FAMILIES OF MOLECULAR DESCRIPTORS (FMDs)

***Lorentz JÄNTSCHI*[1]*, Sorana D. BOLBOACĂ*[2]**

[1] Technical University of Cluj-Napoca
[2] Iuliu Hațieganu University of Medicine and Pharmacy Cluj-Napoca

DEFINITION

Families of molecular descriptors were designed and used to relate the structural information with measured properties and activities for different series of chemical compounds. Since a whole pool of experimental designs are devised for conducting measurements (preferably for series of compounds) of which outcomes depends on a variety of factors (including environmental ones) it is unlikely to expect a unique pattern of relating the structure of the compounds with their measurements. On the other hand, by taking into account that from one series of measurements to another, some settings may have small variations, it is expected that the series of the measurements to be related one to the other as well to be both related with the structure of the compounds. This general idea may have different materializations, one of them being the creation and the use of related structural descriptors - families of molecular descriptors.

The strategy to materialize a family of molecular descriptors is to use a genetic code embedding in it a series of different operators (the genome of the family) applying on the structure of the chemical compound (preferably on both topological and geometrical information) to produce a parameterized numerical outcome.

Several families of molecular descriptors were developed to date: FPIF (from Fragmental Property Index Family), MDF (from Molecular Descriptors Family), MDFV (from Molecular Descriptors Family - Vertex), SAPF (from Structural Atomic Property Family), SMPI (from Szeged Matrix Property Indices) and its extension FMPI (from Fragmental Matrix Property Indices) as well as ChPE, which is an extension of the Characteristic Polynomial intended to be used on molecules (see Char-poly).

KEYWORDS:

FPIF (Fragmental Property Index Family); MDF (Molecular Descriptors Family); MDFV (Molecular Descriptors Family - Vertex); SAPF (Structural Atomic Property Family); SMPI (Szeged Matrix Property Indices); FMPI (Fragmental Matrix Property Indices); ChPE (Characteristic Polynomial Extensive)

HISTORICAL ORIGIN(S)

First steps to the molecular models are recorded in 1861 (Loschmidt 1861[1]). Today molecular modelling involves theoretical methods and computational techniques for pushing further (see Rhinehardt et al. 2015[2]) the knowledge about the molecular structure.

When series of compounds are involved, then the expected result of a model is to provide a function or a relation between the structure and macroscopic observed behaviour of the molecules. Strategies like docking (Taha et al. 2015[3]), assaying (Peng et al. 2015[4]) and mapping (Radwan &

Abdel-Mageed 2014[5]) are involved to better exploit the feature of the systematic experimental observation, but nevertheless strategies to develop families of molecular descriptors (FMDs) began to attract concerns (see Kihl et al. 2015[6]).

Modelling the molecular structure is the way of understanding of the microscopic level and its expression at the macroscopic one level. The accessing of the microscopic level is via measurements (see Figure 1).

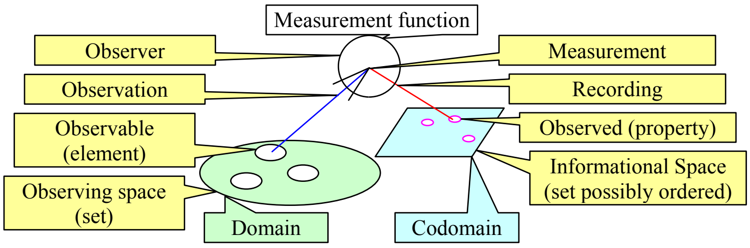


Figure 1. Encoding the information from measurements

In regard of the measurements, there are many ways of expressing the encoded information, differing one from each other by the quality of the representation.

Thus, the primary measurement scale is binomial which encodes (in the informational space) logical values having as allowed operations equality ("=") and negation ("!") providing a structure of Boolean algebra (Boole 1854[7]). Mode and Fisher exact (Fisher 1922[8]) are the allowable statistics on, and examples of measurements associated with the encoded values are distinguishing between dead and alive, and looking for occurrences of the sides of a coin.

(Multi)nomi(n)al scale uses a finite and known series of unordered values to record the observations, being a discrete scale and having allowed the test for equality ("=") providing a structure of a standard set. One statistic have a clear meaning on the values measured on this scale - mode - and comparisons between series of measurements using this scale can be conducted with Chi-square test (Pearson 1900[9]). Examples of measurements expressed with this scale include 'ABO' blood group system, but also the classification of living organisms.

Ordinal scale is encoding discrete values and the allowed operations include the test for equality ("=") and (strict) inequality ("<") providing a structure of commutative algebra (Krull 1935[10]). The allowable statistic is the median and on the information collected with this scale is possible the ranking. An example of information collected using this scale is the number of atoms in molecules.

Interval scale provides continuous values implicitly falling into an interval or domain. As operations is possible to do comparisons using inequality operator ("≤") as well as to do subtractions. It provides a structure of one-dimensional affine space (Berwald 1918[11]), having allowed calculating of the mean, standard deviation, correlation, regression, and ANOVA. Examples include measurements of temperature, distance, time, and energy.

Ratio scale provides too continuous values on non-negative domain having as allowed operations inequality ("≤"), subtraction ("-") and multiplication ("*"). It provides a structure of a one-dimensional vector space (Bolzano 1804[12]) having allowed the most comprehensive list of statistics including geometric and harmonic means, coefficient of variation, doing of logarithms (Napier 1614[13]), and examples include chemical and biological measurements such as pH and sweetness relative to sucrose.

Molecular modelling requires and is feed with measurements. If on one hand stays the measured values (see Figure 1), on the other hand stays the chemical structure (see Figure 2). If the Universe is seen as the whole observing space (see Figure 2) then radiant energy differentiates as having a velocity comparable with light velocity (relativistic velocity) grouping radiations such as β, γ, being differentiated through properties.

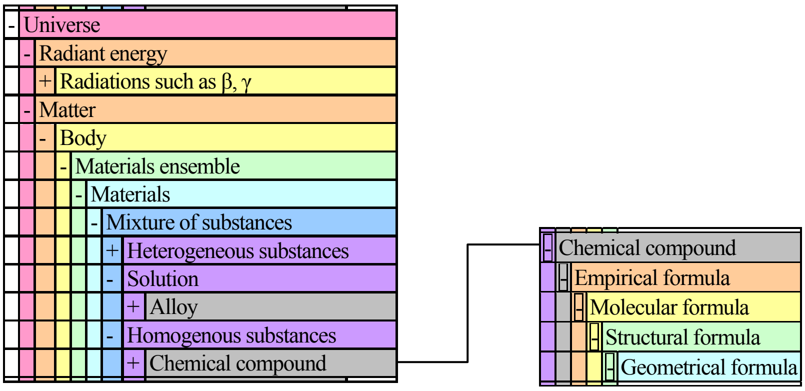


Figure 2. To the layers of the chemical structure

The other main group contains the matter (Figure 2) is seen as the whole non-relativistic observing space in which the body is seen as having the velocity much less than the velocity of light. It contains materials ensemble with possibly variable and discontinue (chemical) composition. Going deeper in the classification, on the next layer stays materials with variable and continue (chemical) composition which generally groups mixtures of substances possessing well defined chemical composition from which homogenous substances have constant (chemical) composition, to finally arrive at chemical compound concept with well defined and unique chemical composition. From this point on we may start to discuss about the chemical structure, and an empirical formula provides the ratio between the atoms in the compound, the molecular formula provides further the number of atoms from each type in the molecule, the structural formula reveals the structural groups in the molecule and finally geometrical formula defines the relative arrangement of the atoms in the molecule. Although it is the last refinement level, sometimes (actually quite often) the geometrical formula may degenerate too being well known the geometrical isomerism (see Warder 1890[14]). Namely, knowing the distances between the atoms and the angles between them we still don't have enough knowledge to define a unique chemical structure, which in some cases may be problematic.

Modelling the molecular structure is a prerequisite for structure-activity inference analysis. Building of a three-dimensional model (3D) is necessary when the calculated descriptors on the structure use the geometry of the molecule. Obtaining the 3D model can be achieved using a molecular modelling program (see Table 1 for a short list of).

Table 1. Molecular modelling software

Name	Provider website	Name	Provider website
Abalone	http://biomolecular-modeling.com	HyperChem	http://hyper.com
ADF	http://scm.com	Materials Studio	http://accelrys.com
ChemBioOffice	http://cambridgesoft.com	Q-Chem	http://q-chem.com
Gaussian	http://gaussian.com	Spartan	http://wavefun.com

When certain software (as given above) is used, sometimes conversions between different formats storing the chemical information are useful, as well as it helps some software for visualising (only) of the obtained models (see Table 2 for a short list of).

Table 2. Molecular modelling auxiliary software

Name	Intend	Open Babel	conversions
GLmol	Browser based visualization	PyMOL	Python application for visualization
Jmol	Java applet for visualization	RasMol	GNU GPL application for visualization
MDL Chime	Browser plugin for visualization	WebQC	conversions

Obtaining of the 3D model of the molecule involves a series of steps, as given below:

÷ Constructing of the topology, namely specification of the atoms by atom type and of the bonds by bond order;

÷ Building of a 3D arrangement, when typical routines possibly including molecular mechanics force fields, such as are CHARM (Brooks et al. 1983[15]), AMBER (Cornell et al., 1995[16]), MMFF94 (Halgren 1996[17]), and OPLS (Jorgensen & Tirado-Rives 1998[18]);

÷ Refining of the 3D arrangement may involve semi-empirical methods, such as are AM1 (Dewar et al. 1985[19]), PM3 (Stewart 1989[20]), RM1 (Rocha et al. 2006[21]) and PM6 (Stewart 2007[22]).

÷ Further refining of the geometry with DFT (density functional theory) approaches including HF (Hartree-Fock, see Hartree 1928[23] & Fock 1930[24]), post-HF - such as are perturbation theory (Møller & Plesset 1934[25]), coupled cluster (Purvis & Bartlett 1982[26]), configuration interaction (Maurice & Head-Gordon 1999[27]), and composite methods (Ohlinger et al. 2009[28]) and KS (Kohn-Sham, see Kohn & Sham 1965[29]) - such as are LDA (Parr & Yang 1994[30]), GGA (Perdew et al. 1992[31]) and PBE (Perdew et al. 1996[32]);

Special precautions at building and of refining of the 3D model should be given to the structures with geometrical isomers, because during the geometrical optimization the passing from one geometrical conformation to another is quite often encountered.

One of the outcomes of the molecular modelling is the charge distribution over the atoms in the molecule, or partial charges. Different approaches are available:

÷ Born (see Born & Goppert-Mayer 1931[33]);

÷ Callen (see Callen 1949[34]);

÷ Szigeti (see Szigeti 1949[35]);

÷ Mulliken (see Mulliken 1955[36] and thereafter);

÷ Coulson (see Coulson et al. 1962[37]);

÷ Politzer (see Politzer 1968[38])

÷ Löwdin (see Löwdin 1970[39]);

÷ Hirshfeld (see Hirshfeld 1977[40]);

÷ Cioslowski (see Cioslowski 1989[41]);

÷ Bader (see Bader 1990[42]);

÷ Optimization method based electrostatic potentials (see for instance Wang & Ford 1994[43]).

Along with the partial charges, the outcome of the molecular modelling includes the (relative) coordinates of the atoms (usually given in Å), the bonds and their types (see Table 3).

Table 3. Typical information from molecular modelling

The list of the atoms			
Label	Type	Coordinates (x, y, z)	Partial charge
The list of the bonds			
Atom Label	Atom Label	Bond type or order	

Usually the methodology for relating the structure with the experimental measurements in series of compounds uses the molecular structure in which the hydrogen atoms are neglected (deleted). Some of the reasons are given in the next:

÷ biological activities determined in vivo have as environment (medium) aqueous solutions in which processes of (partial) dissociation in which the hydrogen atoms pass in the form of protons in solution, leaving the place occupied in the molecular structure;

÷ hydrogen atoms can form a single bond; if they are deleted, excepting their geometrical position information can always be rebuilt;

÷ because form a single bond, the hydrogen atoms do not contribute to the complexity of molecular (not create chains and branches are just terminals for the structure);

deleting of the hydrogen atoms reduces the amount of calculations for a certain structure; considering only an alkane of the general formula C_nH_{2n+2}, removing the hydrogen atoms reduces the complexity of the topology to $^1/_9$ (a topological matrix records values for each pair of atoms and the atoms are about one third less).

NANO-SCIENTIFIC DEVELOPMENT(S)

A genetic code can formally describe the breeding of the molecular descriptors of a family (see Figure 3). The outcome of the breeding process is the population of phenotypes, in which the individuals are labelled with letters encoding their genomic code ($L_n \ldots L_1$ in Figure 3).

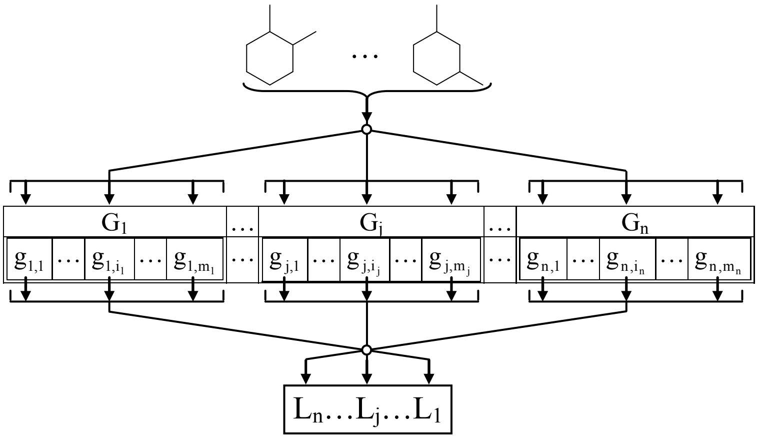

…
G_1
…
G_j
…
G_n
$g_{1,1}$ … g_{1,i_1} … g_{1,m_1}
…
$g_{j,1}$ … g_{j,i_j} … g_{j,m_j}
…
$g_{n,1}$ … g_{n,i_n} … g_{n,m_n}
$L_n \ldots L_j \ldots L_1$

Figure 3. Encoding the information from structure with FMDs

The process of generating the population of descriptors is to be applied for each molecule of the entry dataset in part (see Figure 3) and basically is a for-based procedure (see Figure 4).

```
Input data: series of the molecules ({$Mol_1$, …, $Mol_k$})
For each w from 1 to k //for $Mol_w$
    For each $i_1$ from 1 to $m_1$ //for $G_1$
        … //for the rest of the genes
            For each $i_j$ from 1 to $m_j$ //for $G_j$
                … //for the rest of the genes
                    For each $i_n$ from 1 to $m_n$ //for $G_n$
                        $L_n \ldots L_j \ldots L_1(Mol_w) \leftarrow$ Calculate the descriptor using
                                [$Mol_w$, $g_{1,i_1}$, …, $g_{1,i_j}$, …, $g_{1,i_n}$ ]
                                current values
                    End for
                …
            End for
        …
    End for
End for
Output data: Series of descriptors values for each molecule ({$L_n \ldots L_1(Mol_w)$, $1 \le w \le k$})
```

Figure 4. Obtaining the descriptors of a FMD for a set of molecules

As can be deduced, always a FMD generates exactly the same number of descriptors for any molecule (in the general case, see Figure 3 & Figure 4) the number is $m_1 \cdot \ldots \cdot m_j \cdot \ldots \cdot m_n$.

FPIF (from Fragmental Property Index Family; Jäntschi & Diudea, 2000[44]; see Table 4) is a matrix-based method, in which the matrices collects properties derived from structure for fragments obtained for each pair of atoms.

Table 4. Code of FPIF descriptors

Gene	I_M	D_M	A_P	P_D	F_C	S_M	M_I	L_O
Genome	R	T	M	__p__	si	S	P_	I
	D	G	E	__d__	se	P	P2	R
			C	_1/p_	ji	A	E_	L
			Q	_1/d_	je	G	E2	
				_p*d_	fi	H		
				_p/d_	fe			
				_p/d2				
				p2/d2				

FPIF = $I_M \times D_M \times A_P \times P_D \times F_C \times S_M \times M_I \times L_O$. Examples: RGseCp2/d2SE2, DGjeP_p/d2GP_

FPIF uses $d_M(a,b)$ - the topological distance in structure *M* from atom *a* to atom *b*; $\delta_M(a,b)$ - the topological detour - i.e. longest path - in structure *M* from atom *a* to atom *b*; $W_M(a,b)$ - set of walks;

$P_M(a,b)$ - set of paths; $D_M(a,b)$ - set of distances - i.e. shortest paths; $\Delta_M(a,b)$ - set of detours - i.e. longest paths; M\p - substructure derived from structure M when the atoms inside the path p are removed from *M* together with their connections. Sets (one or more) of atoms of a molecule M for every pair of atoms (a,b) are calculated for every of the following (six) set collecting criteria (called F_C - fragmentation criteria):

- With $Sz_{a,b,d/\delta} = \{c \in M \mid d/\delta_M(c,a) < d/\delta_M(c,b)\}$:
  - F_C = si: $SzDi_{a,b} = \{c \in M \mid d_M(c,a) < d_M(c,b)\}$;
  - F_C = se: $SzDe_{a,b} = \{ c \in M \mid \delta_M(c,a) < \delta_M(c,b)\}$;
- With $CJ_{a,b,p} = \{c \in M \mid d_M(c,a) < d_M(c,b)$ and $\exists w \in W_M(c,a) \mid \{a\}=w \cap p\}$:
  - F_C = ji: $Cj_{a,b,p}$ when $p \in D_M(a,b)$;
  - F_C = je: $Cj_{a,b,p}$ when $p \in \Delta_M(a,b)$;
- With $Cf_{a,b,p} = \{ c \in M \mid d_{G\setminus p}(c,a) < d_{G\setminus p}(c,b)\}$:
  - F_C = fi: $CfDi_{a,b} = Cf_{a,b,p}$ when $p \in D_M(a,b)$;
  - F_C = fe: $CfDe_{a,b} = Cf_{a,b,p}$ when $p \in \Delta_M(a,b)$;

Four atomic properties (A_P an atomic property) are taken into calculation: M (A_P = M) - as relative atomic mass; E (A_P = E) as electronegativity (Sanderson scale, Sanderson 1983[45]); C (A_P = C) as (set) cardinality; P (A_P = P) as partial charge (class I, see Cramer 2002[46], from Mulliken population analysis, Mulliken 1955[36]). Eight property descriptor (P_D a property descriptor) expressions account atomic properties: p (P_D = p) - atomic property; d (P_D = d) - distance; $P_D = 1/p$; $P_D = 1/d$; $P_D = pd$; $P_D = p/d$; $P_D = p/d^2$; $P_D= p^2/d^2$. Five overlapping methods (S_M - superposing method) overlap atomic properties to provide the fragmental property: S (S_M = S) - sum; P (S_M = P) - multiplication; A (S_M = A) - arithmetic mean; G (S_M = G) - geometric mean; H (S_M = H) - harmonic mean. Two models of interaction give transform in a vector a descriptor (I_M - interaction model): R (I_M = R) - rare (uses the assumption that the property of all atoms are approximately located in the fragment centre of property - of which position is consequently obtained and used to express the descriptor vector); D (I_M = D) - dense (the effect of each atom are superposed using vector summation). Two distance metrics (D_M - metric of distance) provides the distance for expressing the descriptor values: T (D_M = T) - topological (from connectivity) and D (D_M = D) - topographical (from 3D model of the molecule obtained from different levels of theory - for a discussion about, see Clark 1985[47]). Four square-matrix based indices (M_I - matrix index) collects overall molecular property: P_ (M_I = P_) - half-sum of matrix elements; P2 (M_I = P2) - half-sum of squared matrix elements; E_ (M_I = E_) - half-sum of Hadamard product of matrix with adjacency matrix; E2 (M_I = E2) - half-sum of squared Hadamard product of matrix with adjacency matrix. Finally, a molecular descriptor is obtained via a linearization operator (L_O - linearization operator) meant to transform nonlinearities to linearity at relationships: I (L_O = I) - identity function; R (L_O = R) - reciprocal function (f(x)=1/x); L (L_O = L) - logarithm function (f(x)=ln(x)). Thus, FPIF family of molecular descriptors puts together a total number of individuals equal with the number of all multiplications described above (2·2·4·8·6·5·4·3 - 46080) - see Table 4.

MDF (from Molecular Descriptors Family; Jäntschi 2004[48]; Jäntschi 2005[49]; see Table 5) is a method based on molecular fragments obtained for pairs of atoms.

Table 5. Code of MDF descriptors

Gene	D_M	A_P	I_D				I_M	F_C	S_M				L_O
Genome	t	C	D	Q	L	F	r	m	m	A	G	H	I
	g	H	d	q	l	f	R	M	M	a	g	h	i
		M	O	J	V	S	m	D	n	B	F	I	A
		E	o	j	E	s	M	P	N	b	f	i	a
		G	P	K	W	T	d		S	P	s		L
		Q	p	k	w	t	D						l

MDF = $D_M \times A_P \times P_D \times I_M \times F_C \times S_F \times L_O$. Examples: lsPRLGg, IhDDDCt

Similarly with FPIF, MDF it uses two distance operators (D_O): topological (t) and geometrical (g), six atomic properties (A_P): cardinality (C), number of directly connected hydrogen atoms (H), relative atomic mass (M), electronegativity (E - Sanderson scale, group electronegativity (G - Diudea

& Silaghi 1989[50]), partial atomic charge (Q - Mulliken, and twenty-four interaction descriptors (I_D) as follows: D(d), d(1/d), O(p_1), o(1/p_1), P(p_1p_2), p(1/p_1p_2), Q($\sqrt{}p_1p_2$), q(1/$\sqrt{}p_1p_2$), J(p_1d), j(1/p_1d), K(p_1p_2d), k(1/p_1p_2d), L(d$\sqrt{}p_1p_2$), l(1/d$\sqrt{}p_1p_2$), V(p_1/d), E(p_1/d_2), W(p_1^2/d), w(p_1p_2/d), F(p_1^2/d^2), f(p_1p_2/d^2), S(p_1^2/d^3), s(p_1p_2/d^3), T(p_1^2/d^4), t(p_1p_2/d^4). Interaction were modelled (I_M) using six functions: R and r - being rare, M and m - being medium, and D and d being dense - the upper letter encoded one having as reference the first atom of the fragment (*a* in the notation given at defining of FPIF) and lower letter nominating the reference on the probe atom (*b* in the notation given at defining of FPIF). Fragmentation is driven by one fragmentation criterion (F_C): m (F_C = m) - defines smallest fragment containing atom *a*; M (F_C = M) - defines largest fragment not containing atom *b*; D (F_C = D) - defines so called Szeged fragments (closer to atom *a* than to atom *b*), P (F_C = P) - Cluj path based fragments (see FPIF definition for the definition of Cluj path based fragments - $CF_{a,b,p}$, p∈D_M(a,b)), nineteen overlapping strategies for fragments interaction (S_F - superposing formula): m (S_F = m) - smallest value; M (S_F = M) - biggest value; n (S_F = n) - smallest absolute value; N (S_F = n) - biggest absolute value; S (S_F = S) - sum of; A (S_F = A) - S divided to number of fragments possessing real value of descriptor; a (S_F = a) - S divided to total number of fragments; B (S_F = B) - S divided to number of atoms; b (S_F = b) - S divided to number of bonds; P (S_F = P) - product of; G geometric mean rooted P as S is divided for A (S_F = A); g (S_F = g) - rooted P as S divided for a (S_F = a); F (S_F = F) - rooted P as S divided for B (S_F = B); f (S_F = f) - rooted P as S divided for b (S_F = b); s (S_F = s) - harmonic sum; H (S_F = H), h (S_F = h), I (S_F = I), i (S_F = i) harmonic means following same procedure from s as G (S_F = G), g (S_F = g), F (S_F = F), f (S_F = f) were derived as geometric means from P and same procedure as for A (S_F = A), a (S_F = a), B (S_F = B), b (S_F = b) derived as arithmetic means from S. Six linearization operators (L_O) being: I (L_O = I) - identity(f(x)=x); i (L_O = i) - inverse (f(x)=1/x), A (L_O = A) - absolute of (f(x)=|x|), a (L_O = a) inverse of absolute of (f(x)=1/|x|), L (L_O = L) - logarithm of (f(x)=ln(x)) and l (L_O = l) - logarithm of absolute of (f(x)=ln(|x|)). Thus, MDF puts together a total number of individuals equal with the number of all multiplications (2·6·6·24·4·19·6 = 787968) - see Table 5.

MDFV (from Molecular Descriptors Family - Vertex; Bolboacă & Jäntschi 2009[51]; see see Table 6) uses atoms in place of pairs of atoms (as FPIF and MDF uses). It implements two distance metrics (D_O): t (topological) and g(geometrical), seven atomic properties (A_P): C (cardinality), H (hydrogen's), M (mass), E (electronegativity, Sanderson scale), Q (partial charge, Mulliken population analysis), L (melting point under normal temperature and pressure conditions), A (electronic affinity), fifty-eight interaction descriptors (I_D, see Table 6).

Table 6. Code of MDFV descriptors

Gene	D_O	A_P	I_D								S_F	S_M	I_T	E_U	L_O
Genome	T	C	J	R	N	Z	V	I	D	A	A	f	D	I
	G	H	j	r	n	z	v	i	d	a	a	F	d	R
		M	O	K	W	S	F	A	0	I	I	c		L
		E	o	k	w	s	f	a	1	i	i	C		
		Q	P	L	X	T	G	B	2	F	F	p		
		L	p	l	x	t	g	b	3	P	P	P		
		A	Q	M	Y	U	H	C	4	C	C	a		
			q	m	y	u	h	c	5			A		
									6			i		
									7			I		

MDFV = D_O×A_P×I_D×S_F×S_M×I_T×E_U×L_O. Examples: TEuIFFDL, GLbIAcDR

Atoms (or vertices in graph theory naming) are cut and fragments (connected atoms) are collected. It is calculated first the fragmental property using one out of ten strategies (I_T - interaction type):

÷ I_T = f - fragment's field - superposes (adds) axial projections of I_D for all pairs of atoms (b,c) from fragment ((b,c) ∈Fr(a)) taken once - giving interactions in the fragment independent of atom cut);

÷ I_T = F - field of the fragment in the cut - superposes (adds) axial projections of I_D for all pairs of atoms (a,b) with one atom in the fragment (b∈Fr(a)) - giving interaction of the fragment in the cut;

÷ I_T = c - fragment's descriptor centre - computes coordinates of the centre of the descriptor using once every pair of atoms of the fragment (b,c)∈Fr(a);

÷ I_T = C - fragmentation descriptor centre - computes coordinates of the centre of the descriptor using all pairs of atoms (a,b) with one atom in the fragment (b∈Fr(a)) - giving the weight of the fragment in the cut;

÷ I_T = p - fragment's potential - uses all pairs (b,c)∈Fr(a) to obtain the average direction (average of the directions) of the field; uses all pairs (b,c)∈Fr(a) to obtain the cumulated value (sums of the effects); gives the intrinsic potential of the fragment;

÷ I_T = P - potential of the fragment relative to the cut - uses all pairs of atoms (a,b) with one atom in the fragment (b∈Fr(a)) for giving the extrinsic potential of the fragment at the cut;

÷ I_T = a - select highest descriptor present in the fragment (from all pairs (b,c)∈Fr(a) of atoms present in the fragment); give strongest interaction in the fragment;

÷ I_T = A - select highest descriptor of the fragment with the cut (from all pairs (a,b) with b∈Fr(a)); give strongest interaction in the cut;

÷ I_T = m - select lowest descriptor present in the fragment (from all pairs (b,c)∈Fr(a) of atoms present in the fragment); give weakest interaction in the fragment;

÷ I_T = M - select highest descriptor of the fragment with the cut (from all pairs (a,b) with b∈Fr(a)); give weakest interaction in the cut.

In general, for a vertex cut more than one fragment may occur. Thus, this fact are accounted using superposing of the descriptors interaction at fragments (between fragments of same cut) level by the superposing at fragment (S_F) formula. When operates in the Minkowski space (using absolute values) two superposing derives: a (S_F = a) - standing for $\max(|(\cdot_x|+|\cdot_y|+|\cdot_z|)$ and i (S_F = i) -standing for $\min(|(\cdot_x|+|\cdot_y|+|\cdot_z|)$. When operates in the Euclidian space (using square values and after squared root of) other two superposing derives: A (S_F = A) - standing for $\max(\sqrt{(}\cdot_x{}^2+\cdot_y{}^2+\cdot_z{}^2))$ and I (S_F = I) - standing for $\min(\sqrt{(}\cdot_x{}^2+\cdot_y{}^2+\cdot_z{}^2))$. When the effects of two or more fragments are superposed, we can superpose it as vectors, and then S_F takes value of F (S_F = F), we can superpose only their directions (and add their values), and then S_F takes the value of P (S_F = P) or weighting their effect, and then S_F takes the value of C (S_F = C). Finally, superposing is conducted at molecular level from all cuts using same procedure described above at superposing at fragments of a cut. Thus, S_M superposes as minimum absolute (when S_M = i), as maximum absolute (when S_M = a), as minimum in Euclidean space (when S_M = I), as maximum in Euclidean space (when S_M = A), weighting effects (when S_M = C), superposing directions (when S_M = P) or vectorial superposing (when S_M = F). All values of the descriptors at molecular level obtained using the procedure described above possess two things: a value and a reference (a coordinate of its position). Thus, we can express as molecular descriptor the value of it (and then E_U = D) or a reference of it (a distance, and then E_U = d) where E_U is the expressing unit). A linearization operator (L_O) serves for linear regression designing of the analysis with MDFV family of descriptors and it takes three values: I (standing for identity with), R (standing for reciprocal or inverse of) and L (standing for logarithm of). Thus, MDFV family of molecular descriptors puts together a total number of individuals equal with the number of all multiplications described above ($2\cdot7\cdot58\cdot7\cdot7\cdot10\cdot2\cdot3 = 2387280$) - see Table 6.

Transforming of MDF to a more complex and large family (as MDFV is) does not provided expected significant improvement of QSAR (quantitative structure-activity relationships) models (with MDFV) as were obtained (with MDF), another approach were developed: SAPF (see Table 7).

SAPF (from Structural Atomic Property Family; Sestraş et al., 2012[52]; see Table 7; calculation details given in Jäntschi 2012[53]) cumulates atomic properties at molecular level. It locates the molecular centre using one (out of three methods, C_F) for this task involving a metric (out of two, M_D) for the distance, a atomic property (out of eight defined till date, A_P), a rising power for the distance (D_P, seven cases), a rising power for the property (P_P, same seven cases). At molecular level one of two sorts of operators (O_M, mean type or sum type) build the molecular property as generalized mean or sum (see O_M) of descriptor's values rising it at a power (G_M, again one out of same seven cases) and the result are subject to linearization (L_O, one out of seven cases).

Table 7. Code of SAPF descriptors

Gene	C_F	D_O	A_P	D_P	P_P	O_M	M_P	L_O
Genome	D	T	C	I	I	S	I	I
	P	G	H	E	E	M	E	A
	C		M	H	H		H	S
			E	G	G		G	T
			A	A	A		A	Q
				Q	Q		Q	R
				S	S		S	L

SAPF = $L_O\times G_M\times O_M\times P_P\times D_P\times A_P\times M_D\times C_F$. Examples: SISHQEGC, TESHIMGP

Thus, SAPF family of molecular descriptors puts together a total number of individuals equal with the number of all multiplications described above (7·7·2·7·7·9(5)·2·3 = 259308 - with 9 atomic properties; 144060 with 5 atomic properties, see Table 7).

SMPI (Szeged Matrix Property Indices; Bolboacă & Jäntschi 2016[54] see Table 8) it has an online interface free to be used (Jäntschi 2014[55]).

Table 8. Code of SMPI descriptors

Gene	A_P	D_M	I_D	M_O	L_O
Genome	A	T	E	m	I
	B	G	U	M	R
	C	U	D	I	L
	D		P	J	
	E			E	
	F			F	
	G				

SMPI = $L_O\times M_O\times I_D\times D_M\times A_P$. Examples: ImETA -first, LFPUG -last

For SMPI distance matrix are calculated, and then for each pair of (distinct) atoms the atoms closer to the first than to the second atom of the pair are collected into (these are fragments; are exactly one fragment associated to a pair of atoms by this way) a matrix (similarly to the unsymmetrical Szeged matrix on paths, but containing sets of atoms in place of their number; for [USzp] matrix definition see Diudea et al. 2001[56]). To each fragment it is assigned an atomic property A_P=A: Atomic mass (a.u.), as sum of; A_P=B: Atomic number (Z), as harmonic sum of; A_P=C: Cardinality (=1), as sum of; A_P=D: Solid state density (kg/m^3), as harmonic mean of; A_P=E: Electronegativity (revised Pauling; for Pauling electronegativity see Pauling 1932[57]; for revised Pauling electronegativity see Allred 1961[58]), as geometrical mean of; A_P=F: First ionization energy (kJ/mol), as average of; A_P=G: Melting point temperature (K), as Euler (PM(p), p=2) mean of. A distance matrix is calculated using three alternatives - D_M=T: Topological distance (bonds); D_M=G: Geometrical distance (Å); D_M=U: Weighted topological distance (as reversed bond order). An interaction descriptor produces the interaction effects matrix operating on the properties and on the distances matrices - I_D=E: $E_{i,j}=P_{i,j}*D_{i,j}$; I_D=U: $U_{i,j}=P_{i,j}/D_{i,j}$; I_D=D: $D_{i,j}=1*D_{i,j}$; I_D=P: $P_{i,j}=P_{i,j}*1$. On the resulted interaction effects matrix a molecular level operator calculates a value - M_O=m: min; M_O=M: max; M_O=I: half-sum($M_{i,j}$); M_O=J: half-sum($M_{i,j}*M_{j,i}$); M_O=E: half-sum($M_{i,j}*Ad_{i,j}$); M_O=F: half-sum($M_{i,j}*M_{j,i}*Ad_{i,j}$). Finally the calculated value is subject to a linearization - L_O=I: I(x)=x; L_O=R: R(x)=1/x; L_O=L: L(x)=Ln(x). A total number of 1512 (7·3·4·6·3) descriptors reflects the molecular structure of a molecule from (slightly) different (from one to another) perspectives (Table 8).

FMPI (from Fragmental Matrix Property Indices; see Table 9) is an improvement were made to SMPI, by extending the principle applied for Szeged fragments (assigned letter: S) to other two matrices collecting fragments from molecule for pairs of atoms, namely to maximal fragments (assigned letter: M) - the largest set containing the first atom of the pair along with all it's connected atoms after removal of the second atom of the pair from molecule and to complements of the maximal fragments (assigned letter: N) - the set containing the second atom of the pair along with the rest of the atoms lost from the molecule when maximal fragments were extracted.

Table 9. Code of FMPI descriptors

Gene	F_C	A_P	D_M	I_D	M_O	L_O
Genome	S	A	T	E	m	I
	M	B	G	U	M	R
	N	C	U	D	I	L
		D		P	J	
		E			E	
		F			F	
		G				

FMPI = L_O×M_O×I_D×D_M×A_P×F_C. Examples: ImETAS -first, LFPUGN -last

Therefore, the gene sequence of FMPI is increased from SMPI with one gene, and the number of descriptors is multiplied with 3 (arriving at 4536; see Table 9).

ChPE (from Characteristic Polynomial Extend; see Table 10) is an extension was recently made to the Characteristic Polynomial in order to be used to produce a family of molecular descriptors (Jäntschi & Bolboacă 2016[59]). The calculations for the extended characteristic polynomial ChPE = $|\lambda\cdot I_A - C_M|$ were conducted diversifying the (atom's) identities (I_A) using atomic properties in 8 levels ('A' - atomic mass, /294.0l; 'B' - cardinality, always 1; 'C' - charges - atomic electrostatic charge, ESP; 'D' - solid state density, in kg/m3, /30000; 'E' - electronegativity (revised Pauling scale, /4.00; 'F' - first ionization potential, in kJ/mol, /1312.0; 'G' - melting point temperature, in K, /3820.0; 'H' - attached hydrogen atoms, /4), diversifying the (molecule's) connectivity (C_M) in 2·3 levels - two: by adjacencies ('g', 'c' and 't') and by distances ('G', 'C', and 'T') and three: diversifying the distance metric ('g' and 'G' - inverse of the geometrical distance; 'c' and 'C' - inverse of the bond order weighted topological distances; 't' and 'T' - inverse of the topological distances) as given in Table 10.

Table 10. Code of ChPE descriptors

Gene	L_O	I_A	C_M	S_A	$\overline{d_1d_2d_3d_4}$
Genome	I	A	G	N	0000..1000
	R	B	C	P	
	L	C	T		
		D	g		
		E	c		
		F	t		
		G			
		H			

ChPE = L_O×I_A×C_M×S_A×$d_1d_2d_3d_4$. Examples: IAGN1000 -first, LHtP1000 -last

Using the selected alternatives for atom's identities (I_A) and for molecule's connectivity the polynomial ChPE = $|\lambda\cdot I_A - C_M|$ is evaluated in 2001 points by giving to the argument (λ) a sign (S_A, 'N' - negative or 'P' - positive) and a value ($\overline{d_1d_2d_3d_4}$, from 0000 to 1000), as $\lambda = \pm\overline{d_1d_2d_3d_4}/1000$. The result of the evaluation is finally is subject to a linearization - L_O=I: I(x)=x; L_O=R: R(x)=1/x; L_O=L: L(x)=Ln(x). A total number of 288144 (3·8·6·2001) descriptors reflects the molecular structure of a molecule from (slightly) different (from one to another) perspectives (see Table 10).

NANO-CHEMICAL APPLICATION(S)

Not implicitly all FMDs are suitable for nano-chemical applications. Namely, a FMD is suitable to be applied at nano-level if the complexity of the calculation did not surpass a polynomial one (e.g. it not involves solving of a hard problem - for hard problems please see Falkenauer 1998[60]). From this point of view, for generating of the fragments, FPIF require the construction of the list of all paths between two atoms, which goes into a hard problem for cycles containing structures. A small modification to the MDF (MDF → MDF2004), namely by replacing Cluj path based fragments with the complement of the maximal fragments makes MDF (MDF →MDF2015) suitable to be used to the nano-level, because the complexity of all others calculations did not surpass $O(N^4)$, where N is the number of the atoms in a molecule. For SMPI and FMPI the complexity of the calculation is $O(N^4)$ too, while for MDFV, SAPF, and ChPE is $O(N^3)$.

SAPF and ChPE are implemented as very fast versions in FreePascal (for FreePascal see Codère et al. 1998-2015[61]), and the rest of the families suitable for nano-level (MDF, MDFV, SMPI, FMPI) were implemented using PHP language (for PHP language see Lerdorf 1994[62] and Gutmans & Suraski 1994-2004[63]).

All implementations were parallelized for exploiting the multi-core feature of the modern computers. Parallelization is portable when is implemented to pass to different tasks different molecules from a set. Thus, Table 11 contains these two implementations (in PHP and in FreePascal).

Table 11. Code of parallelization for processing of sets of molecules (PHP & FreePascal)

PHP	FreePascal
Initialization code:	Initialization code:

define("max_splits",8); define("f_CPU",0);
define("a_CPU","NUMBER_OF_PROCESSORS");
$srv=&$_SERVER; $invoked=$srv["argv"];
$me=&$srv["PHP_SELF"];
if(array_key_exists("1",$srv["argv"]))
$splits=$srv["argv"][1]; else $splits=1;
if(array_key_exists("2",$srv["argv"]))
$c_split=$srv["argv"][2]; else $c_split=0;
if((!is_numeric($splits)) \|\| (!is_numeric($c_split)) ||
($splits>max_splits)){die("wrong execution!\r\n");}
if($splits>1){ echo("$me. in parralel! Task ");
echo(($c_split+1)." of ".$splits.".\r\n");
}else{ $n_CPU=0;
if(array_key_exists(c_CPU,$srv))
$n_CPU=$srv[c_CPU];
if($n_CPU==0){
if($c_split==0){ echo("Run a single task!\r\n");}
else die();
}else{ $splits=$n_CPU-f_CPU;
echo("n-procs: $n_CPU n-splits: $splits."\r\n");
for($i=0;$i<$splits;$i++){
$run_s="START php ".$me." ".$splits." ".$i;
$run[$i]=popen($run_s,"r");
pclose($run[$i]);
} die("All tasks started.\r\n");
}}
if($splits>1)
echo("Split ".($c_split+1)." of ".$splits."\r\n");
	uses SysUtils, Classes, Process, Windows;
	const f_CPUs=0; m_CPUs=8;
	var splits,c_split,i:byte; s:string[255];
	SystemInfo:SYSTEM_INFO; AProcess:TProcess;
	begin
	splits:=1;current_split:=0; //parallelization default state
	if(paramcount=0)then begin
	GetSystemInfo(SystemInfo);
	splits:=SystemInfo.dwNumberOfProcessors-f_CPUs;
	Str(splits,s); AProcess:=TProcess.Create(NIL);
	AProcess.InheritHandles:=FALSE;
	for i:=1 to GetEnvironmentVariableCount do
	AProcess.Environment.Add(GetEnvironmentString(i));
	AProcess.Executable := Paramstr(0);
	AProcess.Parameters.Add(s);
	for c_split:=splits-1 downto 0 do begin
	Str(c_split,s); AProcess.Parameters.Add(s);
	AProcess.Execute; AProcess.Parameters.Delete(1);
	writeln('Starting ',c_split,' child');
	end;
	AProcess.Free;writeln('Started ',splits,' childs.');halt;
	end else begin
	if(paramcount<>2)then halt;//wrong call
	val(paramstr(1),splits,i);if(i<>0)then halt;
	val(paramstr(2),current_split,i);if(i<>0)then halt;
	if((c_split>splits)or(splits>m_CPUs))then halt;
	end;
	writeln('Splits=',splits,' Split=',c_split);

In the program body:	In the program body:

for($i=0;$i<$n_hin;$i++){//for the list of molecules
if($splits>1){if($i % $splits <> $c_split) continue;}
… //do the job for the molecule $i
}
	for i:= m1 downto 0 do begin //for the list of molecules
	if(splits>1)then if(i mod splits <> c_split)then continue;
	… //do the job for the molecule i
	end;

Initially the applications were designed to work with a database and to save the descriptors as well as the later conducted regression analysis on a database; the applications were revised to produce text-based human readable files. The following working plan is to be used for an analysis conducted with FMDs (as the ones described above):

÷ **Stage 0. Preliminary requirements.** This stage is to be applied after a procedure which assumes that the geometry of the molecules is obtained and is saved in '*.hin' - HyperChem format and the partial charges are calculated. Much convenient is to optimize the structures with software which have possibility to parallelize the calculation, such as is Spartan. If it is the case, then conversions from Spartan ('input' and 'output' files) to HyperChem are required. A program (spartan_hin_convert_qsar.php) was designed to do this (and it requires '*.spinput' files to be placed in a directory, '*.txt' Spartan output files to be placed in other one, as well as it requires that the Spartan calculations to be conducted with 'verbose log' in order to contain the partial charges too; then, in a new directory the HyperChem files are generated).

- ÷ **Stage 1. Generation of the descriptors.** A folder containing the structure files is the input data for all programs providing the descriptors in a single file (example: mdf2004_a_generate.php → mdf2004.txt; it applies also for mdf2015_a_generate.php, mdfv2008_a_generate.php, sapf2011.exe, smpi2014.php, fmpi2015.php, and chpf2015.exe). The output files contain matrix-based data, with molecules in columns and descriptors in lines. The values are expressed with 4 significant digits as numbers in general form (in which are expressed with smallest number of characters).
- ÷ **Stage 2. Filtering of the descriptors.** This step is intended only to remove the duplicates - it is possible for simple molecules to have two different descriptors with exactly the same series of values for all molecules in the dataset. Also it is possible that at given precision that the values to be different only in an order of magnitude; thus, the values of the descriptors should be expressed relatively to the order of magnitude of the highest (absolute value). A program compact the outputs (v2_mdf_x_compactize.php for 'mdf*' families and v2_others_compactize.php for the rest); the output files are created as in the following example: mdf2004.txt → mdf2004_r.asc. The 1_sort_all.php program is feed with '*_r.asc' files to produce sorted and distinct series of values (for the descriptors & for the molecules) as '*_t.asc' files as in the following example: mdf2004_r.asc → mdf2004_t.asc
- ÷ **Stage 3. Building of structure - property files.** The properties and/or activities are collected in a file ('properties.asc'), keeping the association with the structure (from '*.hin' files) with the first line having the names of the files containing the structures of the molecules for which the property (or properties) have that value(s). The first column contains the name of the property/activity. The generate_property_files_v2.php program generates files for each property: family_name+"_"+property_name+".txt".
- ÷ **Stage 4. Regression analysis.** From this point on any software may be feed with the data to conduct the regression analysis. A program (_r1v_all.exe) was designed to provide ("r1_"+input_filename) simple linear regressions and other (r2f_all_v2.exe) to account for additive and multiplicative effects with two descriptors.

The use of the FMDs is exemplified in the modelling of C_{42} fullerene isomers continuum elasticity expressed as Total Strain Energy (TSE, in eV), molecules included in the (Bolboacă & Jäntschi 2016[54]) study. The geometries were already available from (Tománek 2015[64]) and here the partial charges were obtained applying HF 6-31G* calculations, and extracted as ground-state Mulliken net atomic charges (from the analysis of SCF wavefunction). The analysis of TSE was conducted by joining the pool of all FMDs described above (except FPIF, for the reason given above).

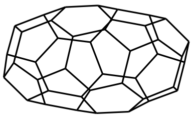
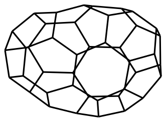
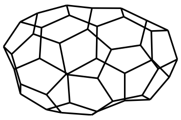
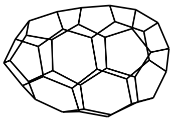
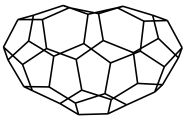
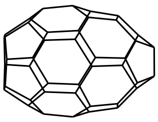
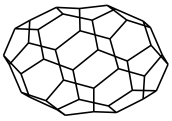
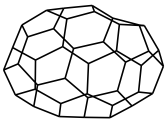
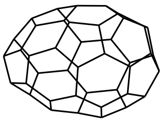
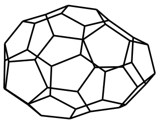
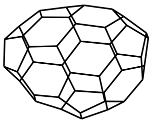
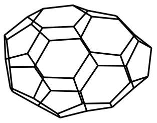
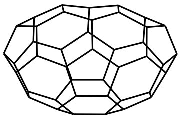
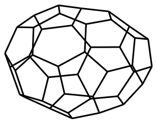
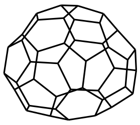
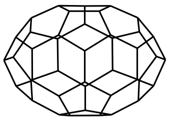
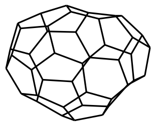
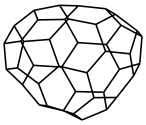
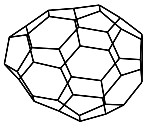
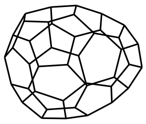
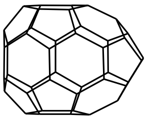
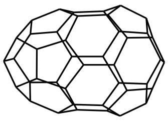
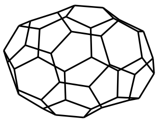
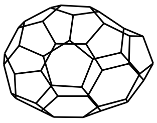
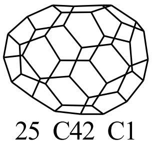
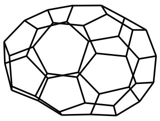
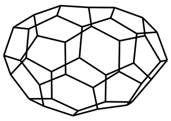
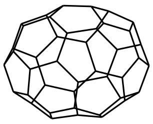
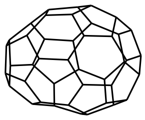
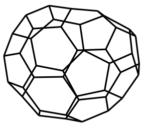
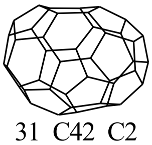
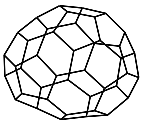
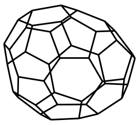
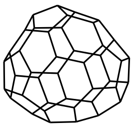
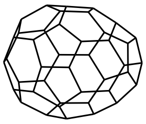
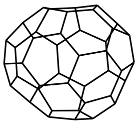
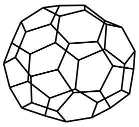
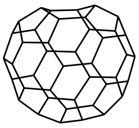
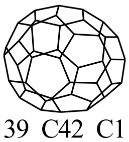
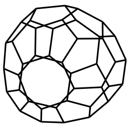
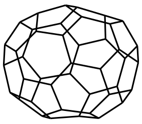
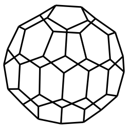
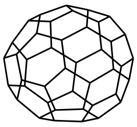
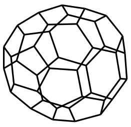
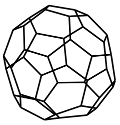
The methodology included the generation of the FMDs descriptors, scaling of the values relative to the highest one, filtering, sorting, joining, and filtering and sorting again.

The size of the valid descriptors from families was as follows: 183050 in ChPE, 1176 in FMPI, 40485 in MDF2004, 188195 in MDF2015, 83187 in MDFV, 6748 in SAPF, 673 in SMPI and 475073 in the joined pool.

The regression analysis included simple linear regression, and multiple linear regression with two structure-based descriptors selected systematically from the whole pool.

The molecules along with TSE values are given in Table 12.

Table 12. All isomers of C_{42} fullerene

TSE: 31.06 eV  01_C42_C2	TSE: 30.537 eV  02_C42_C1	TSE: 29.791 eV  03_C42_C1	TSE: 29.805 eV  04_C42_C1	TSE: 30.618 eV  05_C42_C2
TSE: 29.85 eV  06_C42_C2v	TSE: 30.608 eV  07_C42_C2	TSE: 29.782 eV  08_C42_C1	TSE: 28.527 eV  09_C42_C1	TSE: 29.393 eV  10_C42_C1
TSE: 29.475 eV  11_C42_Cs	TSE: 28.34 eV  12_C42_Cs	TSE: 28.157 eV  13_C42_C2v	TSE: 27.147 eV  14_C42_C1	TSE: 29.955 eV  15_C42_C1
TSE: 28.175 eV  16_C42_C2v	TSE: 28.276 eV  17_C42_C1	TSE: 29.474 eV  18_C42_C1	TSE: 27.408 eV  19_C42_Cs	TSE: 28.175 eV  20_C42_C1
TSE: 27.283 eV  21_C42_C2v	TSE: 29.14 eV  22_C42_Cs	TSE: 28.765 eV  23_C42_C2	TSE: 27.743 eV  24_C42_C1	TSE: 27.487 eV  25_C42_C1
TSE: 28.353 eV  26_C42_C1	TSE: 28.014 eV  27_C42_C2	TSE: 29.051 eV  28_C42_C2	TSE: 27.489 eV  29_C42_C1	TSE: 28.972 eV  30_C42_C1
TSE: 27.484 eV  31_C42_C2	TSE: 26.657 eV  32_C42_C1	TSE: 26.639 eV  33_C42_C1	TSE: 27.371 eV  34_C42_C1	TSE: 26.554 eV  35_C42_Cs
TSE: 27.973 eV  36_C42_C1	TSE: 29.764 eV  37_C42_C1	TSE: 31.101 eV  38_C42_C2	TSE: 26.639 eV  39_C42_C1	TSE: 27.501 eV  40_C42_C2
TSE: 26.672 eV  41_C42_C2	TSE: 28.665 eV  42_C42_Cs	TSE: 28.284 eV  43_C42_C2	TSE: 26.737 eV  44_C42_C1	TSE: 25.661 eV  45_C42_D3

The simple linear regression with highest explanatory power selected a MDF2015 descriptor (lNVRcTu·10^1) with an explained determination of r^2 =0.9886. Also, all other models involving two descriptors (additive, multiplicative and additive-multiplicative) selected MDF2015 descriptors exclusively, this being an important result for MDF2015 in regard to the use of it in nano-chemistry (see Table 13).

Table 13. Best selected additive, multiplicative and additive-multiplicative effects

Model	Equation	r^2
2+1·1v	$\hat{Y} = 0_{t=0} + \text{AMDrgTs}_{\text{MDF2015}}\cdot 10^{6}\cdot 42.53_{t=39} - \text{lDPRgJj}_{\text{MDF2015}}\cdot 133.8_{t=2.38}$ $+ \text{AMDrgTs}_{\text{MDF2015}}\cdot 10^{6}\cdot \text{lDPRgJj}_{\text{MDF2015}}\cdot 42.53_{t=74}$	99.78
1·1v	$\hat{Y} = 195.1_{t=163} - \text{lDBRgKs}_{\text{MDF2015}}\cdot \text{aMTrgTs}_{\text{MDF2015}}\cdot 10^{-7}\cdot 16.23_{t=140}$	99.79
2v	$\hat{Y} = 88.15_{t=12} + \text{AMDrgTs}_{\text{MDF2015}}\cdot 10^{6}\cdot 88.1_{t=97} - \text{lDERgkS}_{\text{MDF2015}}\cdot 31.49_{t=27}$	99.81

The results from Table 13 suggest that an additive model with descriptors from MDF (the 2015 version) provides the highest explanatory power of the association of the TSE with the structure of the C_{42} isomers.

Multi-/Trans- Disciplinary Connection(s)

The using of FMDs to estimate and predict biological activities proved constantly better results when was compared with other alternatives (see for example Bolboacă & Jäntschi 2010[65]).

Open Issues

There is no ideal recipe on about how to design FMDs. Obtained results shown that the refining of the FMDs leads to improvements of the estimation power.

Related List of Abbreviations

Quite often the search for models involving FMDs is conducted using simple linear regressions (SLR) and multiple linear regressions (MLR).

References and Further Reading

[1] Loschmidt, J.; 1861. Konstitutions - formeln der organischen chemie in graphischer derstellung. As: Ostwalds Klassiker der exacten wissenschaften Nr. 190. Leipzig: Richard Anschütz. 384 p.

[2] Rhinehardt, K.L.; Mohan, R.V.; Srinivas, G.; 2015. Computational modeling of peptide-aptamer binding. Methods in Molecular Biology 1268: 313-333.

[3] Taha, M.; Ismail, N.H.; Imran, S.; Selvaraj, M.; Rahim, A.; Ali, M.; Siddiqui, S.; Rahim, F.; Khan, K.M.; 2015. Synthesis of novel benzohydrazone-oxadiazole hybrids as β-glucuronidase inhibitors and molecular modeling studies. Bioorganic and Medicinal Chemistry 23(23): 7394-7404.

[4] Peng, X.; Wang, Q.; Mishra, Y.; Xu, J.; Reichert, D.E.; Malik, M.; Taylor, M.; Luedtke, R.R.; Mach, R.H.; 2015. Synthesis, pharmacological evaluation and molecular modeling studies of triazole containing dopamine D3 receptor ligands. Bioorganic and Medicinal Chemistry Letters 25(3): 519-523.

[5] Radwan, A. A.; Abdel-Mageed, W. M.; 2014. In Silico Studies of Quinoxaline-2-Carboxamide 1,4-di-N-Oxide Derivatives as Antimycobacterial Agents. Molecules 19(2): 2247-2260.

[6] Kihl, O.; Picard, D.; Gosselin, P.-H.; 2015. A unified framework for local visual descriptors evaluation. Pattern Recognition 48(4): 1170-1180.

[7] Boole, G.; 1854. An Investigation of the Laws of Thought. (Reprinted 2003 as Laws of Thought. New York: Prometheus Books), 430 p.

[8] Fisher, R.A.; 1922. On the interpretation of $\chi2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85(1):87-94. doi:10.2307/2340521

[9] Pearson, K.; 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine Series 5 50(302): 157-175. doi:10.1080/14786440009463897

[10] Krull, Wolfgang; 1935. Idealtheorie. As: Ergebnisse der Mathematik, und ihrer Grenzgebiete 4(3). Berlin: Julius Springer. 152 p.

[11] Berwald, L.; 1918. Über affine Geometrie XXVII. Liesche F2, Affinnormale und mittlere Affinkrümmung. Mathematische Zeitschrift 1(1): 63-78.

[12] Bolzano, B.; 1804. Betrachtungen über einige Gegenstände der Elementargeometrie. Prague: Karl Barth. 65 p.

[13] Napier, J.; 1614. Mirifici Logarithmorum Canonis Descriptio. Edinburgh: Andrew Hart. 155 p.

[14] Warder, R. B.; 1890. Recent theories of geometrical isomerism. Science 16(400): 188-188.

[15] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M.; 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comp. Chem. 4(2): 187-217.

[16] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. Jr; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; 1995. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J. Am. Chem. Soc. 117: 5179-5197.

[17] Halgren, T. A.; 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J. Comp. Chem. 17(5-6): 490-519.

[18] Jorgensen, W. L.; Tirado-Rives, J.; 1988. The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. J. Am. Chem. Soc. 110(6): 1657-1666.

[19] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; 1985. Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions. The Journal of the American Chemical Society 107(13): 3902-3909.

[20] Stewart J. J. P.; 1989. Optimization of parameters for semiempirical methods I. Method. The Journal of Computational Chemistry 10(2): 209-220.

[21] Rocha, R. B.; Freire, R. O.; Simas, A. M.; Stewart J. J. P.; 2006. RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. The Journal of Computational Chemistry 27(10): 1101-1111.

[22] Stewart, J. J. P.; 2007. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. The Journal of Molecular Modeling 13(12): 1173-1213.

[23] Hartree, D. R.; 1928. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods. Math. Proc. Cambridge 24(1):89-110. & Hartree, D. R.; 1928. The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion. Math. Proc. Cambridge 24(1):111-132.

[24] Fock, V. A.; 1930. Approximation method for solving the quantum mechanical many-body problem (In German). Zeitschrift für Physik 61(1-2):126-148. & Fock, V. A.; 1930. "Self consistent field" with exchange for sodium (In German). Zeitschrift für Physik 62(11-12):795-805.

[25] Møller, C.; Plesset, M. S.; 1934. Note on an Approximation Treatment form Many-Electron Systems. Physical Review 46(7): 618-622.

[26] Purvis, D. G.; Bartlett, R. J.; 1982. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. The Journal of Chemical Physics 76 (4): 1910-1919.

[27] Maurice, D.; Head-Gordon, M.; 1999. Analytical second derivatives for excited electronic states using the single excitation configuration interaction method: theory and application to benzo[a]pyrene and chalcone. Molecular Physics 96 (10): 1533-1541.

[28] Ohlinger, W. S.; Klunzinger, P. E.; Deppmeier, B. J.; Hehre W. J.; 2009. Efficient Calculation of Heats of Formation. The Journal of Physical Chemistry A 113(10): 2165-2175.

[29] Kohn, W.; Sham, L. J.; 1965. Self-Consistent Equations Including Exchange and Correlation Effects. Physical Review 140(4A): A1133-A1138.

[30] Parr, R.G.; Yang, W.; 1994. Density-Functional Theory of Atoms and Molecules. Oxford: Oxford University Press. 333 p.

[31] Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C.; 1992. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. Physical Review B 46 (11): 6671-6687.

[32] Perdew, J. P.; Burke, K.; Ernzerhof, M.; 1996. Generalized Gradient Approximation Made Simple. Phys. Rev. Lett. 77(18): 3865-3868.

[33] Born, M.; Goppert-Mayer, M.; 1931. Dynamische Gittertheorie der Kristalle. Handbuch der Physik 24(2): 623-794.

[34] Callen, H. B.; 1949. Electric Breakdown in Ionic Crystals. Phys. Rev. 76: 1394-1402.

[35] Szigeti, B.; 1949. Polarisability and dielectric constant of ionic crystals. Trans. Faraday Soc. 45: 155-166.

[36] Mulliken, R. S.; 1955. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. J. Chem. Phys. 23: 1833-1840.

[37] Coulson, C. A.; Redei, L. B.; Stocker, D.; 1962. The Electronic Properties of Tetrahedral Intermetallic Compounds. I. Charge Distribution. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 270(1342): 357-372.

[38] Politzer, P.; 1968. Electron affinities of atoms. Transactions of the Faraday Society 64: 2241-2246.

[39] Löwdin, P. O.; 1970. On the Nonorthogonality Problem. Advances in Quantum Chemistry 5: 185-199.

[40] Hirshfeld, F. L.; 1977. Bonded-Atom Fragments for Describing Molecular Charge Densities. Theoret. Chim. Acta 44: 129-138.

[41] Cioslowski, J.; 1989. General and unique partitioning of molecular electronic properties into atomic contributions. Phys. Rev. Lett. 62(13): 1469-1471.

[42] Bader, R. F. W.; 1990. Atoms in Molecules: A Quantum Theory. Oxford: Oxford University Press. 438 p.

[43] Wang, B.; Ford, G. P.; 1994. Atomic charges derived from a fast and accurate method for electrostatic potentials based on modified AM1 calculations. J. Comput. Chem. 15: 200-207.

[44] Jäntschi, L.; 2000. Predicția proprietăților fizico-chimice şi biologice cu ajutorul descriptorilor matematici. Teză de doctorat (Chimie) - coordonator Prof. Diudea MV. Universitatea "Babeş-Bolyai" din Cluj-Napoca, Cluj-Napoca, Ian. 2000.

[45] Sanderson, R. T.; 1983. Electronegativity and bond energy. Journal of the American Chemical Society 105(8): 2259-2261.

[46] Cramer, C. J.; 2002. Essentials of Computational Chemistry: Theories and Models. New York: J. Wiley. 542 p.

[47] Clark, T.; 1985. A Handbook of Computational Chemistry. New York: Wiley. 332 p.

[48] Jäntschi, L.; 2004. MDF - A New QSAR/QSPR Molecular Descriptors Family. Leonardo Journal of Sciences 3(4): 68-85.

[49] Jäntschi, L.; 2005. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. Leonardo Electronic Journal of Practices and Technologies 4(6): 76-98.

[50] Diudea, M. V.; Silaghi-Dumitrescu, I.; 1989. Valence group electronegativity as a vertex discriminator. Revue Roumaine de Chimie 34(5): 1175-1182.

[51] Bolboacă, S. D.; Jäntschi, L.; 2009. Comparison of QSAR performances on carboquinone derivatives. TheScientificWorldJOURNAL 9(10): 1148-1166.

[52] Sestraş, R. E.; Jäntschi, L.; Bolboacă, S. D.; 2012. Quantum mechanics study on a series of steroids relating separation with structure. JPC - Journal of Planar Chromatography - Modern TLC 25(6): 528-533.

[53] Jäntschi, L.; 2012. Structure vs. Properties: Algorithms and Models. Habilitation thesis in Chemistry. Bucharest: CNATDCU (defended in 2013 in Cluj-Napoca: Babeş-Bolyai University). 175 p.

[54] Bolboacă, S. D.; Jäntschi, L.; 2016. Nano-quantitative structure-property relationship modeling on C_{42} fullerene isomers. Journal of Chemistry. Article in press.

[55] Jäntschi, L.; 2014. Szeged Matrix Property Indices. Online calculation software (release version). URL: http://l.academicdirect.org/Chemistry/SARs/SMPI/

[56] Diudea, M.V.; Gutman, I.; Jäntschi, L.; 2001. Molecular Topology. New York: Nova Science. 335 p.

[57] Pauling, L.; 1932. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. Journal of the American Chemical Society 54(9): 3570-3582.

[58] Allred, A. L.; 1961. Electronegativity Values from Thermochemical Data. Journal of Inorganic and Nuclear Chemistry 17(3-4): 215-221.

[59] Jäntschi, L.; Bolboacă S.D.; 2016. Extending of characteristic polynomial from molecular topology to molecular geometry. Manuscript, author's version of January 25.

[60] Falkenauer, E; 1998. Genetic Algorithms and Grouping Problems. New York: Wiley. 220p.

[61] Codère, C. E.; Mantione, D.; Klämpfl, F.; Maebe, J.; Canneyt, van M.; Vreman, P.; Muller, P.; Voort, van de M.; Boer, de L.; Diehl, A.; Duncan, C.; Gabor, B.; Guenther, S.; Hajny, T.; Lee, J.; May, M.; Neifer, M.; Raab, O.; Schatzl, T.; Scheidler, B.; Sjoholm, N.; Spiegel, M.H.; Tenchio, G.; Wachtmeester, E.; Zago, F.; Schouten, G.; Balogh, K.; 1998. FreePascal: open source compiler for Pascal and Object Pascal (v.0.99.10); 2000: FreePascal v.1.0; 2005: FreePascal v.2.0; 2009: FreePascal v.2.2.4; 2011: FreePascal v.2.4.2; 2012: FreePascal v.2.6.0. 2015: FreePascal v.3.0.0.

[62] Lerdorf, R.; 1994. PHP/FI - initially from Personal Home Page tools (open source online since 1995; PHP v.1.0).

[63] Gutmans, A.; Suraski, Z.; 1997. PHP3 (open source online since 1998; PHP v.3.0); Gutmans, A.; Suraski, Z; 1998. PHP4 (with 'Zend' engine; open source online since 1999; PHP v.4.0); Gutmans, A.; Suraski, Z.; 2004. PHP5 (with 'Zend' engine 2.0; open source online since 1999; PHP v.5.0)

[64] Tománek, D.; 2015. C_{42} Isomers. In: Guide through the Nanocarbon Jungle: Buckyballs, Nanotubes, Graphene, and Beyond. http://nanotube.msu.edu/fullerene/fullerene.php?C=42

[65] Bolboacă, S.D.; Jäntschi, L.; 2010. From molecular structure to molecular design through the molecular descriptors family methodology. In: QSPR-QSAR studies on desired properties for drug design (ed.: Castro E.A.), p. 117-166. Kerala: Research Signpost.