karelson, in addition to supervising the project, helped with documentation. Inna Petrukhina helped with raster graphics and web design. The user's manual was written by Hongfang Yang.

Chapter 2 Background for the development of CODESSA PRO

The wider applicability of the calculations and the stability of the original CODESSA software required further improvement. First of all, the descriptor space was rather limited, as the design for the operating environment could not work efficiently with large memory objects. In some places, the use of a non-standard matrix library resulted in a low performance. In addition, a non-standard object format was restrictive for high performance calculations to store the descriptor matrix. Another burden for the efficient upgrade of the program was the inconvenient way for addition of new descriptors. It was also difficult to convert the original CODESSA code into a client-server methodology (such things must be done at the developmental stage).

Chapter 3 Basic ideas of new version of the CODESSA software package

The basic ideas in the development of the CODESSA PRO software package were based on following principles:

1. Principle of redundancy, which means that all segments, may have multiple realizations;
2. Principle of standardization, i.e. there is no reason to create something that has a standard (well-known, commonly used for such a task, very well tested, free or essentially free) realization;
3. Principle of the highest calculation performance.

Calculations on the CODESSA PRO server parts are made using the BLAS and LAPACK libraries that are standardized and highly optimized methods for working with large matrices. As optimized for most the current processors, they allow the fastest and the most stable results. A free high performance BLAS library for Intel platform was used in programming. All code used for the server portion was written using ANSI standard C++ language with LAPACK and BLAS (part of LAPACK library) calls encapsulated into improved TNT classes. As a result, the addition (or correction) of the codes for calculating descriptors becomes very easy. The realization of these parts of the CODESSSA project has been accomplished by:

1. Using standard ANSI C++(with STL)/Fortran 77 compiler
2. Using the Intel MKL library (LAPACK clone) on the Windows platform and BLAS LAPACK for other for matrix calculations
3. Using improved TNT for connecting LAPACK with C++ code;

The CODESSA PRO software consists of four segments – the molecular descriptor calculator (MDC), the molecular descriptor analyzer (MDA), the molecular descriptor storage (MDS) and the CODESSA visual client (CVI). The main body of CODESSA PRO is the MDS. While the remaining parts can be located on separate machines, they can only interact with the MDS. Interaction with MDS is based on TCP/IP format. Because only small amount of data is usually transferred through the network, it is possible to locate the MDC and MDA routines on machines that have a very limited connection to server, for example, an Internet connection. The CVI that is the user client portion of the CODESSA project, was developed especially for thin connections.

The MDC part of the program retrieves the structures and their geometry from the MDS, calculates molecular descriptors, and returns them to the MDS.

The MDA retrieves molecular descriptors from the MDS, produces (multi)linear regression models, then returns the regression data to the MDS.

The geometry calculation server (GCS) accepts structures with low-level geometry from the MDS, produces the next level geometry, and then stores the structures in the MDS.

The molecule structure predictor (MSP) obtains (multi)linear regression models from the MDS, then searches the entire database to find the molecule that best matches the users conditions. After completing the first task, this segmental routine attempts to predict the geometry that matches best the user conditions using a controllable molecular structure generator.

All server components (MDS, MDC, MDA, GCS and MSP) should operate on almost any high-performance machine and run on a variety of operation systems (principle of redundancy for server). The client portion runs on the Windows machines, and serves as a means to visualize the results of calculations made by the server. The descriptor matrix can be accessed by the standard statistical programs (STATICTICA, SAS, SPSS, etc) (principle of redundancy for client).

Chapter 4 CODESSA PRO Classes of Descriptors

CODESSA PRO Classes of Descriptors

##	Group	Type	Short name	Full name
1	Constitutional	M	N_A	total number of atoms in the molecule
2	Constitutional	A	N_X,$N_{X,r}$	absolute and relative numbers of atoms of certain chemical identity (C, H, O, N, F, etc.) in the molecule
3	Constitutional	M	N_Y,$N_{Y,r}$	absolute and relative numbers of certain chemical groups and functionalities in the molecule
4	Constitutional	M	N_B	total number of bonds in the molecule
5	Constitutional	M	N_S,N_D,N_T, $N_{S,r}$,$N_{D,r}$,$N_{T,r}$	absolute and relative numbers of single, double, triple, aromatic or other bonds in the molecule
6	Constitutional	M	N_R,$N_{R,r}$	total number of rings, number of rings divided by the total number of atoms
7	Constitutional	M	N_{R6},$N_{R6,r}$	total and relative number of 6-atoms aromatic rings
8	Constitutional	M	M, M_r	molecular weight and average atomic weight
9	Topological	M	W	Wiener index
10	Topological	M	χ	Randiċ's molecular connectivity index
11	Topological	M	${}^m\chi$	Randiċ indices of different orders
12	Topological	M	J	Balaban's J index
13	Topological	M	${}^m\chi_v$	Kier and Hall valence connectivity indices
14	Topological	M	${}^m k$	Kier shape indices
15	Topological	M	ϕ	Kier flexibility index
16	Topological	M	${}^k IC$	Mean information content index
17	Topological	M	${}^k SIC$	Structural information content index
18	Topological	M	${}^k CIC$	Complementary information content index
19	Topological	M	${}^k BIC$	Bonding information content index
20	Topological	M	T_n^E	Topological electronic indices
21	Geometrical	M	S_M	Molecular surface area
22	Geometrical	M	S_{SA}	Solvent-accessible molecular surface area

23	Geometrical	M	V_M	Molecular volume
24	Geometrical	M	$V_{M,SE}$	Solvent-excluded molecular volume
25	Geometrical	M	G_p,G_b	Gravitational indexes
26	Geometrical	M	I_X,I_Y,I_Z	Principal moments of inertia of a molecule
27	Geometrical	M	S_{XY},S_{YZ},S_{XZ}	Shadow areas of a molecule
28	Geometrical	M	$S_{XY,r}$, $S_{YZ,r}$, $S_{XZ,r}$	Relative shadow areas of a molecule
29	Electrostatic	A	Q_i	Gasteiger-Marsili empirical atomic partial charges
30	Electrostatic	A	Q_i	Zefirov's empirical atomic partial charges
31	Electrostatic	A	Q_i	Mulliken atomic partial charges
32	Electrostatic	M	Q_{max}, Q_{min}	Minimum (most negative) and maximum (most positive) atomic partial charges
33	Electrostatic	M	P, P', P''	Polarity parameters
34	Electrostatic	M	μ	Dipole moment
35	Electrostatic	M	α	Molecular polarizability
36	Electrostatic	M	β	Molecular hyperpolarizability
37	Electrostatic	M	I_{avg}	Average ionization energy
38	Electrostatic	M	$V_{S,min}$	Minimum electrostatic potential at the molecular surface
39	Electrostatic	M	$V_{S,max}$	Maximum electrostatic potential at the molecular surface
40	Electrostatic	M	Π	Local polarity of molecule
41	Electrostatic	M	δ_{tot}^2	Total variance of the surface electrostatic potential
42	Electrostatic	M	ν	Electrostatic balance parameter
43	CPSA	M	PPSA1	Partial positively charged surface area
44	CPSA	M	PPSA2	Total charge weighted partial positively charged surface area
45	CPSA	M	PPSA3	Atomic charge weighted partial positively charged surface area
46	CPSA	M	PNSA1	Partial negatively charged surface area
47	CPSA	M	PNSA2	Total charge weighted partial negatively charged surface area
48	CPSA	M	PNSA3	Atomic charge weighted partial negatively charged surface area
49	CPSA	M	DPSA1	Difference between partial positively and negatively charged surface areas
51	CPSA	M	DPSA2	Difference between total charge weighted partial positive and negative surface areas

52	CPSA	M	DPSA3	Difference between atomic charge weighted partial positive and negative surface areas
53	CPSA	M	FPSA1	Fractional partial positive surface area
54	CPSA	M	FPSA2	Fractional total charge weighted partial positive surface area
55	CPSA	M	FPSA3	Fractional atomic charge weighted partial positive surface area
56	CPSA	M	FNSA1	Fractional partial negative surface area
57	CPSA	M	FNSA2	Fractional total charge weighted partial negative surface area
58	CPSA	M	FNSA3	Fractional atomic charge weighted partial negative surface area
59	CPSA	M	WPSA1	Surface weighted charged partial positive charged surface area
60	CPSA	M	WPSA2	Surface weighted charged partial positive charged surface area
61	CPSA	M	WPSA3	Surface weighted charged partial positive charged surface area
62	CPSA	M	WNSA1	Surface weighted charged partial negative charged surface area
63	CPSA	M	WNSA2	Surface weighted charged partial negative charged surface area
64	CPSA	M	WNSA3	Surface weighted charged partial negative charged surface area
65	CPSA	M	RPCG	Relative positive charge
66	CPSA	M	RNCG	Relative negative charge
67	CPSA	M	HDSA1	Hydrogen bonding donor ability of the molecule
68	CPSA	M	HDSA2	Area-weighted surface charge of hydrogen bonding donor atoms
69	CPSA	M	HASA1	Hydrogen bonding acceptor ability of the molecule
70	CPSA	M	HASA2	Area-weighted surface charge of hydrogen bonding acceptor atoms
71	CPSA	M	HDCA1	Hydrogen bonding donor ability of the molecule
72	CPSA	M	HDCA2	Area-weighted surface charge of hydrogen bonding donor atoms
73	CPSA	M	HACA1	Hydrogen bonding acceptor ability of the molecule
74	CPSA	M	HACA2	Area-weighted surface charge of hydrogen bonding acceptor atoms

75	CPSA	M	FHDSA1	Fractional hydrogen bonding donor ability of the molecule
76	CPSA	M	FHDSA2	Fractional area-weighted surface charge of hydrogen bonding donor atoms
77	CPSA	M	FHASA1	Fractional hydrogen bonding acceptor ability of the molecule
78	CPSA	M	FHASA2	Fractional area-weighted surface charge of hydrogen bonding acceptor atoms
79	CPSA	M	FHDCA1	Fractional hydrogen bonding donor ability of the molecule
80	CPSA	M	FHDCA2	Fractional area-weighted surface charge of hydrogen bonding donor atoms
81	CPSA	M	FHACA1	Fractional hydrogen bonding acceptor ability of the molecule
82	CPSA	M	FHACA2	Fractional area-weighted surface charge of hydrogen bonding acceptor atoms
83	MO related	M	ε_{HOMO}	Highest occupied molecular orbital (HOMO) energy
84	MO related	M	ε_{LUMO}	Lowest unoccupied molecular orbital (LUMO) energy
85	MO related	M	η	Absolute hardness
86	MO related	M	$\Delta\eta$	Activation hardness
87	MO related	M	E_A	Fukui atomic nucleophilic reactivity index
88	MO related	M	N_A	Fukui atomic electrophilic reactivity index
89	MO related	M	R_A	Fukui atomic one-electron reactivity index
90	MO related	B	P_{AB}	Mulliken bond orders
91	MO related	A	$V_{f,A}$	Free valence
92	Quantum chemical	M	E_{tot}	Total energy of the molecule
93	Quantum chemical	M	E_{el}	Total electronic energy of the molecule
94	Quantum chemical	M	ΔH_0^f	Standard heat of formation
95	Quantum chemical	AT	$E_{ee,A}$	Electron-electron repulsion energy for a given atomic species
96	Quantum chemical	AT	$E_{ne,A}$	Nuclear-electron attraction energy for a given atomic species
97	Quantum chemical	B	$E_{ee,AB}$	Electron-electron repulsion between two given atoms
98	Quantum chemical	B	$E_{ne,AB}$	Nuclear-electron attraction energy between two given atoms

100	Quantum chemical	B	$E_{nn,AB}$	Nuclear repulsion energy between two given atoms
101	Quantum chemical	B	$E_{exc,AB}$	Electronic exchange energy between two given atoms
102	Quantum chemical	BT	$E_{R,AB}$	Resonance energy between given two atomic species
103	Quantum chemical	BT	$E_{C,AB}$	Total electrostatic interaction energy between two given atomic species
104	Quantum chemical	BT	$E_{tot,AB}$	Total interaction energy between two given two atomic species
105	Quantum chemical	M	$E_{ee,tot}$	Total molecular one-center electron-electron repulsion energy
106	Quantum chemical	M	$E_{ne,tot}$	Total molecular one-center electron-nuclear attraction energy
107	Quantum chemical	M	$E_{C,tot}$	Total intramolecular electrostatic interaction energy
108	Quantum chemical	M	K	Electron kinetic energy density
109	Quantum chemical	M	ΔH_{prot}	Energy of protonation
110	Thermodynamic	M	H_V	Vibrational enthalpy of the molecule
111	Thermodynamic	M	H_T	Translational enthalpy of the molecule
112	Thermodynamic	M	S_V	Vibrational entropy of the molecule
113	Thermodynamic	M	S_R	Rotational entropy of the molecule
114	Thermodynamic	M	S_T	Translational entropy of the molecule
115	Thermodynamic	M	C_V	Vibrational heat capacity of the molecule
116	Thermodynamic	M	NAVA	Normal coordinate EigenValues

4.1Constitutional Descriptors

- total number of atoms in the molecule
- absolute and relative numbers of atoms of certain chemical identity (C, H, O, N, F, etc.) in the molecule
- absolute and relative numbers of certain chemical groups and functionalities in the molecule
- total number of bonds in the molecule
- absolute and relative numbers of single, double, triple, aromatic or other bonds in the molecule
- total number of rings, number of rings divided by the total number of atoms

- total and relative number of 6-atoms aromatic rings
- molecular weight and average atomic weight

4.2 Topological Descriptors

4.2.1 Wiener index

Wiener index

Definition:

$$W = \frac{1}{2}\sum_{(i,j)}^{N_{SA}} d_{ij}$$

d_{ij} - the number of bonds in the shortest path connecting the pair of atoms ***i*** and ***j***

N_{SA} - the number of non-hydrogen atom in the molecule

Reference:

H. Wiener, *J. Am. Chem. Soc.*, **1947,** 69, *17*.

4.2.2 Randic's molecular connectivity index

Randic’s molecular connectivity index

Definition:

$$\chi = \sum_{edges\ ij} \left(D_i D_j\right)^{-1/2}$$

D_i and D_j - the edge degrees (atom connectivities) of the molecular graph.

4.2.3 Randic indices of different orders

Randiċ indices of different orders

$$ {}^{m}\chi = \sum_{path} \left(D_i D_j \ldots D_k \right)^{-1/2} $$

References:

1. M. Randiċ, *J. Am. Chem. Soc.*, **1975,** *97*, 6609.
2. L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, J. Wiley & Sons, New York, 1986.

4.2.4 Balaban's *J* index

Balaban's *J* index

Definition:

$$ J = \frac{q}{\mu + 1} \sum_{edges\, ij} \left(S_i S_j \right)^{-1/2} $$

q - number of edges in the molecular graph

m = (*q* – *n* + *1*) - the cyclomatic number of the molecular graph

n – number of atoms in the molecular graph

S_i - distance sums calculated as the sums over the rows or columns of the topological distance matrix of the molecule, ***D***.

References:

1. A. T. Balaban, *Chem. Phys. Lett.*, **1981,** *89*, *399*.
2. A. T. Balaban, *Pure and Appl. Chem.*, **1983,** *55*, *199*.

4.2.5 Kier and Hall valence connectivity indices

Kier and Hall valence connectivity indices

Definition:

$${}^{m}\chi^{v} = \sum_{i=1}^{N_S} \prod_{k=1}^{m+1} \left(\frac{1}{\delta_k^v} \right)^{1/2}$$

$\delta_k^v = \dfrac{(Z_k^v - H_k)}{(Z_k - Z_k^v - 1)}$ - valence connectivity for the *k*-th atom in the molecular graph

Z_k - the total number of electrons in the *k*-th atom

Z_k^v - the number of valence electrons in the *k*-th atom

H_k - the number of hydrogen atoms directly attached to the *k*th non-hydrogen atom

m = 0 - atomic valence connectivity indices

m = 1 - one bond path valence connectivity indices

m = 2 - two bond fragment valence connectivity indices

m = 3 three contiguous bond fragment valence connectivity indices etc.

References:

1. L. B. Kier, L. H. Hall, *Eur. J. Med. Chem.*, **1977**, *12*, *307*.
2. L. B. Kier, L. H. Hall, *J. Pharm. Sci.*, **1981**, *70*, *583*.
3. L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, J. Wiley & Sons, New York, 1986.

4.2.6 Kier shape indices

Kier shape indices

Definition:

$$^{1}\kappa = (N_{SA} + \alpha)(N_{SA} + \alpha - 1)^{2}(^{1}P + \alpha)^{2}$$

$$^{2}\kappa = (N_{SA} + \alpha - 1)(N_{SA} + \alpha - 2)^{2}(^{2}P + \alpha)^{2}$$

$$^{3}\kappa = (N_{S}A + \alpha - 1)(N_{S}A + \alpha - 3)^{2}(^{3}P + \alpha)^{2} \quad \text{if } N_{SA} \text{ is odd}$$

$$^{3}\kappa = (N_{SA} + \alpha - 3)(N_{SA} + \alpha - 2)^{2}(^{3}P + \alpha)^{2} \quad \text{if } N_{SA} \text{ is even}$$

N_{SA} - the number of non-hydrogen atom in the molecule

^{n}P - the number of paths of the length *n*in the molecular graph

$$\alpha = \frac{r_i}{r_{Ci}} - 1$$

r_i - atomic radius of a given atom

r_{Ci} - atomic radius of the carbon atom in the *sp3* hybridization state

References:

1. L. B. Kier, *Quant. Struct.-Act. Relat.*, **1985***, 4*, *109*.
2. L. B. Kier, in: *Computational Chemical Graph Theory*, D. H. Rouvray (Ed.), Nova Science Publishers, New York 1990.

4.2.7 Kier flexibility index

Kier flexibility index

Definition:

$$\Phi = \frac{(^{1}\kappa\, ^{2}\kappa)}{N_{SA}}$$

$^{1}\boldsymbol{k}$ and $^{2}\boldsymbol{k}$ - Kier shape indices

N_{SA} - the number of non-hydrogen atom in the molecule

Reference:

1. L. B. Kier, in: *Computational Chemical Graph Theory*, D. H. Rouvray (Ed.), Nova Science Publishers, New York 1990.

4.2.8 Mean information content index

Mean information content index

Definition:

$$^{k}IC = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

n_i - number of atoms in the *i*th class

n - the total number of atoms in the molecule

k - number of atomic layers in the coordination sphere around a given atom that are accounted for

Reference:

1. L. B. Kier, *J. Pharm. Sci.*, **1980**, *69*, 807.

4.2.9 Structural information content index

Structural information content index

Definition:

$$^{k}SIC = {}^{k}IC / \log_2 n$$

$$^{k}IC = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

n_i - number of atoms in the *i*th class

n - the total number of atoms in the molecule

k – number of atomic layers in the coordination sphere around a given atom that are accounted for

Reference:

1. S.C. Basak, D. K. Harriss, V. R. Magnuson, *J. Pharm. Sci.*, **1984**, *73*, 429.

4.2.10 Complementary information content index

Complementary information content index

Definition:

$$^{k}CIC = \log_2 n - {}^{k}IC$$

$$^{k}IC = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

n_i - number of atoms in the *i*th class

n - the total number of atoms in the molecule

k - number of atomic layers in the coordination sphere around a given atom that are accounted for

Reference:

1. S.C. Basak, D. K. Harriss, V. R. Magnuson, *J. Pharm. Sci.*, **1984,** *73,* 429.

4.2.11 Bonding information content index

Bonding information content index

Definition:

$$^{k}BIC = {}^{k}IC / \log_2 q$$

$$^{k}IC = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

n_i - number of atoms in the *i*th class

n - the total number of atoms in the molecule

k - number of atomic layers in the coordination sphere around a given atom that are accounted for

q - number of edges in the molecular graph

Reference:

1. S.C. Basak, D. K. Harriss, V. R. Magnuson, *J. Pharm. Sci.*, **1984,** *73*, 429 ()

4.2.12 Topological electronic indices

Topological electronic indices

Definition:

$$T_1^E = \sum_{(i<j)}^{N_{SA}} \frac{|q_i - q_j|}{r_{ij}^2}$$

$$T_2^E = \sum_{(i<j)}^{N_b} \frac{|q_i - q_j|}{r_{ij}^2}$$

$\boldsymbol{q_i}$- partial charge on the *i*-th atom

r_{ij} – distance between *i*-th and *j*-th atoms

N_{SA} – number of non-hydrogen atom in the molecule

N_b – number of bonds between non-hydrogen atom in the molecule

Reference:

1. K. Osmialowski, J. Halkiewicz, R. Kaliszan, *J. Chromatogr.*, **1986**, *63*, *361*.

4.3 Geometrical Descriptors

4.3.1 Molecular surface area

Molecular surface area

Definition:

$$S_M = \sum_i S_{VW}^{(i)} - S_{ov}$$

$S_{VW}^{(i)}$ - van der Waals area of the *i*-th constituent atom of a molecule

S_{ov} – van der Waals area of atoms inside overlapping atomic envelopes

Reference:

1. M. Karelson, *Molecular Descriptors in QSAR/QSPR*, J. Wiley & Sons, New York, 2000.

4.3.2 Solvent-accessible molecular surface area

Solvent-accessible molecular surface area

Definition:

$$S_{SA} = A_+ + A_s + A_-$$

A_+ - convex areas of a molecule

A_s - saddle areas of a molecule

A_- - concave areas of a molecule

Reference:

M. L. Connolly, *J. Appl. Crystallogr.,* **1983***, 16*, 548-558.

4.3.3 Molecular volume

Molecular volume

Definition:

$$V_M = \sum_i V_{VW}^{(i)} - V_{ov}$$

$V_{VW}^{(i)}$ - van der Waals volume of the *i*-th constituent atom of a molecule

V_{ov} – volume of overlapping van der Waals atomic envelopes

Reference:

1. F. M. Richards, *Annu. Rev. Biophys. Bioeng.*, **1977,** *6*, 151-176

4.3.4 Solvent-excluded molecular volume

Solvent-excluded molecular volume

Definition:

$$V_{mol(SE)} = V_p + \sum V_+ + \sum V_s + \sum V_- + V_{ac} + V_{nc}$$

V_p - volume of internal polyhedron

$\sum V_+$ - volume pieces between the center of an atom and the convex face of the solvent-accessible surface

$\sum V_s$ - volume of saddle pieces

$\sum V_-$ - volume of concave pieces

V_{ac}, V_{nc} - cusp volume pieces

Reference:

1. M. L. Connolly, *J. Am. Chem. Soc.*, **1985**, *107*, 1118-1124

4.3.5 Gravitational indexes

Gravitational indexes

Definition:

$$G_p = \sum_{i<j}^{N_a} \frac{m_i m_j}{r_{ij}^2}$$

$$G_b = \sum_{i<j}^{N_b} \frac{m_i m_j}{r_{ij}^2}$$

m_i, m_j - atomic masses of atoms i and j

r_{ij} - interatomic distance of atoms i and j

N_a - number of atoms in the molecule

N_b - number of chemical bonds in the molecule

Reference:

1. A. R. Katritzky, L. Mu, V. S. Lobanov, M. Karelson, *J. Phys. Chem.*, **1996**, *100*, 10400-10407.

4.3.6 Principal moments of inertia of a molecule

Principal moments of inertia of a molecule

Definition:

$$I_k = \sum_i m_i r_{ik}^2$$

m_i - atomic weights of constituent atoms of a molecule

r_{ik} - distance of the *i*-th atomic nucleus from the *k*-th main rotational axes ($k = X$, Y or Z)

Reference:

1. *Handbook of Chemistry and Physics*, CRC Press, Cleveland OH, 1974, p. F-112.

4.3.7 Shadow areas of a molecule

Shadow areas of a molecule

Definition:

$$S_k = \frac{1}{2} \oint_{(C)} \left(\nu d\rho - \rho d\nu \right)$$

C – contour of the projection of the molecule on the plane defined by two principal axes of the molecule ($k = XY, XZ$ or YZ)

ν - x or y

p - *y* or *z*

Reference:

1. R. H. Rohrbaugh, P. C. Jurs, *Anal. Chim. Acta*, **1987,** *199*, *99*.

4.3.8 Relative shadow areas of a molecule

Relative shadow areas of a molecule

Definition:

$$S_k^r = \frac{\oint\limits_{(C)} (\nu d\rho - \rho d\nu)}{S^{(k)}}$$

C – contour of the projection of the molecule on the plane defined by two principal axes of the molecule (*k* = *XY, XZ* or *YZ* plane)

$\boldsymbol{\nu}$- *x* or *y*

p - *y* or *z*

$S^{(k)} = X \bullet Y; X \bullet Z$ or $Y \bullet Z$

Reference:

1. M. Karelson, *Molecular Descriptors in QSAR/QSPR*, J. Wiley & Sons, New York, 2000.

4.4 Electrostatic Descriptors

4.4.1 Gasteiger-Marsili empirical atomic partial charges

Gasteiger-Marsili empirical atomic partial charges

Definition:

$$Q_i = \sum_{\alpha} q_i^{<\alpha>}$$

$q_i^{<\alpha>} = \left(\frac{1}{2}\right)^{\alpha} \sum_{\nu \in i} \left[\sum_{\mu \in j} \frac{\chi_{j\mu}^{<\alpha>} - \chi_{i\nu}^{<\alpha>}}{\chi_{i\nu}^{+}} + \sum_{\lambda \in k} \frac{\chi_{k\lambda}^{<\alpha>} - \chi_{i\nu}^{<\alpha>}}{\chi_{k\lambda}^{+}} \right]$ - the contribution to the atomic charge on the *a-th* step of iteration of charge

$\chi_{i\nu} = a_{i\nu} + b_{i\nu} Q_i + c_{i\nu} Q_i^2$ - electronegativity of *n*–th orbital on *i*-th atom

$$a_{i\nu} = \frac{I_{i\nu}^0 + E_{i\nu}^0}{2}$$

$$b_{i\nu} = \frac{I_{i\nu}^0 + E_{i\nu}^{+} - E_{i\nu}^0}{4}$$

$$b_{i\nu} = \frac{I_{i\nu}^{+} - I_{i\nu}^0 + E_{i\nu}^{+} - E_{i\nu}^0}{4}$$

$I_{i\nu}^0$, $I_{i\nu}^{+}$, $E_{i\nu}^0$, and $E_{i\nu}^{+}$ - the ionization potentials and electron affinities of the neutral atom (superscript *0*) and of the positive ion (superscript +), respectively.

References:

1. J. Gasteiger, M. Marsili, *Tetrahedron Lett.*, **1978,** 3181
2. J. Gasteiger, M. Marsili, *Tetrahedron*, **1980,** *36*, 3219-3228

4.4.2 Zefirov's empirical atomic partial charges

Zefirov's empirical atomic partial charges

Definition:

$$Q_i = f(\chi_i)$$

χ_i – atomic electronegativities

$$\chi_i = \left(\chi_i^0 \prod_{k=1}^{n} \chi_k \right)^{1/(n+1)}$$

χ_i^0 - electronegativities of isolated atoms

n – atoms in the first coordination sphere of a given atom

References:

1. N. S. Zefirov, M. A. Kirpichenok, F. F. Izmailov, M. I. Trofimov., *Dokl. Akad. Nauk SSSR*, **1987,** *296*, 883.
2. M. A. Kirpichenok, N. S. Zefirov, *Zh. Org. Khim.*, **1987,** *23*, 4 .

4.4.3 Mulliken atomic partial charges

Mulliken atomic partial charges

Definition:

$$Q_A = Z_A - \left(\sum_{k \in A} P_{kk} + \frac{1}{2} \sum_{l \neq k} P_{kl} + \frac{1}{2} \sum_{l \neq k} P_{lk} \right) = Z_A - \sum P_{kl}$$

Z_A - atomic nuclear charge

P_{kl} - atomic population matrix elements

Reference:

1. R. S. Mulliken, *J. Chem. Phys.*,**1955** *23*, 1833-1840.
2. I. G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.

4.4.4 Minimum (most negative) and maximum (most positive) atomic partial charges

Minimum (most negative) and maximum (most positive) atomic partial charges

Definition:

$$Q_{min} = min\ (Q^{-})$$

$$Q_{max} = max\ (Q^{+})$$

Q^{-}- negative atomic partial charges

Q^{+} - positive atomic partial charges

Reference:

1. I. G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.

4.4.5 Polarity parameters

Polarity parameters

Definitions:

$$P = Q_{max} - Q_{min}$$

$$P' = \frac{Q_{max} - Q_{min}}{R_{mm}}$$

$$P'' = \frac{Q_{max} - Q_{min}}{R_{mm}^{2}}$$

Q_{max} - the most positive atomic partial charge in the molecule

Q_{min} - the most negative atomic partial charge in the molecule

R_{mm} - distance between the most positive and the most negative atomic partial charges in the molecule

Reference:

1. K. Osmialowski, J. Halkiewicz, A. Radecki, R. Kaliszan, *J. Chromatogr.*, **1985,** *346*, 53.

4.4.6 Dipole moment

Dipole moment

Definition:

$$\mu = -\sum_{i=1}^{occ} \int_{(V)} \phi_i \hat{r} \phi_i dv + \sum_{a=1}^{M} Z_a \vec{R}_a$$

ϕ_i - molecular orbitals

$\hat{r}$ - electron position operator

Z_a - a-th atomic nuclear charge

$\vec{R}_a$ - position vector of a-th atomic nucleus

Reference:

1. P. W. Atkins, *Quanta*, Oxford University Press, Oxford, 1991.

4.4.7 Molecular polarizability, α

Molecular polarizability, α

Definition:

$$\mu' = \mu + \alpha \boldsymbol{E} + \frac{1}{2} \beta \boldsymbol{E}^2 + \ldots$$

μ- permanent dipole moment of the molecule

μ' - induced dipole moment of the molecule

$\boldsymbol{E}$ - external electric field

Reference:

1. P. W. Atkins, *Quanta*, Oxford University Press, Oxford, 1991.

4.4.8 Molecular hyperpolarizability, β

Molecular hyperpolarizability, β

Definition:

$$\mu' = \mu + \alpha\boldsymbol{E} + \frac{1}{2}\beta\boldsymbol{E}^2 + \ldots$$

μ- permanent dipole moment of the molecule

μ' - induced dipole moment of the molecule

$\boldsymbol{E}$ - external electric field

Reference:

1. P. W. Atkins, *Quanta*, Oxford University Press, Oxford, 1991.

4.4.9 Average ionization energy

Average ionization energy

Definition:

$$\bar{I}(\mathbf{r}) = \frac{\sum_i \rho_i(\mathbf{r})|\varepsilon_i|}{\rho(\mathbf{r})}$$

p(r) - electron density of the *i*th molecular orbital at the point **r**

ε_i - *i*th molecular orbital energy

Reference:

1. T. Brinck, J. S. Murray, P. Politzer, *Int. J. Quant. Chem.*, **1993,** *48*, 73-88

4.4.10 Minimum electrostatic potential at the molecular surface

Minimum electrostatic potential at the molecular surface

Definition:

$$V_{S,\boldsymbol{min}} = \boldsymbol{min}\left[V(\mathbf{r})\right] = \boldsymbol{min}\left[\sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} - \int \frac{\rho(\mathbf{r}')d\mathbf{r}'}{|\mathbf{r}' - \mathbf{r}|}\right]$$

Z_A - charge on atomic nucleus A at point $\mathbf{R}_A$

p(r') - total electron density of the molecule

Reference:

1. P. Politzer, J. S. Murray, *Rev. Comput. Chem.*, **1991**, *2*.

4.4.11 Maximum electrostatic potential at the molecular surface

Maximum electrostatic potential at the molecular surface

Definition:

$$V_{S,\boldsymbol{max}} = \boldsymbol{max}\left[V(\mathbf{r})\right] = \boldsymbol{max}\left[\sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} - \int \frac{\rho(\mathbf{r}')d\mathbf{r}'}{|\mathbf{r}' - \mathbf{r}|}\right]$$

Z_A - charge on atomic nucleus A at point $\mathbf{R}_A$

p(r') - total electron density of the molecule

Reference:

1. J. S. Murray, P. Lane, T. Brinck, P. Politzer, *J. Phys. Chem.*, **1990,** *94*, 844

4.4.12 Local polarity of molecule

Local polarity of molecule

Definition:

$$\Pi = \frac{1}{A}\int_S \left|V(\mathbf{r}) - \overline{V}_S\right| dS \approx \frac{1}{n}\sum_{i=1}^{n}\left|V_i(\mathbf{r}) - \overline{V}_S\right|$$

A - molecular surface area

$\overline{V}_S$ - average value of the electrostatic potential in the molecule

V (r) - electrostatic potential in the molecule

n – number of integration points

Reference:

1. T. Brinck, J. S. Murray, P. Politzer, Mol. Phys., **1992,** *76*, 609.

4.4.13 Total variance of the surface electrostatic potential

Total variance of the surface electrostatic potential

Definition:

$$\sigma_{tot}^2 = \sigma_+^2 - \sigma_-^2 = \frac{1}{m}\sum_{i=1}^{m}\left[V^+(\mathbf{r}_i) - \overline{V}_S^+\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[V^-(\mathbf{r}_i) - \overline{V}_S^-\right]^2$$

$\overline{V}_S^+$ - average value of the positive electrostatic potential in the molecule

$\overline{V}_S^-$ - average value of the negative electrostatic potential in the molecule

$V^+(r_i)$ - positive electrostatic potential in the molecule

$V^-(r_i)$ - negative electrostatic potential in the molecule

m,n - number of integration points

Reference:

1. P. Politzer, P. Lane, J. S. Murray, T. Brinck, *J. Phys. Chem.*, **1992,** *96*, 7938.

4.4.14 Electrostatic balance parameter

Electrostatic balance parameter

Definition:

$$\nu = \frac{\sigma_+^2 \sigma_-^2}{\left[\sigma_{tot}^2\right]^2}$$

σ_+^2 - variance of the positive electrostatic potential in the molecule

σ_-^2 - variance of the negative electrostatic potential in the molecule

σ_{tot}^2 - total variance of the electrostatic potential in the molecule

Reference:

1. J. S. Murray, P. Lane, T. Brinck, P. Politzer, *J. Phys. Chem.*, **1993,** *97*, 5144

4.5 CPSA Descriptors

4.5.1Partial positively charged surface area

Partial positively charged surface area PPCSA

Definition:

$$PPSA1 = \sum_{A} S_A \qquad A \in \{\delta_A > 0\}$$

S_A- positively charged solvent-accessible atomic surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***,* *62*, 2323 .; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306 .

4.5.2 Total charge weighted partial positively charged surface area

Total charge weighted partial positively charged surface area

Definition:

$$PPSA2 = \sum_{A} q_A \cdot \sum_{A} S_A \qquad A \in \{\delta_A > 0\}$$

S_A - positively charged solvent-accessible atomic surface area

q_A - atomic partial charge

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323 .; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306 .

4.5.3 Atomic charge weighted partial positively charged surface area

Atomic charge weighted partial positively charged surface area

Definition:

$$PPSA3 = \sum_{A} q_A \cdot S_A \qquad A \in \{\delta_A > 0\}$$

S_A - positively charged solvent-accessible atomic surface area

q_A - atomic partial charge

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323 ; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.4 Partial negatively charged surface area

Partial negatively charged surface area *PNCSA*

Definition:

$$PNSA1 = \sum_{A} S_A \qquad A \in \{\delta_A < 0\}$$

S_A - negatively charged solvent-accessible atomic surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.5 Total charge weighted partial negatively charged surface area

Total charge weighted partial negatively charged surface area

Definition:

$$PNSA2 = \sum_A q_A \cdot \sum_A S_A \qquad A \in \{\delta_A < 0\}$$

S_A - negatively charged solvent-accessible atomic surface area

q_A - atomic partial charge

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.6 Atomic charge weighted partial negatively charged surface area

Atomic charge weighted partial negatively charged surface area

Definition:

$$PNSA3 = \sum_A q_A \cdot S_A \qquad A \in \{\delta_A < 0\}$$

S_A - negatively charged solvent-accessible atomic surface area

q_A - atomic partial charge

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.7 Difference between partial positively and negatively charged surface areas

Difference between partial positively and negatively charged surface areas

Definition:

$$\mathrm{DPSA1} = \mathrm{PPSA1} - \mathrm{PNSA1}$$

PPSA1 - positively charged solvent-accessible molecular surface area

PNSA1 - negatively charged solvent-accessible molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.8 Difference between total charge weighted partial positive and negative surface areas

Difference between total charge weighted partial positive and negative surface areas

Definition:

$$\mathbf{DPSA2 = PPSA2 - PNSA2}$$

PPSA2 – total charge weighted partial positively charged molecular surface area

PNSA2 - total charge weighted partial negatively charged molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.9 Difference between atomic charge weighted partial positive and negative surface areas

Difference between atomic charge weighted partial positive and negative surface areas

Definition:

$$\mathbf{DPSA2 = PPSA2 - PNSA2}$$

PPSA2 – total charge weighted partial positively charged molecular surface area

PNSA2 - total charge weighted partial negatively charged molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.10 Fractional partial positive surface area

Fractional partial positive surface area

Definition:

$$FPSA1 = \frac{PPSA1}{TMSA}$$

PPSA1– partial positively charged molecular surface area

TMSA- total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.11 Fractional total charge weighted partial positive surface area

Fractional total charge weighted partial positive surface area

Definition:

$$FPSA2 = \frac{PPSA2}{TMSA}$$

PPSA2 – total charge weighted partial positively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.12 Fractional atomic charge weighted partial positive surface area

Fractional atomic charge weighted partial positive surface area

Definition:

$$FPSA3 = \frac{PPSA3}{TMSA}$$

PPSA3– total charge weighted partial positively charged molecular surface area

TMSA -total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.13 Fractional partial negative surface area

Fractional partial negative surface area

Definition:

$$FNSA1 = \frac{PNSA1}{TMSA}$$

PNSA1 – partial negatively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.14 Fractional total charge weighted partial negative surface area

Fractional total charge weighted partial negative surface area

Definition:

$$FNSA2 = \frac{PNSA2}{TMSA}$$

PNSA2 – total charge weighted partial negatively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.15 Fractional atomic charge weighted partial negative surface area

Fractional atomic charge weighted partial negative surface area

Definition:

$$FNSA3 = \frac{PNSA3}{TMSA}$$

PNSA3 – total charge weighted partial negatively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.16 Surface weighted charged partial positive charged surface area WPSA1

Surface weighted charged partial positive charged surface area WPSA1

Definition:

$$WPSA1 = \frac{PPSA1 \cdot TMSA}{1000}$$

PPSA1 – partial positively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.17 Surface weighted charged partial positive charged surface area WPSA2

Surface weighted charged partial positive charged surface area WPSA2

Definition:

$$WPSA2 = \frac{PPSA2 \cdot TMSA}{1000}$$

PPSA2 – total charge weighted partial positively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***, 62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.18 Surface weighted charged partial positive charged surface area WPSA3

Surface weighted charged partial positive charged surface area WPSA3

Definition:

$$WPSA3 = \frac{PPSA3}{TMSA}$$

PPSA3 – total charge weighted partial positively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***, 62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.19 Surface weighted charged partial negative charged surface area WNSA1

Surface weighted charged partial negative charged surface area WNSA1

Definition:

$$WNSA1 = \frac{PNSA1 \cdot TMSA}{1000}$$

PNSA1 – partial negatively charged molecular surface area

TMSA -total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.20 Surface weighted charged partial negative charged surface area WNSA2

Surface weighted charged partial negative charged surface area WNSA2

Definition:

$$WNSA2 = \frac{PPSA2 \cdot TMSA}{1000}$$

PNSA2 – total charge weighted partial negatively charged molecular surface area

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** 62, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** *32*, 306.

4.5.21 Surface weighted charged partial negative charged surface area WNSA3

Surface weighted charged partial negative charged surface area WNSA3

Definition:

$$WNSA3 = \frac{PNSA3}{TMSA}$$

PNSA3 – total charge weighted partial negatively charged molecular surface area

TMSA -total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992** 32, 306.

4.5.22 Relative positive charge

Relative positive charge

Definition:

$$RPCG = \frac{\delta^{+}_{\max}}{\sum_{A} \delta_{A}} \qquad A \in \{\delta_{A} > 0\}$$

$\delta^{+}_{\max}$ - maximum atomic positive charge in the molecule

δ_A - positive atomic charge in the molecule

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.23 Relative negative charge

Relative negative charge

Definition:

$$RNCG = \frac{\delta^{-}_{\max}}{\sum_{A} \delta_A} \qquad A \in \{\delta_A > 0\}$$

$\delta^{-}_{\max}$ - maximum atomic negative charge in the molecule

δ_A - negative atomic charge in the molecule

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.5.24 Hydrogen bonding donor ability of the molecule *HDSA1*

Hydrogen bonding donor ability of the molecule *HDSA1*

Definition:

$$HDSA1 = \sum_{D} s_D \qquad D \in H_{H\text{-}donor}$$

s_D - solvent-accessible surface area of H-bonding donor H atoms

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990**, *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992**, *32*, 306.

4.5.25 Area-weighted surface charge of hydrogen bonding donor atoms *HDSA2*

Area-weighted surface charge of hydrogen bonding donor atoms *HDSA2*

Definition:

$$HDSA2 = \sum_{D} \frac{q_D \sqrt{s_D}}{\sqrt{S_{tot}}} \qquad D \in H_{H\text{-donor}}$$

s_D- solvent-accessible surface area of H-bonding donor H atoms

q_D -partial charge on H-bonding donor H atoms

S_{tot} - total solvent-accessible molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.26 Hydrogen bonding acceptor ability of the molecule *HASA1*

Hydrogen bonding acceptor ability of the molecule *HASA1*

Definition:

$$HASA1 = \sum_{A} s_A \qquad A \in X_{H-acceptor}$$

s_D - solvent-accessible surface area of H-bonding acceptor atoms

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***,* *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***,* *32*, 306

4.5.27 Area-weighted surface charge of hydrogen bonding acceptor atoms *HASA2*

Area-weighted surface charge of hydrogen bonding acceptor atoms *HASA2*

Definition:

$$HASA2 = \sum_{A} \frac{q_A \sqrt{S_A}}{\sqrt{S_{tot}}} \qquad A \in X_{H\text{-}acceptor}$$

S_A -solvent-accessible surface area of H-bonding acceptor atoms

q_A - partial charge on H-bonding acceptor atoms

S_{tot} - total solvent-accessible molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***,* *32*, 306.

4.5.28 Hydrogen bonding donor ability of the molecule *HDCA1*

Hydrogen bonding donor ability of the molecule *HDCA1*

Definition:

$$HDCA1 = \sum_{D} s_D \qquad D \in H_{H\text{-}donor}$$

s_D - solvent-accessible surface area of H-bonding donor H atoms, selected by threshold charge

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.29 Area-weighted surface charge of hydrogen bonding donor atoms *HDCA2*

Area-weighted surface charge of hydrogen bonding donor atoms *HDCA2*

Definition:

$$HDCA2 = \sum_{D} \frac{q_D \sqrt{S_D}}{\sqrt{S_{tot}}} \qquad D \in H_{H\text{-}donor}$$

S_D - solvent-accessible surface area of H-bonding donor H atoms, selected by threshold charge

q_D - partial charge on H-bonding donor H atoms, selected by threshold charge

S_{tot} -total solvent-accessible molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.5.30 Hydrogen bonding acceptor ability of the molecule *HACA1*

Hydrogen bonding acceptor ability of the molecule *HACA1*

Definition:

$$HACA1 = \sum_{A} s_A \qquad A \in X_{H-acceptor}$$

S_D - solvent-accessible surface area of H-bonding acceptor atoms, selected by threshold charge

Reference:

D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***, 62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.5.31 Area-weighted surface charge of hydrogen bonding acceptor atoms *HACA2*

Area-weighted surface charge of hydrogen bonding acceptor atoms *HACA2*

Definition:

$$HACA2 = \sum_A \frac{q_A \sqrt{s_A}}{\sqrt{S_{tot}}} \qquad A \in X_{H\text{-}acceptor}$$

s_A - solvent-accessible surface area of H-bonding acceptor atoms, selected by threshold charge

q_A - partial charge on H-bonding acceptor atoms, selected by threshold charge

S_{tot} - total solvent-accessible molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***, 62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.32 Fractional hydrogen bonding donor ability of the molecule *FHDSA1*

Fractional hydrogen bonding donor ability of the molecule *FHDSA1*

Definition:

$$FHDSA1 = \frac{HDSA1}{TMSA}$$

HDSA1 - hydrogen bonding donor ability

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***,62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.33 Fractional area-weighted surface charge of hydrogen bonding donor atoms *FHDSA2*

Fractional area-weighted surface charge of hydrogen bonding donor atoms *FHDSA2*

Definition:

$$FHDSA2 = \frac{HDSA2}{TMSA}$$

HDSA2 - area-weighted surface charge on hydrogen bonding donor atoms

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.5.34 Fractional hydrogen bonding acceptor ability of the molecule *FHASA1*

Fractional hydrogen bonding acceptor ability of the molecule *FHASA1*

Definition:

$$FHASA1 = \frac{HASA1}{TMSA}$$

HASA1 - hydrogen bonding acceptor ability

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990**, *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.35 Fractional area-weighted surface charge of hydrogen bonding acceptor atoms *FHASA2*

Fractional area-weighted surface charge of hydrogen bonding acceptor atoms *FHASA2*

Definition:

$$FHASA2 = \frac{HASA2}{TMSA}$$

HASA2 - area-weighted surface charge on hydrogen bonding acceptor atoms

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***, 62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.5.36 Fractional hydrogen bonding donor ability of the molecule *FHDCA1*

Fractional hydrogen bonding donor ability of the molecule *FHDSA1*

Definition:

$$FHDSA1 = \frac{HDSA1}{TMSA}$$

HDSA1 - hydrogen bonding donor ability

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990***,62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.37 Fractional area-weighted surface charge of hydrogen bonding donor atoms *FHDCA2*

Fractional area-weighted surface charge of hydrogen bonding donor atoms *FHDCA2*

Definition:

$$FHACA2 = \frac{HACA2}{TMSA}$$

HDCA2 - area-weighted surface charge on hydrogen bonding donor atoms, selected by threshold charge

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992,** *32*, 306.

4.5.38 Fractional hydrogen bonding acceptor ability of the molecule *FHACA1*

Fractional hydrogen bonding acceptor ability of the molecule *FHASA1*

Definition:

$$FHASA1 = \frac{HASA1}{TMSA}$$

$HASA1$ – hydrogen bonding acceptor ability

$TMSA$ – total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990**, *62*, 2323
2. D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992**, *32*, 306

4.5.39 Fractional area-weighted surface charge of hydrogen bonding acceptor atoms *FHACA2*

Fractional area-weighted surface charge of hydrogen bonding acceptor atoms *FHACA2*

Definition:

$$FHACA2 = \frac{HACA2}{TMSA}$$

HACA2 - area-weighted surface charge on hydrogen bonding acceptor atoms, selected by threshold charge

TMSA - total molecular surface area

Reference:

1. D.T. Stanton, P.C. Jurs, Anal. Chem., **1990,** *62*, 2323; D.T. Stanton, L.M. Egolf, P.C. Jurs, M.G. Hicks, J. Chem. Inf. Comput. Sci., **1992***, 32*, 306.

4.6 MO Related Descriptors

4.6.1 Highest occupied molecular orbital (HOMO) energy

Highest occupied molecular orbital (HOMO) energy

Definition:

$$\varepsilon_{HOMO} = \langle \phi_{HOMO} | \hat{\mathbf{F}} | \phi_{HOMO} \rangle$$

ϕ_{HOMO} - highest occupied molecular orbital

$\hat{\mathbf{F}}$ - Fock operator

Reference:

1. I. G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.
2. B.W. Clare, *Theoret. Chim. Acta*, **1994,** *87*, 415-430.

4.6.2 Lowest unoccupied molecular orbital (LUMO) energy

Lowest unoccupied molecular orbital (LUMO) energy

Definition:

$$\varepsilon_{LUMO} = \langle \phi_{LUMO} | \hat{\mathbf{F}} | \phi_{LUMO} \rangle$$

ϕ_{LUMO}- lowest unoccupied molecular orbital

$\hat{\mathbf{F}}$ - Fock operator

References:

1. I.G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.
2. B.W. Clare, *Theoret. Chim. Acta*, **1994,** *87*, 415-430.

4.6.3 Absolute hardness

Absolute hardness

Definition:

$$\eta = \left(\varepsilon_{LUMO} - \varepsilon_{HOMO}\right)/2$$

ε_{LUMO} - lowest unoccupied molecular orbital energy

ε_{HOMO} - highest occupied molecular orbital energy

References:

1. Z. Zhou, R. G. Parr, *J. Am. Chem. Soc.*, **1990,** *112*, 5720.
2. R. G. Pearson, *J. Org. Chem.*, **1989,** 54, 1423.

4.6.4 Activation hardness

Activation hardness

Definition:

$$\Delta\eta = \eta_R - \eta_T$$

η_R - absolute hardness of the reactant

η_T - absolute hardness of the transition state

References:

1. Z. Zhou, R.G. Parr, *J. Am. Chem. Soc.*, **1990,** *112*, 5720.
2. R.G. Pearson, *J. Org. Chem.*, **1989,** *54*, 1423.

4.6.5 Fukui atomic nucleophilic reactivity index

Fukui atomic nucleophilic reactivity index

Definition:

$$N_A = \sum_{i \in A} c_{iHOMO}^2 / (1 - \varepsilon_{HOMO})$$

or simplified

$$N'_A = \sum_{i \in A} c_{iHOMO}^2$$

ε_{HOMO} - highest occupied molecular orbital energy

c_{iHOMO} - highest occupied molecular orbital MO coefficients

Reference:

1. R. Franke, *Theoretical Drug Design Methods*. Elsevier, Amsterdam, 1984

4.6.6 Fukui atomic electrophilic reactivity index

Fukui atomic electrophilic reactivity index

Definition:

$$E_A = \sum_{j \in A} c_{jLUMO}^2 / (\varepsilon_{LUMO} + 10)$$

or simplified

$$E'_A = \sum_{j \in A} c^2_{jLUMO}$$

ε_{LUMO} - lowest unoccupied molecular orbital energy

c_{jLUMO} - lowest unoccupied molecular orbital MO coefficients

Reference:

1. R. Franke, *Theoretical Drug Design Methods*. Elsevier, Amsterdam, 1984

4.6.7 Fukui atomic one-electron reactivity index

Fukui atomic one-electron reactivity index

Definition:

$$R_A = \sum_{i \in A} \sum_{j \in A} c_{iHOMO} c_{jLUMO} / (\varepsilon_{LUMO} - \varepsilon_{HOMO})$$

or simplified

$$R'_A = \sum_{i \in A} \sum_{j \in A} c_{iHOMO} c_{jLUMO}$$

c_{iHOMO} - highest occupied molecular orbital MO coefficients

c_{jLUMO} - lowest unoccupied molecular orbital MO coefficients

ε_{LUMO} - lowest unoccupied molecular orbital energy

ε_{HOMO} - highest occupied molecular orbital energy

Reference:

1. R. Franke, *Theoretical Drug Design Methods*. Elsevier, Amsterdam, 1984.

4.6.8 Mulliken bond orders

Mulliken bond orders

Definition:

$$P_{AB} = \sum_{i=1}^{occ} \sum_{\mu \in A} \sum_{\nu \in B} n_i c_{i\mu} c_{j\nu}$$

n_i- the occupation number of the *i*-th MO

$c_{i\mu}, c_{j\nu}$ - MO coefficients for atomic orbitals *m* and*n*

References:

1. I.G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.
2. A.B. Sannigrahi, *Adv. Quant. Chem.*, **1992,** 23, 301-351.

4.6.9 Free valence

Free valence

Definition:

$$V_{fA} = V_{max} - P_A$$

V_{max} - maximum valence of atom *A*

P_A - total electronic population on atom *A*

References:

1. I.G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976.
2. A.B. Sanniraghi, *Adv. Quant. Chem.*, **1992,** *23*, 301-351.

4.7 Quantum Chemical Descriptors

4.7.1 Total energy of the molecule

Total energy of the molecule

Definition:

$$E_{tot} = E_{el} + \sum_{A \neq B} Z_A Z_B / R_{AB}$$

E_{el} - total electronic energy of the molecule

Z_A, Z_B - nuclear charges of atoms *A* and *B*

R_{AB} - distance between nuclei *A* and *B*

Reference:

1. I. G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976
2. M. Bodor, Z. Gabanyi, C.-K. Wong, *J. Am. Chem. Soc.*, **1989,** 111, 3783

4.7.2 Total electronic energy of the molecule

Total electronic energy of the molecule

Definition:

$$Eel= 2Tr(\mathbf{RF}) - Tr(\mathbf{RG})$$

R - first order density matrix

F - matrix representation of the Hartree-Fock operator

G - matrix representation of the electron repulsion energy

Reference:

1. I. G. Csizmadia, *Theory and Practice of MO Calculations on Organic Molecules*, Elsevier, Amsterdam, 1976

4.7.3 Standard heat of formation

Standard heat of formation

Definition:

$$\Delta H_f^0 = H_f - \sum_a H_f^a$$

H_f - quantum-chemically calculated total energy of the molecule

H_f^a - quantum-chemically calculated energies of isolated atoms, *a*

Reference:

1. P. W. Atkins, *Physical Chemistry*, 3rd Edition, Oxford University Press, Oxford, 1988

4.7.4 Electron-electron repulsion energy for a given atomic species

Electron-electron repulsion energy for a given atomic species

Definition:

$$E_{ee}(A) = \sum_{B \neq A} \sum_{\mu,\nu \in A} \sum_{\lambda,\sigma \in B} P_{\mu\nu} P_{\lambda\sigma} \langle \mu\nu | \lambda\sigma \rangle$$

A – given atomic species

B – other atoms

$P\mu\nu$ $P\lambda\delta$ - density matrix elements over atomic basis $\{\mu\nu\lambda\sigma\}$

$\langle \mu\nu | \lambda\sigma \rangle$ - electron repulsion integrals on atomic basis $\{\mu\nu\lambda\sigma\}$

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980

4.7.5 Nuclear-electron attraction energy for a given atomic species

Nuclear-electron attraction energy for a given atomic species

Definition:

$$E_{ne}(A) = \sum_{B} \sum_{\mu,\nu \in A} P_{\mu\nu} \langle \mu | \frac{Z_B}{R_{iB}} | \nu \rangle$$

A - given atomic species

B - other atoms

$P_{\mu\nu}$ - density matrix elements over atomic basis {μν}

Z_B - charge of atomic nucleus, *B*

R_{iB} - distance between the electron and atomic nucleus, *B*

$\langle \mu | \frac{Z_B}{R_{iB}} | \nu \rangle$ - electron-nuclear attraction integrals on atomic basis {μν}

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980

4.7.6 Electron-electron repulsion between two given atoms

Electron-electron repulsion between two given atoms

Definition:

$$E_{ee}(AB) = \sum_{\mu,\nu\in A}\sum_{\lambda,\sigma\in B} P_{\mu\nu}P_{\lambda\sigma}\langle\mu\nu|\lambda\sigma\rangle$$

A – given atomic species

B – another atomic species

$P_{\mu\nu}$, $P_{\lambda\delta}$ - density matrix elements over atomic basis $\{\mu\nu\lambda\sigma\}$

$\langle\mu\nu|\lambda\sigma\rangle$ - electron repulsion integrals on atomic basis $\{\mu\nu\lambda\sigma\}$

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980

4.7.7 Nuclear-electron attraction energy between two given atoms

Nuclear-electron attraction energy between two givenatoms

Definition:

$$E_{ne}(AB) = \sum_{B}\sum_{\mu,\nu\in A} P_{\mu\nu}\langle\mu|\frac{Z_B}{R_{iB}}|\nu\rangle$$

A – given atomic species

B – another atomic species

$P_{\mu\nu}$- density matrix elements over atomic basis $\{\mu\nu\}$

Z_B - charge of atomic nucleus, *B*

R_{iB} - distance between the electron and atomic nucleus, *B*

$\langle\mu|\frac{Z_B}{R_{iB}}|\nu\rangle$ - electron-nuclear attraction integrals on atomic basis $\{\mu\nu\}$

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980

4.7.8 Nuclear repulsion energy between two given atoms

Nuclear repulsion energy between two given atoms

Definition:

$$E_{nn}(AB) = \frac{Z_A Z_B}{R_{AB}}$$

A – given atomic species

B – another atomic species

Z_A - charge of atomic nucleus, *A*

Z_B - charge of atomic nucleus, *B*

R_{iB} - distance between the atomic nuclei, *A* and *B*

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.9 Electronic exchange energy between two given atoms

Electronic exchange energy between two given atoms

Definition:

$$E_{exc}(AB) = \sum_{\mu,\nu\in A}\sum_{\lambda,\sigma\in B} P_{\mu\lambda}P_{\nu\sigma}\left\langle \mu\lambda \middle| \nu\sigma \right\rangle$$

A – given atomic species

B – another atomic species

$\boldsymbol{P}_{\mu\nu}$□□, $\boldsymbol{P}_{\lambda\sigma}$, - density matrix elements over atomic basis $\{\mu\nu\lambda\sigma\}$

$\left\langle \mu\nu \middle| \lambda\sigma \right\rangle$ - electron repulsion integrals on atomic basis $\{\mu\nu\lambda\sigma\}$

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980

4.7.10 Resonance energy between given two atomic species

Resonance energy between given two atomic species

Definition:

$$E_R(AB) = \sum_{\mu\in A}\sum_{\nu\in B} P_{\mu\nu}\beta_{\mu\nu}$$

A – given atomic species

B – another atomic species

$P_{\mu\nu}$ - density matrix elements over atomic basis $\{\mu\nu\}$

$\beta_{\mu\nu}$□- resonance integrals on atomic basis $\{\mu\nu\}$

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.11 Total electrostatic interaction energy between two given atomic species

Total electrostatic interaction energy between two given atomic species

Definition:

$$E_C(AB) = E_{ee}(AB) + E_{ne}(AB) + E_{nn}(AB)$$

A – given atomic species

B – another atomic species

$\boldsymbol{E_{ee}(AB)}$- electronic repulsion energy between two atomic species

$\boldsymbol{E_{ne}(AB)}$- electron-nuclear attraction energy between two atomic species

$\boldsymbol{E_{nn}(AB)}$- nuclear repulsion energy between two atomic species

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.12 Total interaction energy between two given two atomic species

Total interaction energy between two given two atomic species

Definition:

$$E_{tot}(AB) = E_C(AB) + E_{exc}(AB)$$

A – given atomic species

B – another atomic species

$E_C(AB)$- electrostatic interaction energy between two atomic species

$E_{exc}(AB)$- electronic exchange energy between two atomic species

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.13 Total molecular one-center electron-electron repulsion energy

Total molecular one-center electron-electron repulsion energy

Definition:

$$E_{ee}(tot) = \sum_A E_{ee}(A)$$

A - given atomic species

$E_{ee}(A)$ - electron-electron repulsion energy for atom *A*

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.14 Total molecular one-center electron-nuclear attraction energy

Total molecular one-center electron-nuclear attraction energy

Definition:

$$E_{ne}(tot) = \sum_A E_{ne}(A)$$

A – given atomic species

$E_{ne}(A)$ - electron-nuclear attraction energy for atom *A*

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.15 Total intramolecular electrostatic interaction energy

Total intramolecular electrostatic interaction energy

Definition:

$$E_C(tot) = \frac{1}{2}\sum_A E_C(A)$$

A - given atomic species

$E_C(A)$ - electrostatic energy for atom *A*

Reference:

1. E. Clementi, *Computational Aspects of Large Chemical Systems*, Springer Verlag, New York, 1980.

4.7.16 Electron kinetic energy density

Electron kinetic energy density

Definition:

$$K = -\frac{N}{4}\int\left(\Psi^*\nabla^2\Psi + \Psi\nabla^2\Psi^*\right)d\vec{r}'$$

$$G = \frac{N}{2}\int\nabla\Psi^*\nabla\Psi d\vec{r}'$$

N - number of electrons in the molecule

Ψ - electronic wave function of the molecule

Reference:

1. C. M. Breneman, M. Martinov, in: *Molecular Electrostatic Potentials: Concepts and Applications*, Theoretical and Computational Chemistry, Volume 3, J. S. Murray, K. Sen (Eds.), Elsevier Science B.V., Amsterdam, 1996.

4.7.17 Energy of protonation

Energy of protonation

Definition:

$$\Delta H_{prot} = E_{MH^+} - E_{MH} - E_{H^+}$$

E_{MH} – quantum-chemically calculated energy of neutral molecular species

$E_{MH}{}^{+}$ – quantum-chemically calculated energy of protonated species

$E_{H}{}^{+}$ – energy of proton in the given reference system

Reference:

1. G. Trapani, A. Carotti, M. Franco, A. Latrofa, G. Genchi, G. Liso, G. *Eur. J. Med. Chem.*, **1993,** 28, 13

4.8 Thermodynamic Descriptors

4.8.1 Vibrational enthalpy of the molecule

Vibrational enthalpy of the molecule

Definition:

$$H_{vib} = \frac{1}{2}\sum_{j=1}^{a} h\nu_j + \frac{h\nu_j \exp\left(-h\nu_j/2kT\right)}{1-\exp\left(-h\nu_j/2kT\right)}$$

n_j - frequencies of normal vibrations in the molecule

h - Planck's constant

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.
2. A.I. Akhiezer, S.V. Peltminskii, *Methods of Statistical Physics,* Pergamon Press, Oxford, 1981.

4.8.2 Translational enthalpy of the molecule

Translational enthalpy of the molecule

Definition:

$$H_{tr} = \int_{-\infty}^{\infty} \frac{p^2}{2m} e^{-\frac{p^2}{2mkT}} dp$$

p - momentum of the molecule

m - mass of the molecule

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.
2. A.I. Akhiezer, S.V. Peltminskii, *Methods of Statistical Physics,* Pergamon Press, Oxford, 1981.

4.8.3 Vibrational entropy of the molecule

Vibrational entropy of the molecule

Definition:

$$S_{vib} = \sum_{j=1}^{\alpha} \left\{ \frac{h\nu_j \exp\left(-h\nu_j/2kT\right)}{kT\left[1-\exp\left(-h\nu_j/2kT\right)\right]} - \ln\left[1-\exp\left(-h\nu_j/2kT\right)\right] \right\}$$

n_j - frequencies of normal vibrations in the molecule

h - Planck's constant

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.
2. A.I. Akhiezer, S.V. Peltminskii, *Methods of Statistical Physics,* Pergamon Press, Oxford, 1981.

4.8.4 Rotational entropy of the molecule

Rotational entropy of the molecule

Definition:

$$S_{rot} = Nk\ln\left[\frac{\pi^{1/2}}{\sigma}\prod_{j=1}^{3}\left(\frac{8\pi^2 I_j kT}{h^2}\right)^{1/2}\right]$$

I_j - principal moments of inertia of the molecule

s - symmetry number of the molecule

h - Planck's constant

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.

4.8.5 Translational entropy of the molecule

Translational entropy of the molecule

Definition:

$$S_{tr} = \mathbf{ln}\left(\frac{2\pi mkT}{h^2}\right)^{1/2} \frac{Ve^{5/2}}{N_A}$$

V - volume of the system

N_A - Avogadro's number

m - mass of the molecule

h - Planck's constant

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.
2. A.I. Akhiezer, S.V. Peltminskii, *Methods of Statistical Physics,* Pergamon Press, Oxford, 1981.

4.8.6 Vibrational heat capacity of the molecule

Vibrational heat capacity of the molecule

Definition:

$$c_{v,vib} = k\sum_{j=1}^{\alpha}\left(\frac{h\nu_j}{kT}\right)^2 \frac{\exp\left(-h\nu_j/2kT\right)}{1-\exp\left(-h\nu_j/2kT\right))}$$

n_j - frequencies of normal vibrations in the molecule

h - Planck's constant

k - Boltzmann's constant

T - absolute temperature (K)

References:

1. D.A. McQuarrie, *Statistical Thermodynamics*, Harper & Row Publishers, New York, 1973.
2. A.I. Akhiezer, S.V. Peltminskii, *Methods of Statistical Physics,* Pergamon Press, Oxford, 1981.

4.8.7 Normal coordinate EigenValues (EVA)

Normal coordinate EigenValues (EVA)

Definition:

$$EVA_x = \sum_{i=1}^{3N-6}\frac{1}{\sigma\sqrt{2\pi}}\,exp\left[-\frac{(x-\nu_i)^2}{2\sigma^2}\right]$$

n_j - frequencies of normal vibrations in the molecule

x - sampling point on frequency scale

s- fixed standard deviation for all Gaussian functions characterizing the shape of the vibrational peak

References:

1. T.W. Heritage, A.M. Ferguson, D.B. Turner, P. Willett, in: *3D QSAR in Drug Design*, Volume 2, H. Kubinyi, G. Folkers, Y.C. Martin (Eds.), Kluwer/Escom, Dordrecht, The Netherlands, 1998.

Chapter 5 Methods

The methods for QSRP/QSAR analyses are essentially statistical methods

5.1Multilinear Regression

The basic method for QSPR analysis is essentially the solution of a multilinear regression problem. This can be expressed compactly and conveniently using matrix notation.[1, 2, 3] Suppose that there are *n* property values in **Y** and *n* associated calculated values for each *k* molecular descriptor in **X** columns. Then Y_i, X_{ik}, and e_i can represent the *i*th value of the **Y** variable (property), the *i*th value of each of the **X** descriptors, and the *i*th unknown residual value, respectively. Collecting these terms into matrices we have:

$$\mathbf{Y}=\begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X}=\begin{bmatrix} 1 & X_{11} & \cdots & \cdots & \cdots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & \cdots & \cdots & X_{nk} \end{bmatrix}, \quad \mathbf{e}=\begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ e_n \end{bmatrix}$$

The multiple regression model in matrix notation then can be expressed as

$$\mathbf{Y}=\mathbf{X}\mathbf{b}+\mathbf{e}$$

where **b** is a column vector of coefficients (b_1 is for the intercept) and *k* is the number unknown regression coefficients for the descriptors. We recall that the goal of multiple regression is to minimize the sum of the squared residuals:

$$\min_{b} \|\mathbf{e}\|_2$$

Regression coefficients that satisfy this criterion are found by solving the system of linear equations (multiplying both sides by **X**' from left)

$$\mathbf{X}'\mathbf{Y}=\mathbf{X}'\mathbf{X}\mathbf{b}$$

When the **X**variables are linearly independent (an **X'X** matrix which is of full rank), there is a unique solution to the system of linear equations. One of the ways for solving the system above is to premultiply both sides of the matrix formula for the normal equations by the inverse matrix **X'X** to give

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The other way is to solve directly the system above using LS (underdetermined, n < k) or QR factorization for the overdetermined (n > k) system. This method is more general and does not require time-consuming matrix inversion. Singular value decomposition methods can also be used, but usually such methods are significantly more time-consuming and only advantageous when a strong linear dependence exists that would diminish quality of models.

The third way to solve the problem of linear dependency of variables (determinant of the **X'X** matrix is above zero) is by general matrix inversion, but this is usually outside the sphere of QSPR.

A fundamental principle of least squares methods, the multiple linear regression in particular, is that variance of the dependent variable can be partitioned (divided into parts) according to the source. Suppose that a dependent variable (property) is regressed on one or more descriptors and, for convenience, the dependent variable is scaled so that its mean is 0. Next, a basic least squares identity is calculated in which the total sum of squared values on the dependent variable equals the sum of squared predicted values plus the sum of squared residual values. Stated more generally,

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

where the term on the left is the total sum of squared deviations of the observed values on the dependent variable from the dependent variable mean, and the terms on the right are:

(i) the sum of the squared deviations of the predicted values for the dependent variable from the dependent variable mean and

(ii) the sum of the squared deviations of the observed values on the dependent variable from the predicted values, that is, the sum of the squared residuals.

Stated yet another way,

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Note that the SS_{Total} is always the same for any particular data set, but SS_{Model} and the SS_{Error} vary with the regression equation. Assuming again that the dependent variable is scaled so that its mean is 0, the $SS_{Model}$and SS_{Error} can be computed using

$$SS_{Model} = \mathbf{b'\,X'\,Y'}$$

$$SS_{Error} = \mathbf{Y'\,Y - b'\,X'Y}$$

Assuming that *X'X*is full-rank,

$$r^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

$$s^2 = \frac{SS_{Error}}{n-k-1}$$

$$F(k, n-k-1) = \frac{SS_{Model}}{ks^2}$$

where $r^2$is squared correlation coefficient which is the measure of the quality of model fitnessto the property, s^2 is an unbiased estimate of the residual or error variance, and *F* is Fisher criteria of (*k, n* - *k* - 1) degrees of freedom. If **X'X**is not full rank, *rank(***X'X***)* + 1is substituted for *k*.

References:

1. http://www.statsoft.com/textbook/stathome.html
2. Darlington, R. B. *Regression and linear models*. New York: McGraw-Hill, 1990.
3. Neter, J.; Wasserman, W.; Kutner, M. H. *Applied linear regression models* (2nd ed.). Homewood, IL: Irwin, 1989.

5.2 Selection of descriptors

A rigorously correct solution for descriptor selection requires a full search procedure of the discrete descriptor space. Unfortunately, combinatorial explosion does not allow the application of a full search procedure to real tasks. For example, if we search for a 5-parameter correlation on a space of 1000 descriptors (numbers are realistic for a typical search), we would have to test over $8*10^{12}$ correlations for their ability to match some criterion (usually the squared correlation coefficient). Modern machines have achieved sufficient productivity to calculate one correlation each 0.0001-0.0002 seconds using highly optimized linear algebra libraries (CODESSA PRO uses an Intel MKL – LAPACK [3] clone), and highly optimized low level code. Even with this high level of optimization, the time required for the solution of the aforementioned task, using a full search procedure, is $8*10^{12}*0.0001$ seconds (about 26 years).

Because it is impossible to solve typical tasks in a reasonable amount of time using full search, methods for simplification have been developed. Such methods of descriptor selection can be categorized as either deterministic or stochastic.

Throughout years of research, many algorithms for non-full searches were developed. The best known deterministic algorithms [1] are forward entry/backward removal of effects (in our case – descriptors). The methods of forward and backward stepwise searches combine the entering and removal of the effects at each step. Each of the methods mentioned above have many limitations, [2] the majority of which are concerned with the absence of a consistent set of the correlations (models) which represent the upper segment of the search space. The best-subset methods (proposed in this work) are the next alternative to the full-search procedure, bur possess such limitations.

Two methods for reducing full search procedure were utilized by our group: [4] the heuristic method and the best multi-linear regression.

The heuristic method for descriptor selection proceeds with a pre-selection of descriptors by sequentially eliminating descriptors that do not match any of the following criteria: (i) Fisher *F*-criteria greater than 1.0; (ii) R^2 value less than a value defined at the start; (iii) Student's *t*-criterion less than a defined value; (iv) duplicate descriptors having a higher squared intercorrelation coefficient than a predetermined level (retaining the descriptor with higher R^2 with reference to the property). The descriptors that remain are then listed in decreasing order of correlation coefficients when used in global search for 2-parameter correlations. Each significant 2-parameter correlation by *F*-criteria is recursively expanded to an *n*-parameter correlation till the normalized *F*-criteria remains greater than the startup value. The best *N* correlations by R^2, as well as by *F*-criterion, are saved.

The best multilinear regression method is based on the (i) selection of the orthogonal descriptor pairs, (ii) extension of the correlation (saved on the previous step) with the addition of new descriptors until the *F*-criteria becomes less than that of the best 2-parameter correlation. The best *N* correlations (by R^2) are saved.

Both methods successfully solve the initial selection problem by reducing the number of pairs of descriptors in the "starting set". The major limitations of these methods are the pairwise selection on the first step and the low consistence of the presentation of the upper (according to the selected criteria) segment of the search (N in both cases is 400) due to the small size of the correlation selection.

A review of the stochastic methods, the genetic algorithm (GA) in particular, was recently published by Leardi.[5] The same author also published the first application of the genetic algorithm, [6] and in a review, mentioned the two major disadvantages of using GA: the repeatability of the optimization and the unpredictable coverage of the search space. The repeatability problem is a failure of all stochastic methods by definition and is therefore unacceptable for an industrial strength system. The possibility of chance correlations is a disadvantage to all methods of effects (descriptors) selection and it is surely the most important factor, which limits generalized and extensive use of GA. [7]

References:

1. Darlington, R. B. *Regression and linear models*. New York: McGraw-Hill, 1990.

2. Neter, J.; Wasserman, W.; Kutner, M. H. *Applied linear regression models* (2nd ed.). Homewood, IL: Irwin, 1989.
3. Anderson, E; Bai, Z.; Bischof, C.; Demmel, J.; Dongarra, J.; Ducroz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK: A portable linear algebra library for high-performance computers.* Computer Science Dept. Technical Report CS-90-105, University of Tennessee: Knoxville, TN, 1990
4. Katritzky, A.R.; Lobanov V.; Karelson, M. CODESSA Reference Manual. University of Florida, Gainesville, 1996.
5. Leardi, R. Genetic algorithm in chemometrics and chemistry: a review. *J. Chemometrics* **2001**, *15*, 559-569.
6. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithm as a strategy for feature selection, *J. Chemometrics* **1992**, *6*, 267-281.
7. Leardi, R.; Gonzalez, A. L. Generic algorithm applied to feature selection on PLS regression: how and when to use them. *Chemometr. Intell. Lab.* **1998**, *41*, 195-207.

5.3 Multivariate methods

The principal component analysis (PCA) is generally described as an ordination technique for describing the variation in a multivariate data set.[1, 2, 3] The first axis (the first principal component, or PC1) describes the maximum variation in the whole data set; the second describes the maximum variance remaining, and so forth, with each axis orthogonal to the preceding axis. Principal components are eigenvectors of a covariance **X'X** or correlation **X'Y** matrix. The number of principal components that can be extracted will typically exceed the maximum of the number of **Y**and **X**variables.

The principal component analysis and factor analysis are based on the separation of the original matrix **X** into two matrixes: factor scores matrix **T** and loading matrix **Q**. In matrix form:

$$\mathbf{X} = \mathbf{TQ}$$

The columns in a factor score matrix are linear independent. Usually, the columns in **X** and **Y** matrix are centered (by subtracting their means) and scaled (by dividing by their standard deviations). Suppose we have a data set with response variables **Y** (in matrix form) and a large number of predictor variables **X** (in matrix form), some of which are highly correlated. A regression, using factor extraction for this type of data, computes the factor score matrix

$$\mathbf{T} = \mathbf{XW}$$

for an appropriate weight matrix **W**, and then considers the linear regression model

$$\mathbf{Y} = \mathbf{TQ} + \mathbf{E}$$

where **Q** is a matrix of regression coefficients (loadings) for **T**, and **E** is an error (noise) term. Once the loadings **Q** are computed, the above regression model is equivalent to

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\mathbf{B} = \mathbf{WQ}$$

which can be used as a predictive regression model.

The factor scores and loadings can be obtained in many different ways. NIPALS algorithm was developed in 1923, [4] later modified in 1966, [5] and SIMPLS algorithm [6] resulted from work by de Jong in 1993. Singular value decomposition is another commonly used method for calculating scores and loading.[7]

Principal components regression (PCR) and partial least squares (PLS) regression differ in the methods used for extracting factor scores.[1] PLR produces the weight matrix **W** reflecting the covariance structure between the predictor variables, while PLS regression produces the weight matrix **W** reflecting the covariance structure between the predictor and response variables. In PLSregression, prediction functions are represented by factors extracted from the **Y'XX'Y** matrix.

For establishing the model, PLS regression produces a weight matrix **W** for **X** such that **T=XW**, i.e., the columns of **W** are weight vectors for the **X** columns producing the corresponding factor score matrix **T**. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of **Y** on **T** are then performed to produce **Q**, the loadings for **Y**(or weights for **Y**) such that **Y**=**TQ**+**E**. Once **Q** is computed, we have **Y**=**XB**+**E**, where **B**=**WQ**, and the prediction model is complete.

One additional matrix which is necessary for a complete description of partial least squares regression procedures is the factor loading matrix **P** which gives a factor model **X**=**TP**+**F**, where **F** is the unexplained part of the **X** scores.

References:

1. http://www.statsoft.com/textbook/stathome.html
2. Manly, B. F. J. *Multivariate Statistical Methods. A Primer*. London-NY: Chapman and Hall, 1986.
3. Nilsson, J. *Multiway calibration in 3D QSAR: applications to dopamine receptor ligands*; Groningen: University Library Groningen, 1998, Online Resource.
4. Fisher, R.; MacKensie, W. Journal of Agricaltural Science **1923**, *13*, 311-320.
5. Wold, H., In: Research papers in Statistics, ed. David, F., NY: Wiley & Sons, 1966, pp.411-444.

6. de Jong, S. SIMPLS: An Alternative Approach to Partial Least Squares Regression, *Chemometrics and Intelligent Laboratory Systems* **1993**, *18*, 251-263
7. Mandel, J. American Statistician **1982**, *36*, 15-24 .

Chapter 6 Test

Most QSPR models are useful, but occasionally models are produced that not at all reliable.[1] The problems of reliability can validly be classified as (i) overfitting and (ii) models by chance. The latter problem can only be solved by subjective human judgment based on the justifiability of any assessment of the uncertainty of a particular prediction.[2, 3]

The problem of overfitting is one of the more common problems in the development of the any kind of model and QSPR is no exception. The problem is usually solved using objective validation criterions. The most common method of validation in chemometrics is crossvalidation. [4]

Referenses:

1. Eriksson, L.; Johansson, E.; Muller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered *J. Chemometrics* **2000**, *14*, 599-616.
2. Stone M.; Jonathan P. Statistical Thinking and Technique for QSAR and related studies. 1. Genaral Theory *J. Chemometrics* **1993**, *7*, 455-475.
3. Stone M.; Jonathan P. Statistical Thinking and Technique for QSAR and related studies. 2. Specific Methods *J. Chemometrics* **1994**, *6*, 1-20.
4. Xu Q.-S.; Liang Y.-Z. Monte Carlo cross validation *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1-11.

6.1 External validation set

The easiest method of the correlation testing is use of an external validation set. [1] In this method, the correlation is used for predict a property value for a chemical structure that was not used in the creation of the correlation; some test statistics are calculated for the external dataset; the difference between the test statistics in the training and validation datasets is a measure of the reliability of the correlation. The widely used measure is the prediction error sum of squares (*PRESS*) and is defined as

$$PRESS = \sum_i \left(y_{e,i} - y_{p,i}\right)^2$$

where $y_{e,i}$ are experimental values of the property and $y_{p,i}$ are predicted values for external validation test.

Often the *RMSPE* criterion is prefered:

$$RMSPE = \sqrt{\frac{PRESS}{n}}$$

because it gives error on a ‘per compound’ basis. [1]

The method is a particular case of the leave-many-out cross-validation method.

References:

1. Stone M.; Jonathan P. Statistical Thinking and Technique for QSAR and related studies. 1. Genaral Theory *J. Chemometrics* **1993**, *7*, 455-475.

6.2 Leave-One-Out crossvalidation

The simplest, and a commonly used method of crossvalidation in chemometrics is the “leave-one-out” method. The idea behind this method is to predict the property value for a compound from the data set, which is in turn predicted from the regression equation calculated from the data for all other compounds. For evaluation, predicted values can be used for *PRESS*, *RMSPE*, and squared correlation coefficient criteria (r^2_{cv}).

The method tends to include unnecessary components in the model, and has been provided [2] to be asymptotically incorrect. Furthermore, the method does not work well for data with strong clusterization, [1] and underestimates the true predictive error. [3]

References:

1. Eriksson, L.; Johansson, E.; Muller, M.; Wold, S. On the selection of the training set in environmental QSAR analysis when compounds are clustered *J. Chemometrics* **2000**, *14*, 599-616.
2. Stone M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion *J. R. Stat. Soc., B* **1977**, *38*, 44-47.
3. Martens, H.A.; Dardenne, P. Validation and verification of regression in small data sets **1998**, *44*, 99-121.

6.3 Leave-Many-Out crossvalidation

The “leave-many-out” crossvalidation method was firstly described in 1975. [3] Later the asymptotic consistence of the method was proved. [2] Because of combinatorial complexity of the calculation resulted in low productivity, some simplifications were developed for the method. The evaluation of the results of “leave-many-out” crossvalidation can be done using Monte Carlo approach. [1]

References:

1. Xu Q.-S.; Liang Y.-Z. Monte Carlo cross validation *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1-11.
2. Stone M. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion *J. R. Stat. Soc., B* **1977**, *38*, 44-47.
3. Geisser, S. The Predictive Sample Reuse Method with Application *J. Amer. Stat.Ass.* **1975**, *70*, 320-328.

6.4 Randomization test

Randomization tests [1] are statistical tests in which the data are repeatedly elaborated; a test statistic (***r^2***, ***t***-criteria, ***F***-criteria, etc.) is computed for each data permutation and the proportion of the data permutations, with test statistics values as large as the value for the obtained results, determines the significance of the results. For the testing of the multilinear correlation, the vector **Y** permutations are processed through multilinear regression procedures with fixed columns of matrix **X**. Due to a factorial increase in time spent from the size of the vector **Y**, Monte-Carlo method is often be used for producing randomization test.

Chapter 7 Publications

References:

1. Edington E. S. *Randomization tests*: Marsel Dekker, Inc.: New York and Basel, 1980, pp.195-216.
2. **QSPR Correlation of Free-Radical Polymerization Chain-Transfer Constants for Styrene**
   (A. R. Katritzky, F.H. Ignatz-Hoover, R. Petrukhin, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 295.
3. **Correlation of the Solubilities of Gases and Vapors in Methanol and Ethanol with Their Molecular Structures**
   (A. R. Katritzky, D.B. Tatham, U. Maran) *J. Chem. Inf. Comput. Sci.*, **2001***, 41*, 358.
4. **CODESSA-Base Theoretical QSPR Model for Hydantoin HPLC-RT Lipophilicities**
   (A. R. Katritzky, S. Perumal, R. Petrukhin, E. Kleinpeter) *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 569.
5. **A QSRR-Treatment of Solvent Effects on the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylates Employing Molecular Descriptors**
   (A. R. Katritzky, S. Perumal, R. Petrukhin) *J. Org. Chem.*, **2001**, *66*, 4036.
6. **Interpretation of Quantitative Structure - Property and -Activity Relationships**
   (A. R. Katritzky, R. Petrukhin, D. Tatham, S. Basak, E. Benfenati, M. Karelson, U. Maran) *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 679.
7. **Theoretical Descriptors for the Correlation of Aquatic Toxicity of Environmental Pollutants by Quantitative Structure-Toxicity Relationships**
   (A. R. Katritzky, D.B. Tatham, U. Maran) *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1162.
8. **QSPR Analysis of Flash Points**
   (A. R. Katritzky, R. Petrukhin, R. Jain, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **2001,** *41*, 1521.
9. **Perspective on the Relationship between Melting Points and Chemical Structure**
   (A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin, U. Maran, M. Karelson) *Crystal Growth & Design*, **2001***, 1*, 261.
10. **QSAR Correlations of the Algistatic Activity of 5-Amino-1-aryl-1H-tetrazoles**
    (A. R. Katritzky, R. Jain, R. Petrukhin, S.N. Denisenko, T. Schelenz) *S. Q. Env. Res.*, **2001,** *12*, 259.
11. **QSPR Correlation and Predictions of GC Retention Indexes for Methyl-Branched Hydrocarbons Produced by Insects**
    (A. R. Katritzky, K. Chen, U. Maran, D.A. Carlson) *Anal. Chem.*, **2000***, 72*, 101.
12. **Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties**

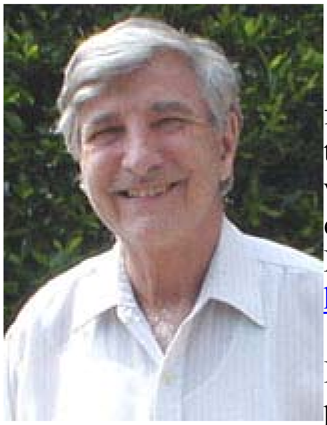
(A. R. Katritzky, U. Maran, V.S. Lobanov, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **2000** *40*, 1.

13. **Prediction of Liquid Viscosity for Organic Compounds by a Quantitative Structure-Property Relationship**
(A. R. Katritzky, K. Chen, Y. Wang, M. Karelson, B. Lucic, N. Trinajstic, T. Suzuki, G. Shuurmann) *J. Phys. Org. Chem.*, **2000**, *13*, 80.
14. **A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines**
(A. R. Katritzky, U. Maran, M. Karelson) *Quant. Struct. Act. Relat. Pharmacol. Chem. Biol.*, **1999**, *18*, 3.
15. **A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors**
(A. R. Katritzky, B. Lucic, N. Trinajstic, S. Sild, M. Karelson,) *J. Chem. Inf. Comput. Sci.*, **1999,** *39*, 610.
16. **QSPR Treatment of Solvent Scales**
(A. R. Katritzky, T. Tamm, Y. Wang, S. Sild, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1999,** *39*, 684.
17. **A Unified Treatment of Solvent Properties**
(A. R. Katritzky, T. Tamm, Y. Wang, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 692.
18. **QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of CODESSA Applications**
(A. R. Katritzky, M. Karelson, U. Maran, Y. Wang) *Collect. Czech. Chem. Commun.*, **1999**, *64*, 1551.
19. **Insights into Sulfur Vulcanization from QSPR Quantitative Structure-Property Relationship Studies**
(A. R. Katritzky, F. Ignatz-Hoover, V.S. Lobanov, M. Karelson) *Rubber Chem. Technol.*, **1999**, *72*, 318.
20. **Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure - Property Relationship** (A. R. Katritzky, V.S. Lobanov, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 28.
21. **Quantitative Structure - Property Relationship (QSPR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers**
(A. R. Katritzky, S. Sild, V. Lobanov, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 300.
22. **Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure**
(A. R. Katritzky, P.D.T. Huibers) *J. Chem. Inf. Comput. Sci.*, **1998***, 38*, 283.
23. **Relationship of Critical Temperatures to Calculated Molecular Properties**
(A. R. Katritzky, L. Mu, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 293 .
24. **QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficients**
(A. R. Katritzky, Y. Wang, S. Sild, T. Tamm, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 720.

25. **General Quantitative Structure-Property Relationship Treatment of the Refractive Index of Organic Compounds**
    (A. R. Katritzky, S. Sild, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 840.
26. **Correlation and Prediction of the Refractive Indices of Polymers by QSPR**
    (A. R. Katritzky, S. Sild, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1998,** *38*, 1171.
27. **Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 2. Anionic Surfactants**
    (A. R. Katritzky, P.D.T. Huibers, V.S. Lobanov, D.O. Shah, M. Karelson) *J. Colloid and Interface Sci.*, **1997***, 187*, 113.
28. **QSPR as a Means of Predicting and Understanding Chemical and Physical Properties in Terms of Structure**
    (A. R. Katritzky, M. Karelson, V.S. Lobanov) *Pure Appl. Chem.*, **1997***, 69*, 245.
29. **Predicting Surfactant Cloud Point from Molecular Structure**
    (A. R. Katritzky, P.D.T. Huibers, D.O. Shah) *J. Colloid and Interface Sci.*, **1997,** *193*, 132.
30. **Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach**
    (A. R. Katritzky, U. Maran, M. Karelson, V.S. Lobanov) *J. Chem. Inf. Comput. Sci.*, **1997,** *37*, 913.
31. **QSPR Treatment of the Unified Nonspecific Solvent Polarity Scale**
    (A. R. Katritzky, L. Mu, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1997,** *37*, 756.
32. **Prediction of Critical Micelle Concentration Using a Quantitative Structure-Property Relationship Approach. 1. Nonionic Surfactants**
    (A. R. Katritzky, P.D.T. Huibers, V.S. Lobanov, D.O. Shah, M. Karelson) *Langmuir*, **1996,** 12, 1462.
33. **Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics**
    (A. R. Katritzky, L. Mu, V.S. Lobanov, M. Karelson) *J. Phys. Chem.*, **1996***, 100*, 10400.
34. **Quantum-Chemical Descriptors in QSAR/QSPR Studies**
    (A. R. Katritzky, M. Karelson, V.S. Lobanov) *Chem. Rev.*, **1996,** 96, 1027.
35. **Prediction of Polymer Glass Transition Temperatures Using A General Quantitative Structure-Property Relationship Treatment**
    (A. R. Katritzky, P. Rachwal, K.W. Law, M. Karelson, V.S. Lobanov) *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 879.
36. **A QSPR Study of the Solubility of Gases and Vapors in Water**
    (A. R. Katritzky, L. Mu, M. Karelson) *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 1162.
37. **Comprehensive Descriptors for Structural and Statistical Analysis. 1 Correlations Between Structure and Physical Properties of Substituted Pyridines**
    (A. R. Katritzky, V.S. Lobanov, M. Karelson, R. Murugan, M.P. Grendze, J.E. Toomey Jr.) *Rev. Roum. Chim.*, **1996***, 41*, 851.

38. **QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure**
(A. R. Katritzky, V.S. Lobanov, M. Karelson) *Chem. Soc. Rev.*, **1995,** 279.
39. **Predicting Physical Properties from Molecular Structure**
(A. R. Katritzky, R.Murugan, M.P. Grendze, J.E. Toomey, Jr, M. Karelson, V. Lobanov, P. Rachwal) *Chem. Tech.*, **1994,** *24*, 17.
40. **Prediction of Gas Chromatogrphic Retention Times and Response Factors Using a General Quantitative Structure-Property Relationship Treatment**
(A. R. Katritzky, E.S. Ignatchenko, R.A. Barcock, V.S. Lobanov, M. Karelson) *Anal. Chem.*, **1994,** *66*, 1799.

Chapter 8 Authors



Professor Alan R. Katritzky

Alan Katritzky is Kenan Professor and Director of the Center for Heterocyclic Compounds in the Department of Chemistry at the University of Florida. He is known internationally for his work in many areas of organic and physical organic chemistry, especially concerning Heterocyclic chemistry on QSPR/QSAR. For further details please see his home page at http://ark.chem.ufl.edu.

In 1999 he conceived the Charitable Trust ARKAT-USA which publishes the electronic journal: "Archive of Organic Chemistry" (ARKIVOC) completely free to users (no access or downloading fees) and authors (no page charges) throughout the world: see http://www.arkat-usa.org

E-mail: Katritzky@chem.ufl.edu
Telephone: 352-392-0554
Fax: 352-392-9199
Mailing Address: Department of Chemistry
University of Florida
PO Box 117200
Gainesville, FL 32611

Professor Mati Karelson

Prof. Mati Karelson, born in 1948, studied chemistry and obtained Ph.D. in chemistry at the Tartu State University, Estonia, in 1975. His professional expertise includes the university teaching and research, being full professor in theoretical chemistry at the University of Tartu from 1992. Dr. Mati Karelson has published over 150 scientific articles, numerous monographs and monographic reviews (*e.g.* M. Karelson, Molecular Descriptors in QSAR/QSPR, J. Wiley & Sons, New York, 2000.). In 1994, he was appointed the Adjunct Professor in Chemistry at the University of Florida and in 1998 received the Honorary Fellowship at the Florida Center for Heterocyclic Compounds. In 2001, he was awarded with Estonian State Prize in Science.

Interests: quantum theory of molecules in condensed media, structure-property relationships, artificial intelligence in chemistry, molecular design.

Consulting: quantitative structure-property/activity relationships (CODESSA) quantum mechanical calculations of molecular properties and spectra in disordered condensed media.

E-mail:	mati@chem.ut.ee
Telephone:	+37/27 375 264
Fax:	+37/27 375 255
Mailing Address:	University of Tartu 2 Jakobi Street Tartu 51014 Estonia

Dr. Ruslan O. Petrukhin

Dr. Ruslan O. Petrukhin, born in 1967, studied chemistry at Ukrainian State University of Chemical Technology, Dniepropetrovsk, Ukraine and obtained his Ph.D. at Tartu State University, Estonia, in 2001. His professional expertise includes chemical scientific software development since 1987, and various QSPR/QSAR research. He published over ten scientific articles and reviews. Since 1999, he works at Prof. Alan R. Katritzky Center for Heterocyclic Compounds at the University of Florida as a leader of computational chemistry group. In 1999 he initiated the CODESSA PRO project and then made major contribution into development of this software package.

Interests: chemical scientific software development, high throughput and distributed computation, chemical database technologies, QSPR/QSAR, molecular design.

E-mail:	ruslanp@chem.ufl.edu
Home Page:	http://chem.ufl.edu/~ruslanp
Telephone:	352-392-0554
Fax:	352-392-9199
Mailing Address:	Department of Chemistry University of Florida PO Box 117200 Gainesville, FL 32611