SCOPUS

SCOPUS **Documents**

Jäntschi, L., Katona, G., Diudea, M.V.

Modeling Molecular Properties by Cluj Indices

(2000) *Match*, 41, pp. 151-188.

[a] Fac. of Sci. and Eng. of Materials, Technical University

[b] Bios Research and Production Center

[c] Fac. of Chem. and Chem. Engineering, Babes-Bolyai University, 3400 Cluj, Romania

Abstract

A new approach, leading to a fragmental property index family, FPIF, is presented. Indices are calculated as local descriptors of some molecular fragments and the global values are then obtained by summing the fragmental contributions. The modeling ability of FPIF is demonstrated by modeling some physico-chemical properties and biological activities on selected sets of organic compounds. The results are compared to those reported in some previous works.

References

- Free, S.M., Wilson, J.W.
  A mathematical contribution to structure-activity studies
  (1964) *J. Med. Chem.*, 7, p. 395.
- Gao, C., Govind, R., Tabak, H.H.
  Application of the group contribution method for predicting the toxicity of organic chemicals
  (1992) *Environmental Toxicol. Chem.*, 11, pp. 631-636.
- Kalivas, J.H., Sutter, J.M., Roberts, N.
  Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry
  (1989) *Anal. Chem.*, 61, pp. 2024-2030.
- Kalivas, J.H.
  Generalized simulated annealing for calibration sample selection from an existing set and orthoginization of undesigned experiments
  (1991) *J. Chemometrics*, 5, pp. 37-48.

- Sutters, J.M., Dixon, S.L., Jurs, P.C.
  Automated descriptor selection for Quantitative Structure-Activity Relationships, using generalized simulated annealing
  (1995) *J. Chem. Inf. Comput. Sci.*, 35, pp. 77-84.

- Leardi, R., Boggia, R., Terrile, M.
  Genetic algorithms as a strategy for feature selection
  (1992) *J. Chemom.*, 6, p. 267.

- Lucasius, C.B., Kateman, G.
  Undersanding and using genetic algorithms. Part 1. Concepts, properties and context
  (1993) *Chemom. Intell. Lab. Sys.*, 19, p. 1.

- Diudea, M.V.
  Cluj matrix CJu : Source of various graph descriptors
  (1997) *Commun. Math. Comput. Chem. (MATCH)*, 35, pp. 169-183.

- Diudea, M.V., Minailiuc, O., Katona, G., Gutman, I.
  Szeged matrices and related numbers
  (1997) *Commun. Math. Comput. Chem. (MATCH)*, 35, pp. 129-143.

- Diudea, M.V.
  Cluj matrix invariants
  (1997) *J. Chem. Inf. Comput. Sci.*, 37, pp. 300-305.

- Diudea, M.V., Pârv, B., Topan, M.I.
  Derived Szeged and Cluj indices
  (1997) *J. Serb. Chem. Soc.*, 62, pp. 267-276.

- Diudea, M.V., Gutman, I.
  Wiener-type tpological indices
  (1998) *Croat. Chem. Acta*, 71, pp. 21-51.

- Kiss, A.A., Katona, G., Diudea, M.V.
  Szeged and Cluj matrices within the matrix operator W (M1,M2,M3)
  (1997) *Coll. Sci. Papers Fac. Sci. Kragujevac*, 19, pp. 95-107.

- Gutman, I., Diudea, M.V.
  Defining Cluj matrices and Cluj matrix invariants
  (1998) *J. Serb. Chem. Soc.*, 63, pp. 497-504.

- Diudea, M.V., Parv, B., Gutman, I.
  Detour-Cluj matrix and derived invariants
  (1997) *J.Chem.Inf.Comput.Sci.*, 37, pp. 1101-1108.

- Diudea, M.V., Katona, G., Lukovits, I., Trinajstić, I.
  Detour and Cluj-detour indices
  (1998) *Croat. Chem. Acta*, 71, pp. 459-471.

- Minailiuc, O., Katona, G., Diudea, M.V., Strunje, M., Graovac, A., Gutman, I.
  Szeged fragmental indices
  (1998) *Croat. Chem. Acta*, 71, pp. 473-488.

- Horn, R.A., Johnson, C.R.
  (1985) *Matrix Analysis*,

- Diudea, M.V., Gutman, I., Jäntschi, L.
  Molecular Topology,

- Sears, F.W., Zemansky, M.W., Young, H.D.
  (1976) *University Physics, Fifth Edition, Addison - Wesley Publishing Company*,

- Golender, V., Vesterman, B., Vorpagel, E.
  APEX-3D Expert system for drug design
  (1996) *Network Science*,

- Rose, V.S., Wood, J.
  Generalized cluster significance analysis and stepwise cluster significance analysis with conditional probabilities
  (1998) *Quant. Struct.-Act. Relat.*, 17, pp. 348-356.

- Young, H.D.
  (1962) *Statistical Treatment of Experimental Data*,

- Reif, F.
  (1965) *Fundamentals of Statistical and Thermal Physics*,

- Diudea, M.V., Silaghi-Dumitrescu, I.
  Valence group electronegativity as a vertex discriminator
  (1989) *Rev. Roumaine Chim.*, 34, pp. 1175-1182.

- Diudea, M.V., Kacso, I.E., Topan, M.I.
  Molecular topology. 18. A Qspr/Qsar study by using new valence group carbon-related electronegativities
  (1996) *Rev. Roum. Chim.*, 41, pp. 141-157.

- Crawford Jr., F.S.
  (1968) *Waves*, 3.

- Ivanciuc, O.
  3D QSAR models

QSPR/QSAR Studies by Molecular Descriptors,

- Hopfield, J.J.
  Neural networks and physical systems with emergent collective computational abilities
  (1982) *Proc. Natl. Acad. Sci. U.S.A.*, 79, pp. 2554-2558.
- Zupan, J., Gasteiger, J.
  Neural networks: A new method for solving chemical problems or just a passing phase?
  (1991) *Anal. Chim. Acta*, 248, pp. 1-30.
- Gasteiger, J., Zupan, J.
  Neural networks in chemistry
  (1993) *Angew. Chem. Int. Ed. Engl.*, 32, pp. 503-527.
- Zupan, J., Gasteiger, J.
  (1993) *Neural Networks for Chemists*,
- Bulsari, A.B.
  (1995) *Neural Networks for Chemical Engineers*,
- Devillers, J.
  (1996) *Neural Networks in QSAR and Drug Design*, p. 279.
- Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A., Doucet, J.P.
  13C NMR chemical shift prediction of sp2 carbon atoms in acyclic alkenes using neural networks
  (1996) *J. Chem. Inf. Comput. Sci.*, 36, pp. 644-653.
- Ivanciuc, O.
  Molecular graph descriptors used in neural network models
  (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*, pp. 697-777.
- Gakh, A.A., Gakh, E.G., Sumpter, B.G., Noid, D.W.
  Neural network-graph theory approach to the prediction of the physical properties of organic compounds
  (1994) *J. Chem. Inf. Comput. Sci.*, 34, pp. 832-839.
- Baskin, I.I., Palyulin, V.A., Zefirov, N.S.
  A Neural device for searching direct correlations between structures and properties of chemical compounds
  (1997) *J. Chem. Inf. Comput. Sci.*, 37, pp. 715-721.
- Kireev, D.B.
  ChemNet: A Novel neural network based method for

graph/property mapping
(1995) *J. Chem. Inf. Comput. Sci.*, 35, pp. 175-180.

- Topliss, J.G., Edwards, R.P.
Chance factors in in studies of Quantitative Structure-Activity Relationships
(1979) *J. Med. Chem.*, 22, p. 1238.

- Nikolić, S., Medić-Sarić, M., Matijević-Sosa, J.
A QSAR study of 3-(Phtalimidoalkyl)-pyrazolin-5-ones
(1993) *Croat. Chem. Acta*, 66, pp. 151-160.

- Nagy, P.I., Tokarski, J., Hopfinger, A.J.
Molecular shape and QSAR analysis of a family of substituted dichlorodiphenyl aromatase inhibitors
(1994) *J. Chem. Inf. Comput. Chem.*, 34, pp. 1190-1197.

- Wessel, M.D., Jurs, P.C.
Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure
(1995) *J. Chem. Inf. Comput. Sci.*, 35, pp. 841-850.

- Stanton, D.T., Jurs, P.C.
Development and use of charged partial surface area structural descriptors for Quantitative Structure-Property Relationships studies
(1990) *Anal. Chem.*, 62, p. 2323.

- Goll, E.S., Jurs, P.C.
Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model
(1999) *J. Chem. Inf. Comput. Chem.*, 39, pp. 974-983.

- Stuper, A.J., Brugger, W.E., Jurs, P.C.
(1979) *Computer-assisted Studies of Chemical Structure and Biological Function*,

- Trebst, A., Draber, W.
Advan. Pest. Sci
(1978) *Symp. Papers IV-th Int. Congress Pest. Chem.*, p. 223.

- Benigni, R., Gallo, G., Giorgi, F., Giuliani, A.
On the equivalence between different descriptions of molecules: Value for computational approaches

(1999) *J. Chem. Inf. Comput. Sci.*, 39, pp. 575-578.

Correspondence Address
Diudea M.V.; Fac. of Chem. and Chem. Engineering; Babes-Bolyai University 3400 Cluj, Romania

ISSN: 03406253
Language of Original Document: English
Abbreviated Source Title: Match

Modeling Molecular Properties by Cluj Indices

Lorentz Jäntschi,[a] Gabriel Katona[b] and Mircea V. Diudea[c*]

[a]Faculty of Science and Engineering of Materials, Technical University

[b]Bios Research and Production Center

[c]Faculty of Chemistry and Chemical Engineering, "Babeş-Bolyai" University

3400 Cluj, Romania

Abstract. A new approach, leading to a *fragmental property index family,* ***FPIF***, is presented. Indices are calculated as local descriptors of some molecular fragments and the global values are then obtained by summing the fragmental contributions. The modeling ability of ***FPIF*** is demonstrated by modeling some physico-chemical properties and biological activities on selected sets of organic compounds. The results are compared to those reported in some previous works.

Commun. Math. Comput. Chem. (*MATCH*), **2000**, *41*, 151-188

Introduction

QSPRs (Quantitative Structure-Property Relationships) link quantitatively the physico-chemical properties of chemical compounds with the molecular structure. They provide mathematical models aimed to accurately predict a certain property from the structural attributes.

Some molecular properties (i.e. those of which numerical value vary with changes in the molecular structure) such as the normal boiling point, critical parameters, viscosity, solubility, retention chromatographic index, are often used for characterizing chemicals in databases. However, a certain property is not always available in tables or other reference sources. It is just the case of newly synthesized compounds. As a consequence, methods of estimation / prediction of physico-chemical properties from the structural features of organic

molecules become very important. The advent of combinatorial chemistry in the last decade required automated procedures for predicting various molecular properties.

Monitoring the environmental pollution needs the prediction of toxicity of chemicals in air, waste waters and sole. **QSAR**s (Quantitative Structure-Property Relationships) can be used to predict the toxicity accurately, without using more expensive experimental methods. Drug research and production is also related to the **QSAR** techniques.

QSPRs/**QSAR**s thus relate a molecular property, shown by a series of chemicals, to the structure encoded by a set of descriptors.

In the past works, investigators looked for easily obtainable descriptors.[1,2] As the computer technology developed, the attention of the descriptor designers turned toward more elaborated descriptors, with enhanced ability in modeling a certain molecular property /activity. Large pools of descriptors were thus created.

Nevertheless, a new problem arised: how to select, in real time, a subset of descriptors suitable for the optimal modeling of the chosen property? An algorithm doing such a task must be a simple one, rapid and convergent to a global optimum. Simulated annealing **SA** and genetic algorithms **GA** are already verified procedures.[3-7]

In this paper a new approach, leading to a *fragmental property index family,* ***FPIF***, is presented. These indices are calculated as local descriptors of some fragments of the molecule and, a global index is then obtained by summing the fragmental contributions. The modeling ability of ***FPIF*** is demonstrated on selected sets of organic compounds.

Fragmentation Criteria

The fragmentation criteria define the basic topological descriptors: ***CJ***, ***CF*** and ***Sz***. The fragments are just the entries in the Cluj and Szeged matrices,[8-17] respectively. Before defining these matrices, some graph-theoretical background is needed.

Let $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$ be a connected graph, with $\boldsymbol{V}$ being the set of vertices and $\boldsymbol{E} \subset \boldsymbol{V} \times \boldsymbol{V}$ the set of edges.

A *walk* $\boldsymbol{w}$ is an alternating string of vertices and edges, $\boldsymbol{w}_{1,n} = (\boldsymbol{v}_1, \boldsymbol{e}_1, \boldsymbol{v}_2, \boldsymbol{e}_2, ..., \boldsymbol{v}_{n-1}, \boldsymbol{e}_m, \boldsymbol{v}_n)$, $\boldsymbol{v}_i \in \boldsymbol{V}(\boldsymbol{G})$, $\boldsymbol{e}_i \in \boldsymbol{E}(\boldsymbol{G})$, $\boldsymbol{m} \geq \boldsymbol{n} - \mathbf{1}$, such that any subsequent pair of vertices $(\boldsymbol{v}_{i-1}, \boldsymbol{v}_i) \in \boldsymbol{E}(\boldsymbol{G})$. Revisiting of vertices and edges is allowed. Then $\boldsymbol{V}(\boldsymbol{w}_{1,n}) = \{\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_{n-1}, \boldsymbol{v}_n\}$ is the set of vertices of $\boldsymbol{w}_{1,n}$. Similarly, $\boldsymbol{E}(\boldsymbol{w}_{1,n}) = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_{m-1}, \boldsymbol{e}_m\}$ is the set of edges of $\boldsymbol{w}_{1,n}$

The *length* of a walk, $l(w_{1,n}) = |E(w_{1,n})| \geq |V(w_{1,n})| - 1$, equals to the number of its traversed edges. The walk is *closed* if $v_1 = v_n$ (i.e. its *endpoints* coincide) and is *open* otherwise. The *set of all walks* in G is denoted by $W(G)$.

A *path* p is a walk having all its vertices and edges distinct: $v_i \neq v_j$, $(v_{i-1}, v_i) \neq (v_{j-1}, v_j)$ for any $1 \leq i < j \leq n$. As a consequence, the revisiting of vertices and edges, as well as branching, is prohibited. The *length* of a path is $l(p_{1,n}) = |E(p_{1,n})| = |V(p_{1,n})| - 1$. A closed path is a *cycle* (i.e. *circuit*). The *set of all paths* in G is denoted by $P(G)$.

A *terminal path* $tp_{1,n}$ is the path involving a walk $w = v_1, e_1, v_2, \ldots, v_n, e_n, v_k$, that *is no more a path* in G, for any $v_k \in V(G)$ such that $(v_n, v_k) = e_n \in E$.

A path is *Hamiltonian* if $n = |V(G)|$. In words, a Hamiltonian path visits once all the vertices in G. If such a path is a closed one, then it is a *Hamiltonian circuit.*

The *distance*, d_{ij}, between two vertices v_i and v_j is the length of a *shortest* path joining them, if exists: $d_{ij} = \min l(p_{ij})$; otherwise $d_{ij} = \infty$. A shortest path is often called a *geodesic*. The *eccentricity* of a vertex i, ecc_i, is the maximum distance between i and any vertex j of G: $ecc_i = \max d_{ij}$. The *radius* of a graph, $r(G)$, is the minimum eccentricity among all vertices i in G: $r(G) = \min ecc_i = \min \max d_{ij}$. Conversely, the *diameter*, $d(G)$, is the maximum eccentricity in G: $d(G) = \max ecc_i = \max \max d_{ij}$. The *set of all geodesics* (i.e. distances) in G is denoted by $D(G)$.

The *detour*, δ_{ij}, between two vertices v_i and v_j is the length of a *longest* path joining these vertices, if exists: $\delta_{ij} = \max l(p_{ij})$; otherwise $\delta_{ij} = \infty$. The *set of all detours* (i.e. longest paths) in G is denoted by $\Delta(G)$.

Resuming to the Cluj and Szeged matrix definition.

Let $p \in D(G)$ **or** $p \in \Delta(G)$; the *Cluj Fragments,* CJ and CF represent the sets of vertices obeying the relations

$$CJ_{i,j,p} = \{v \mid v \in V(G);\ d(G)_{v,i} < d(G)_{v,j};\ \text{and}\ \exists\, w \in W_{v,i},\ V(w) \cap V(p) = \{i\}\} \quad (1)$$

$$CF_{i,j,p} = \{v \mid v \in V(G);\ d(G_p)_{v,i} < d(G_p)_{v,j};\ G_p = G - p \quad (2)$$

Here, $G_p = G - p$ is the spanning subgraph, resulted by deleting the path p joining the vertices i and j (except its endpoints), $d(G)$ and $d(G_p)$ denote the topological distances measured in G and G_p, respectively, and $D(G)$ and $\Delta(G)$ have the above mentioned meaning.

The sets $CJ_{i,j,p}$ and $CF_{i,j,p}$ represent subgraphs (connected or not) in G, related to the endpoint i and referred to j and path p.

In *Cluj fragmentation criteria*, the *path* p plays the central role in selecting the fragments. In cycle-containing graphs, more than one path could join the pair (i,j) thus resulting more than one fragment related to i, so that we define the nondiagonal entries $[\mathbf{UM}]_{ij}$ in the Cluj matrices as the maximum cardinality of the sets defined by eqs 1 or 2

$$[\mathbf{UM}]_{ij} = \max_{p} \left| V_{i,j,p} \right| \tag{3}$$

where **M** = **CJD** (Cluj-Distance, $p \in D(G)$); **CJΔ** (Cluj-Detour, $p \in \Delta(G)$), **CFD** (Cluj-Fragmental-Distance, $p \in D(G)$) and **CFΔ** (Cluj-Fragmental-Detour, $p \in \Delta(G)$), and $\left| V_{i,j,p} \right|$ is the cardinality of the set $CJ_{i,j,p}$ or $CF_{i,j,p}$. The diagonal entries are zero. The above definitions hold for any connected graph.

The Cluj matrices are square arrays, of dimension **N** x **N,** usually *unsymmetric* (excepting some symmetric regular graphs). They can be symmetrized, e.g., by the Hadamard product with their transposes

$$\mathbf{SM_p} = \mathbf{UM} \bullet (\mathbf{UM})^{\mathrm{T}} \tag{4}$$

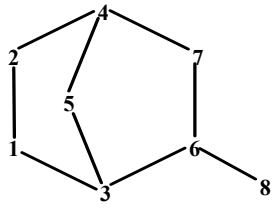
$$\mathbf{SM_e} = \mathbf{SM_p} \bullet \mathbf{A} \tag{5}$$

The symbol • indicates the Hadamard (pairwise) matrix product[18] ($[\mathbf{M}_a \bullet \mathbf{M}_b]_{ij} = [\mathbf{M}_a]_{ij} [\mathbf{M}_b]_{ij}$). For the symmetric matrices, the letter **S** is usually missing. In eq 5, the Hadamard product between the path-defined matrix $\mathbf{SM}_p$ and the adjacency matrix **A (**i.e. the matrix having the non-diagonal entries unity for two adjacent vertices and zero otherwise**)** provides the corresponding edge-defined matrix, $\mathbf{SM}_e$, which is a weighted adjacency matrix.

In trees, **CJD** , **CJΔ , CFΔ** and **CFΔ,** are identical, due to the uniqueness of the path joining a pair of vertices (i,j). Some special properties of Cluj matrices were exposed elsewhere.[10,12,15]

As noted above, the sets $CJ_{i,j,p}$ and $CF_{i,j,p}$ represent subgraphs in G, either connected or not, related to the endpoint i and referred to j and path p. The connectivity of CF sets was

demonstrated elswhere.[19] A case in which $CJD_{i,j,p}$ is disconnected while $CFD_{i,j,p}$ is connected is illustrated in Figure 1. Along with the Cluj matrices, the sets $CJ_{i,j,p}$ and $CF_{i,j,p}$ are presented as well.



G_1

UCJD(G_1)

	1	2	3	4	5	6	7	8
1	0	4	2	2	2	2	2	4
2	3	0	2	2	2	2	2	**3**
3	5	4	0	4	4	4	3	5
4	3	5	3	0	3	3	4	4
5	3	4	2	2	0	3	3	4
6	3	3	3	3	3	0	4	7
7	3	3	2	3	3	3	0	4
8	1	**1**	1	1	1	1	1	0

UCFD(G_1)

	1	2	3	4	5	6	7	8
1	0	4	2	2	2	2	3	5
2	3	0	2	2	2	3	2	**4**
3	5	4	0	4	4	4	3	6
4	3	5	3	0	3	3	4	5
5	5	5	2	2	0	4	4	5
6	3	4	3	3	3	0	4	7
7	3	3	2	3	3	3	0	6
8	1	**1**	1	1	1	1	1	0

$CJD_{i,j,p}$	$CFD_{i,j,p}$
(2, 8) [2, 4, 7, 6, 8] {2,1,5} (disconnected) (2, 8) [2, 1, 3, 6, 8] { 2, 4, 5} (8, 2) [8, 6, 3, 1, 2] { 8} (8, 2) [8, 6, 7, 4, 2] { 8}	(2, 8) [2, 4, 7, 6, 8] { 2,1, 3, 5} (2, 8) [2, 1, 3, 6, 8] { 2, 4, 5, 7} (8, 2) [8, 6, 3, 1, 2] { 8} (8, 2) [8, 6, 7, 4, 2] { 8}

Figure 1.Unsymmetric Cluj matrices and fragmentation for the graph G_1.

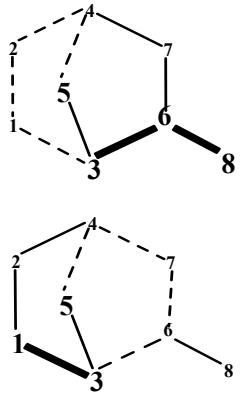
UCJΔ(G_1)

	1	2	3	4	5	6	7	8
1	0	1	1	1	1	2	1	2
2	1	0	1	1	1	1	2	1
3	2	2	0	3	**3**	2	2	2
4	2	2	2	0	2	2	2	3
5	1	1	**1**	1	0	1	1	1
6	3	2	2	2	2	0	2	7
7	1	3	1	1	1	1	0	1
8	1	1	1	1	1	1	1	0

UCFΔ(G_1)

	1	2	3	4	5	6	7	8
1	0	1	1	1	1	2	1	2
2	1	0	1	1	1	1	2	1
3	2	2	0	3	**4**	2	2	2
4	2	2	2	0	4	2	2	3
5	1	1	**1**	1	0	1	1	1
6	3	2	2	2	2	0	2	7
7	1	3	1	1	1	1	0	1
8	1	1	1	1	1	1	1	0

$CJ\Delta_{i,j,p}$	$CF\Delta_{i,j,p}$
(3, 5) [3,1,2,4,5] {3,6,8}	(3, 5) [3,1,2,4,5] {3,6,7,8}
(3, 5) [3,6,7,4,5] {3,1}	(3, 5) [3,6,7,4,5] {3,1,2}
(5, 3) [5,4,2,1,3] {5}	(5, 3) [5,4,2,1,3] {5}
(5, 3) [5,4,7,6,3] {5}	(5, 3) [5,4,7,6,3] {5}



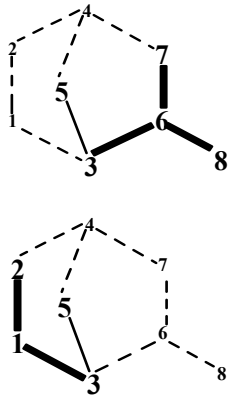


Figure 1. (continued)

An interesting property is shown by the detour-based matrices: $\mathbf{CJ\Delta}_p$ and $\mathbf{CF\Delta}_p$. Let consider the vertices 8 (of degree 1) and 5 (of degree 2) in $\boldsymbol{G_1}$, Figure 1. The vertex 8 is an *external* vertex (with a terminal path ending in it) while the vertex 5 is an *internal* one (usually a terminal path not ending in it). An external vertex, like 8, shows all its entries in the Cluj matrices equal to 1 (see Figure 1). The same entries are shown by the internal vertex 5. This unusual property is called *the internal ending of all detours* joining a vertex ***i*** and the remaining vertices in ***G***. Such a vertex is called an *internal endpoint*.[15] There exist graphs with all the vertices internal endpoints and their detours are *Hamiltonian paths* now. This kind of graph we call the *full Hamiltonian detour graph*, ***FHΔ***.[19]

According to eq 3, sets of *maximal fragments* $\boldsymbol{CJDS^{M}_{i,j}}$, $\boldsymbol{CJ\Delta S^{M}_{i,j}}$, $\boldsymbol{CFDS^{M}_{i,j}}$ and $\boldsymbol{CF\Delta S^{M}_{i,j}}$ are expected in some graphs. This observation is important in the calculation of Cluj property indices (see the next section).

Szeged Fragments, $\boldsymbol{SZD_{i,j}}$ and $\boldsymbol{SZ\Delta_{i,j}}$ are defined by the equations:

$$\boldsymbol{SZD_{i,j}} = \{v \mid v \in V(G);\ \ d(G)_{v,i} < d(G)_{v,j}\} \tag{6}$$

$$\boldsymbol{SZ\Delta_{i,j}} = \{v \mid v \in V(G);\ \ \delta(G)_{v,i} < \delta(G)_{v,j}\} \tag{7}$$

These fragments represent the entries in the unsymmetric Szeged matrices, USZD and USZΔ.[19] In eq 5, $\boldsymbol{\delta(G)_{v,i}}$ represents the detour between the vertices *i* and *v*.

Note that in the definition of the Szeged fragments, the path joining the vertices *i* and *j* is irrelevant. Thus, in *Szeged fragmentation criteria*, each pair (*i,j*) effects one and only one fragment. It was demonstrated elsewhere[19] that the Szeged sets $\boldsymbol{SZD_{i,j}}$ represent connected subgraphs (i.e. fragments) while $\boldsymbol{SZ\Delta_{i,j}}$ are not necessary connected.

In any graph, $\mathbf{CJD_e = CFD_e = SZD_e}$. In cyclic graphs, $\mathbf{CJD_p \neq CFD_p \neq SZD_p}$, $\mathbf{CJ\Delta_p \neq CF\Delta_p \neq SZ\Delta_p}$.

Fragmental Property Indices

Model Parameters

It is well known that the physical laws govern the natural phenomena. Macroscopic interactions are interactions of field-type. The field is produced by a scalar function of potential. Let $\boldsymbol{f(x, y, z)}$ be such a scalar function. This function induces a field given in terms of the gradient of *f*:

$$\vec{\nabla} \cdot \boldsymbol{f} = \left(\frac{\partial}{\partial \boldsymbol{x}} \vec{\boldsymbol{i}} + \frac{\partial}{\partial \boldsymbol{y}} \vec{\boldsymbol{j}} + \frac{\partial}{\partial \boldsymbol{z}} \vec{\boldsymbol{k}} \right) \cdot \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \vec{\boldsymbol{i}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}} \vec{\boldsymbol{j}} + \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{z}} \vec{\boldsymbol{k}} \tag{8}$$

For the potential of type

$$\boldsymbol{f(x, y, z) = pz} \tag{9}$$

and applying eq 8 we obtain the associated field, in the form:

$$\vec{\nabla}\cdot f = \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial y}\vec{j} + \frac{\partial f}{\partial z}\vec{k} = \frac{\partial (pz)}{\partial x}\vec{i} + \frac{\partial (pz)}{\partial y}\vec{j} + \frac{\partial (pz)}{\partial z}\vec{k} = \\ = 0\vec{i} + 0\vec{j} + p\vec{k} = p\vec{k} = \vec{p} \tag{10}$$

This is the case of the well-known uniform gravitational field:

$$\vec{G} = m\vec{g} \tag{11}$$

the potential of which is given by

$$E_p = E_p(z) = mgz \tag{12}$$

where m is the mass of the probe and z is the reference coordinate.

Note that eq 10 is applicable not only to the Newtonian (gravitational) interactions but also to the Coulombian (electrostatic) interactions. In both cases the relation is valid if the mass m (or the charge q) that generates the potential f and associated field $\vec{\nabla}\cdot f$ is far enough ($r >> z$) for the approximation $(r + z)^2/r^2 = (r^2 + 2rz + z^2)/r^2 = 1 + 2z/r + (z/r)^2 \cong 1$ be applied in the equation of field produced by m or q (see below).

For the potential of type:

$$f(x, y, z) = p/z \tag{13}$$

eq 8 leads to the associated field:

$$\vec{\nabla}\cdot f = \frac{\partial f}{\partial x}\vec{i} + \frac{\partial f}{\partial y}\vec{j} + \frac{\partial f}{\partial z}\vec{k} = \frac{\partial (p/z)}{\partial x}\vec{i} + \frac{\partial (p/z)}{\partial y}\vec{j} + \frac{\partial (p/z)}{\partial z}\vec{k} = \\ = 0\vec{i} + 0\vec{j} + \frac{-p}{z^2}\vec{k} = -\frac{p}{z^2}\vec{k} = -\frac{p}{z^3}\vec{z} = -\frac{\vec{p}}{z^2} \tag{14}$$

This is the case of well-known (non-uniform) gravitational field given by:

$$\vec{G} = \vec{G}(m, r) = -k\frac{m}{r^3}\vec{r} \tag{15}$$

and the associated potential of the form:

$$U = U(m, r) = k\frac{m}{r} \tag{16}$$

where m is the mass of the probe and r is the position relative to the location of the point producing the field.

For the Coulombian field eq 14 becomes:

$$\vec{F}_C = \vec{F}_C(r) = -k\frac{q}{r^3}\vec{r} \tag{17}$$

and the potential associated to the Coulombian field:

$$U = U(q,r) = k\frac{q}{r} \tag{18}$$

For *fragmental property indices* four models of interaction are implemented:[19] two of them are *topological* (dense topological and rare topological) and two others are *geometric* (dense geometric and rare geometric).

The *models* are related to two types of field interactions: one of *weak dependence on distance* for the potential of the type (9) generating a uniform field (10), and the second, of *strong dependence on distance* for the potential of the type (13) that generates a non-uniform field (14).

The variables in the models are *metrics of distance* $\boldsymbol{d}$ (topological $\boldsymbol{d_T}$ and geometric $\boldsymbol{d_E}$), *properties* $\boldsymbol{\Phi}$ (mass $\boldsymbol{M}$, electronegativity $\boldsymbol{E}$, cardinality $\boldsymbol{C}$, partial charge or any other atomic property $\boldsymbol{P}$), *property descriptors* $\boldsymbol{\Omega}$ (**p, d, pd, 1/p, 1/d, p/d, p/d2, p2/d2**) and method of *superposition* $\boldsymbol{\Psi}$ (**S, P, A, G, H**).

Let $\boldsymbol{x_1, \ldots, x_n}$ be rational numbers; the (mathematical) *superposition* is

$$\boldsymbol{\Psi}: \; S = \sum_{i=1}^{n} x_i \; ; \; P = \prod_{i=1}^{n} x_i \; ; \; A = S/n \; ; \; G = \left(sgn(P)\right)^n \cdot \sqrt[n]{abs(P)}; \; H = \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1} \tag{19}$$

The expressions for the *property descriptors* are:

$$\boldsymbol{\Omega}: p = p \; ; \; d = d \; ; \; pd = p \cdot d \; ; \; 1/p = \frac{1}{p} \; ; \; 1/d = \frac{1}{d} \; ; \; p/d = \frac{p}{d} \; ; \; p/d2 = \frac{p}{d^2} \; ; \; p2/d2 = \frac{p^2}{d^2} \tag{20}$$

where $\boldsymbol{p}$ is any property ($\boldsymbol{p \in \Phi}$) and d is any metric of distance.

These variables are most frequently used in our models by the following reasons:

- The expressions of the *property descriptor* $\boldsymbol{\Omega}$ simulate the most occurring physical interactions (e.g. **p, pd, p/d, p/d2, p2/d2**)[20] and the most usual descriptors in topological and geometric models. The property descriptor is used either in the calculation of the *vertex descriptor* (when d is the distance from the vertex $\boldsymbol{v}$ to $\boldsymbol{j}$ and $\boldsymbol{p}$ is any atomic property) or in the evaluation of the *fragment descriptor* (when d is the distance between the center of property of the fragment and j, while p is a calculated fragment property).

- The (mathematical) superposition is applied upon a string of vertex descriptors for giving a fragment descriptor. Note that $\boldsymbol{\Psi}$ means $\boldsymbol{S}$ = *sum*; $\boldsymbol{P}$ = *product*; $\boldsymbol{A}$ = *arithmetic mean*; $\boldsymbol{G}$ = *geometric mean* and $\boldsymbol{H}$ = *harmonic sum*. The summation is suitable in the case of any additive property (mass, volume, partial charges, electric capacities, etc.).[21] The multiplication

occurs in concurrent phenomena (probabilistically governed).[22] The arithmetic mean is useful in evaluating some mean contributions (corresponding to some uniform probabilistic distribution).[23,24] The geometric mean is used in calculating the group electronegativities.[25,26] Finally, the harmonic sum is important in connection with the elastic forces, electric fields and group mobility in viscous media.[27]

Model Descriptions

Let (***i,j***) be a pair of vertices and $\boldsymbol{Fr_{i,j}}$ any fragment related to ***i*** and referred to ***j***.

Dense Topological Model

Let ***v*** be a vertex in the fragment $\boldsymbol{Fr_{i,j}}$. The property descriptor applies to the vertex property $\boldsymbol{p_v}$ and topological distance $\boldsymbol{d_{Tv,j}}$. The fragmental *property descriptor* ***PD***, resulting by the vertex descriptor superposition, gives the interaction of all the points belonging to the fragment $\boldsymbol{Fr_{i,j}}$ with the point *j*:

$$\boldsymbol{PD(Fr_{i,j}) = \underset{v \in Fr_{i,j}}{\Psi} (\Omega(d_{Tv,j}, p_v))} \qquad (21)$$

The *j* point can be conceived as an *internal probe atom* (see the *CoMFA* approach).[28] However, the chemical identity of *j* is not considered.

Rare Topological Model

Within this model, the property descriptor applies to the fragmental property and topological distance $\boldsymbol{d_{T\ i,j}}$. The fragmental property descriptor models the interaction of the whole fragment $\boldsymbol{Fr_{i,j}}$ with the point *j* and looks the global property being *concentrated* in the vertex *i*:

$$\boldsymbol{PD(Fr_{i,j}) = \Omega(d_{T\,i,j}, \underset{v \in Fr_{i,j}}{\Psi} (p_v))} \qquad (22)$$

Dense Geometric Model

The fragmental property descriptor is the vector sum of the vertex descriptor vectors. It applies the property descriptor to the vertex property $\boldsymbol{p_v}$ and the Euclidean distance $\boldsymbol{d_{E,v,j}}$ in providing a *point of equivalent* (*fragmental*) *property* located at the Euclidean distance

$d_{E,CP_i,j}$ (with $d_{E,CP_i,j}$ being the distance between the center of fragmental property *CP*, of $Fr_{i,j}$, and the vertex j). The vector of the fragmental property has the orientation of this *distance vector*. The model simulates the interactions in non-uniform fields (gravitational, electrostatic, etc):

$$PD(Fr_{i,j}) = \left\| \sum_{v \in Fr_{i,j}} \vec{\Omega}\left(d_{E\,v,j}, p_v\right) \right\|; \quad \vec{\Omega} = \Omega \frac{\vec{d}_{E\,v,j}}{d_{E\,v,j}}; \quad P(Fr_{i,j}) = \underset{v \in Fr_{i,j}}{\Psi}(p_v);$$

$$d_{E,CP_i,j} = \Omega_p^{-1}(DG(Fr_{i,j}), P(Fr_{i,j})), \tag{23}$$

where $d_{E,CP_i,j}$ is the distance that satisfies: $\Omega(d_{E,CP_i,j}, P(Fr_{i,j})) = PD(Fr_{i,j})$

Rare Geometric Model

The scalar fragmental descriptor applies the property descriptor to the *center of fragmental property* and Euclidean distance between this center and the vertex *j*.

The model simulates the interactions in uniform fields (uniform gravitational, electrostatic, etc.):

$$PD(Fr_{i,j}) = \Omega(d_{E,CP_i,j}, \underset{v \in Fr_{i,j}}{\Psi}(p_v));$$

$$CP_i(x_{CP_i,j}, y_{CP_i,j}, z_{CP_i,j}); \quad x_{CP_i,j} = \sum_{v \in Fr_{i,j}} x_v \cdot p_v \Big/ \sum_{v \in Fr_{i,j}} p_v \tag{24}$$

$$y_{CP_i,j} = \sum_{v \in Fr_{i,j}} y_v \cdot p_v \Big/ \sum_{v \in Fr_{i,j}} p_v; \quad z_{CP_i,j} = \sum_{v \in Fr_{i,j}} z_v \cdot p_v \Big/ \sum_{v \in Fr_{i,j}} p_v$$

Some Particular Fragmental Property Models

Let i, j be two vertices in $V(G)$ and $Fr_{i,j}$ any fragment related to i with respect to j.

Fragmental Mass

In evaluating the fragmental mass, the chosen property is $\Phi = M$, descriptor $\Omega = p$, superposition $\Psi = S$, and the model is *rare topological*, ***RT***. The fragmental mass descriptor takes the form:

$$PD(Fr_{i,j}) = \sum_{v \in Fr_{i,j}} M_v \tag{25}$$

It models the molecular mass of the fragment. The *name* of the associated property matrix is **RTcDdM__p__S**, with the known meaning for ***c*** and ***Dd***.

If ***c*** = ***s*** and ***Dd*** = ***Di*** then **RTsDiM__p__S,** it models the molecular mass of the Szeged Distance Fragments (eq 6). If ***c*** = ***f*** and ***Dd*** = ***Di*** then the matrix **RTfDiM__p__S** collects mean values of mass of all the fragments belonging to ***i*** (with respect to ***j***) according to the ***CF*** criterion (eqs 2 and 3).

Fragmental Electronegativity

The well known equalizing principle of electronegativity *E*, is here considered: the fragment electronegativity is the geometric mean of electronegativities of the *s* atoms joined to form that fragment.

Let the property $\Phi = E$ (electronegativity); descriptor $\Omega = p$; superposition $\Psi = G$; the model is *rare topological*, ***RT***. The fragmental electronegativity descriptor is:

$$PD(Fr_{i,j}) = |Fr_{i,j}| \sqrt{\prod_{v \in Fr_{i,j}} E_v} \tag{26}$$

It models the electronegativity of the fragment $Fr_{i,j}$. The *name* of the property matrix associated with it is **RTcDdE__p__G** . Note that E_v is the group electronegativity for vertex ***v*** calculated with formula:

$$E_v = \sum_{j \in \Gamma_v}^{b(v,j)} \sqrt{\prod_{j \in \Gamma_v} E_j^{b(v,j)}} \tag{27}$$

where $b(v, j)$ is the *conventional bond order* between ***v*** and ***j*** (e.g. 1, 1.5, 2, 3 for single, aromatic, double and triple bonding, respectively), E_a is the atomic electronegativity (Sanderson) and $j \in \Gamma_v$ is any atom (hydrogen atoms included) consisting the group Γ_v.

Fragmental Numbers

The property $\Phi = C$ (cardinality) was introduced for recovering some graph-theoretical quantities and/or graph theoretical analogue indices (see below).

For descriptor $\Omega = p$, superposition $\Psi = \{P, A, G\}$, and the model *rare topological*, ***RT***, the cardinal numbering descriptor of $Fr_{i,j}$ is:

$$PD(Fr_{i,j}) = \prod_{v \in Fr_{i,j}} 1 = \frac{\sum_{v \in Fr_{i,j}} 1}{|Fr_{i,j}|} = |Fr_{i,j}| \sqrt{\prod_{v \in Fr_{i,j}} 1} = 1 \tag{28}$$

The arithmetic mean ***A***, geometric mean ***G*** and product ***P*** applied to **1** (value for vertex property) leave it unchanged. The mean value for all fragments belonging to ***i*** vs. ***j*** (***CJ*** and ***CF*** only) is also 1. All matrices **RTcDdC__p__P**, **RTcDdC__p__A** and **RTcDdC__p__G** have all their entries unity, except the main diagonal elements that are zero.

The corresponding path-calculated indices give the number of edges in the complete graph having the same number of vertices *N*, as the considered molecular graph:

$$RTcDdC\_\_p\_\_PP\_ = RTcDdC\_\_p\_\_AP\_ = RTcDdC\_\_p\_\_GP\_ = N(N-1)/2$$

Similarly, the edge-calculated indices, ***RTcDdC__p__PE_***, ***RTcDdC__p__AE_***, ***RTcDdC__p__GE_*** give the number of edges in the molecular structure.

Let now the property $\Phi = C$, descriptor $\Omega = p$, superposition $\Psi = S$ and *rare topological* model. The value of cardinal numbering descriptor for $Fr_{i,j}$ is:

$$PD(Fr_{i,j}) = \sum_{v \in Fr^{\sigma_t}{}_{i,j}} 1 = \left| Fr^{\sigma_t}{}_{i,j} \right| \tag{29}$$

It models the number of atoms in the fragment. The associated matrices are of the form **RTcDdC__p__S**:

RTfDiC__p__S = CFD ; **RTfDeC__p__S = CFΔ**

RTjDiC__p__S = CJD ; **RTjDeC__p__S = CJΔ**

RTsDiC__p__S = SZD ; **RTsDeC__p__S = SZΔ**

and the corresponding indices are exactly the graph-theoretical descriptors corresponding to the Cluj and Szeged criteria (eqs 1-3, 6, 7).

Uniform Field Gravity

Let the property $\Phi = M$, descriptor $\Omega = p/d2$, superposition $\Psi = S$ and *rare geometrical* model.

The uniform gravity descriptor of $Fr_{i,j}$ is calculated by:

$$PD(Fr_{i,j}) = \sum_{v \in Fr_{i,j}} \frac{M_v}{d_{v,j}^2} \tag{30}$$

It models the value of the gravitational field induced by the fragment $\boldsymbol{Fr_{i,j}}$ in the point $\boldsymbol{j}$. Values given by eq 30 are collected in the matrix **RGsDdM_p/d2S** while averaged values are considered in **RGfDdM_p/d2S** and **RGjDdM_p/d2S** matrices.

Non-Uniform Field Gravity

Let the property $\boldsymbol{\Phi = M}$, descriptor $\boldsymbol{\Omega = p/d2}$, superposition $\boldsymbol{\Psi = S}$ and *dense geometrical* model. The *distance* (vs. $\boldsymbol{j}$) of the *center of equivalent fragmental gravity* of $\boldsymbol{Fr_{i,j}}$ is:

$$\boldsymbol{d_{E,CP_i,j}} = \sqrt{\left(\sum_{v \in Fr_{i,j}} \boldsymbol{M_v}\right) \Bigg/ \left(\left(\sum_{v \in Fr_{i,j}} \frac{\boldsymbol{M_v}}{\boldsymbol{d_{v,j}^2}} \cdot \frac{\vec{\boldsymbol{d}}_{\boldsymbol{v,j}}}{\boldsymbol{d_{v,j}}}\right) \cdot \left(\sum_{v \in Fr_{i,j}} \frac{\boldsymbol{M_v}}{\boldsymbol{d_{v,j}^2}} \cdot \frac{\vec{\boldsymbol{d}}_{\boldsymbol{v,j}}}{\boldsymbol{d_{v,j}}}\right)\right)^{\frac{1}{2}}} \qquad (31)$$

It models the distance at which a point mass equal to the fragment mass $\sum_{v \in Fr_{i,j}} \boldsymbol{M_v}$ should be located vs. *j* such that the gravitational field induced by $\boldsymbol{Fr_{i,jj}}$ in $\boldsymbol{j}$ be equal to the field induced by all atoms of the fragment. The associated matrix is of the form **DGcDdM_p/d2S**.

Uniform Electrostatic field

Let the property $\boldsymbol{\Phi = P}$ ($\boldsymbol{Q_P}$ implicitly, in the Cluj Program), descriptor $\boldsymbol{\Omega = p/d2}$, superposition $\boldsymbol{\Psi = S}$ and *rare geometrical* model. The uniform electrostatic field descriptor of $\boldsymbol{Fr_{i,j}}$ is:

$$\boldsymbol{PD(Fr_{i,j})} = \sum_{v \in Fr_{i,j}} \frac{\boldsymbol{Q_{P_v}}}{\boldsymbol{d_{v,j}^2}} \qquad (32)$$

It models the value of electrostatic field induced by the fragment in $\boldsymbol{j}$**.** The property matrix is of the form: **RGcDdP_p/d2S**.

Non-Uniform Electrostatic Field

For the property $\boldsymbol{\Phi = P}$ ($\boldsymbol{Q_P}$ implicitly), descriptor $\boldsymbol{\Omega = p/d2}$, superposition $\boldsymbol{\Psi = S}$ and *dense geometrical* model, the *distance* (vs. $\boldsymbol{j}$) of the *center of equivalent electrostatic field* of $\boldsymbol{Fr_{i,j}}$ is:

$$d_{E,CP_i,j} = \sqrt{\sum_{v \in Fr_{i,j}} Q_{Pv} \Big/ \left(\left(\sum_{v \in Fr_{i,j}} \frac{Q_{Pv}}{d_{v,j}^2} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}} \right) \cdot \left(\sum_{v \in Fr_{i,j}} \frac{Q_{Pv}}{d_{v,j}^2} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}} \right) \right)^{\frac{1}{2}}} \quad (33)$$

It models the distance at which a point charge equal to the fragment charge $\sum_{v \in Fr_{i,j}} Q_{P_v}$ be located vs. ***j*** such that the electrostatic field induced by it in ***j*** be equal to the field induced by the all atoms of the fragment. The associated matrix is of the form: **DGcDdP_p/d2S**.

Uniform Field Gravitational Potential

It is obtained for the property $\Phi = M$, descriptor $\Omega = p/d$, superposition $\Psi = S$ and *rare geometrical* model. The property descriptor of $Fr_{i,j}$ is:

$$PD(Fr_{i,j}) = \sum_{v \in Fr_{i,j}} \frac{M_v}{d_{v,j}} \quad (34)$$

It models the value of the gravitational potential induced by the fragment in ***j***. The property matrix is of the form: **RGcDdM_p/d_S**.

Non-*Uniform Field-Type Gravitational Potential*

For the property $\Phi = M$; descriptor $\Omega = p/d$; superposition $\Psi = S$; dense *geometrical model*, the *distance* (vs. ***j***) of the *center of equivalent fragmental gravity* of $Fr_{i,j}$ is:

$$d_{E,CP_i,j} = \sqrt{\left(\sum_{v \in Fr_{i,j}} M_v \right) \Big/ \left(\left(\sum_{v \in Fr_{i,j}} \frac{M_v}{d_{v,j}} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}} \right) \cdot \left(\sum_{v \in Fr_{i,j}} \frac{M_v}{d_{v,j}} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}} \right) \right)^{\frac{1}{2}}} \quad (35)$$

It models the distance at which a point mass equal to the fragment mass ($\sum_{v \in Fr_{i,j}} M_v$) should be located vs. ***j*** such that the gravitational potential induced by it in ***j*** be equal to the potential induced by the all atoms of the fragment. The associated matrix is of the form **DGcDdM_p/d_S**.

Uniform Field Coulombian Potential

It is obtained for the property $\Phi = P$ (Q_p implicitly), descriptor $\Omega = p/d$, superposition $\Psi = S$ and *rare geometrical* model. The electrostatic potential descriptor of $Fr_{i,j}$ is:

$$PD(Fr_{i,j}) = \sum_{v \in Fr_{i,j}} \frac{Q_{Pv}}{d_{v,j}} \tag{36}$$

It models the value of the electrostatic potential induced by the fragment in ***j***. The property matrix is of the form: **RGcDdP_p/d_S**.

Non-Uniform Field Electrostatic Potential

For the property $\Phi = P$ (Q_P implicitly); descriptor $\Omega = p/d$; superposition $\Psi = S$ and *dense geometrical* model, the *distance* (vs. ***j***) of the *center of equivalent electrostatic potential* of $Fr_{i,j}$ is:

$$d_{E,CP_i,j} = \sqrt{\left(\sum_{v \in Fr_{i,j}} Q_{Pv}\right) \Bigg/ \left(\left(\sum_{v \in Fr_{i,j}} \frac{Q_{Pv}}{d_{v,j}} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}}\right) \cdot \left(\sum_{v \in Fr_{i,j}} \frac{Q_{Pv}}{d_{v,j}} \cdot \frac{\vec{d}_{v,j}}{d_{v,j}}\right)\right)^{\frac{1}{2}}} \tag{37}$$

It models the distance at which a point charge equal to the fragment charge ($\sum_{v \in Fr_{i,j}} Q_{Pv}$) should be located vs. ***j*** such that the electrostatic potential induced by it in *j* be equal to the potential induced by all the atoms of the fragment. The associated matrix is of the form **DGcDdP_p/d_S**.

In all the above models, *j* appears as a *virtual probe atom* . In the opposite to the CoMFA approach, whose descriptors are calculated as interactions of the molecule with external grid probe atoms, our approach makes use of internal probe atoms: the property of fragment $Fr_{i,j}$ is viewed as the interaction of atoms forming the fragment $Fr_{i,j}$ with the atom ***j*** (with no chemical identity, however). Other particular fragmental property models the reader can find in ref.[19]

Fragmental Property Matrices

The fragmental property matrices are square matrices of the order ***N*** (i.e. the number of non-hydrogen atoms in the molecule). The non-diagonal entries in such matrices are fragmental properties corresponding to any pair of vertices (***i,j***) by a chosen model.

In case of Cluj criteria, the fragmentation can supply more than one maximal fragment for the pair (***i,j***). In such a case, the matrix entry is the arithmetic mean of the individual values.

Thus, if $i,j \in V(G)$, $i \neq j$ and $P_{i,j} = \{ p^1_{i,j}, p^2_{i,j}, ..., p^k_{i,j} \}$ paths joining i and j, then cf. ***CJ*** or ***CF*** definition (eqs 1-3), the fragments $Fr^1_{i,j}, Fr^2_{i,j}, ..., Fr^k_{i,j}$ are generated. Let m be the number of maximal fragments among all the k fragments, $1 \leq m \leq k$, and let $\sigma_1, \ldots, \sigma_m$ be the index for the maximal fragments. By applying any of the above models (eqs 21-29) for all the m maximal fragments we obtain m values (for example, by eq 29):

$$PD(Fr^{\sigma_1}_{i,j}),\ PD(Fr^{\sigma_2}_{i,j}),\ \ldots,\ PD(Fr^{\sigma_m}_{i,j})$$

and consequently, the matrix entry associated to the pair (i,j) is the mean value:

$$PD_{i,j} = \frac{\sum_{t=1}^{m} PD(Fr^{\sigma_t}_{i,j})}{m} \tag{38}$$

The resulting matrices are in general *unsymmetric* but they can be symmetrized (see eqs 4, 5). The symbols for the fragmental property matrices will be detailed below.

Operators for Calculating Fragmental Property Indices

Fragmental property indices are calculated at any fragmental property matrices above discussed, by applying four types of index operators: ***P_***, ***P2***, ***E_***, ***E2*** according to the relations:

$$P\_(M) = \tfrac{1}{2}\sum\sum [\mathbf{M}]_{i,j} \quad ; \quad P2(M) = \tfrac{1}{2}\sum\sum [\mathbf{M}]_{i,j}[\mathbf{M}]_{j,i};$$

$$E\_(M) = \tfrac{1}{2}\sum\sum [\mathbf{M}]_{i,j}[\mathbf{A}]_{i,j} \quad ; \quad E2(M) = \tfrac{1}{2}\sum\sum [\mathbf{M}]_{i,j}[\mathbf{M}]_{j,i}[\mathbf{A}]_{i,j} \tag{39}$$

where **M** is any property matrix, symmetric or unsymmetric.

Name of the Fragmental Property Matrices and Indices

The name of *fragmental property matrices* is of the general form:

ABcDdEfffffG (40)

where:

A ∈ {**D**, **R**}; D = Dense; R = Rare;

B ∈ {**T**, **G**}; **T** = Topological; ***G*** = Geometric;

c ∈ {**f**, **j**, **s**}; ***f*** = ***CF***-type; ***j*** = ***CJ***-type; ***s*** = ***Sz***-type;

Dd ∈ {**Di**, **De**}; ***Di*** = Distance; ***De*** = Detour;

E ∈ Φ (i.e. **E**∈{**M**, **E**, **C**, **P**} where ***M*** = *mass*; ***E*** = *electronegativity*; ***C*** = *cardinality*; ***P*** = other atomic property - implicitly, *partial charge*; explicitly, a property given by manual input);

fffff ∈ Ω (i.e. **fffff**∈{**__p__**, **_1/p_**, **__d__**, **_1/d_**, **_p.d_**, **_p/d_**, **_p/d2**, **p2/d2**}

G ∈ Ψ (i.e. **G**∈{**S**, **P**, **A**, **G**, **H**} with the known meaning (see above).

The name of *fragmental property indices* is of the general form:

ABcDdEfffffGii (41)

where:

ii ∈ {***P_***, ***P2***, ***E_***, ***E2***} with the known meaning (eq 39).

If an operator, such as ***f*(*x*) = 1/*x*** (inverse operator) or ***f*(*x*)** = **ln(*x*)**, is applied the indices are labeled as follows:

ln*ABcDdEfffffGii* := ln(*ABcDdEfffffGii*);

$$\textbf{\textit{1/ABcDdEfffffGii}} := \frac{1}{\textbf{\textit{ABcDdEfffffGii}}} \quad (42)$$

For example, index ***lnDGfDeM__p__SP_*** is the logarithm of index ***DGfDeM__p__SP_*** computed on the property matrix **DGfDeM__p__S**. The model used is dense, geometric, on fragment of type ***CF***, with the cutting path being detour. The chosen property is the mass, the descriptor for property is even the property (mass) and the sum operator counts the vertex descriptors.

Model Degeneration and Computational Features

The degeneration in the above models may occur in cases when the values of property are not diverse enough, like is case of cardinality (see *Fragmental Numbers,* this Section). Another degeneration is in the case: **RTfDiC__p__H = TfDiC_1/p_S**.

The fragmental analysis was made by the aid of four original 16-bit windows computer programs. First program, ***ClujTheor*** calculates *topological descriptors* of *Cluj* and *Szeged* type and generates the fragments for the molecules. Second, ***ClujProp*** calculates the *fragmental properties*. The third one, ***StatMon*** makes *monovariate regressions* and sorts indices according to the correlation score. The forth program, ***StatQ*** performs *multi-linear regression* (2-variate, 4-variate, etc.) and saves on disk the best couples of indices. The total number of indices is given by: 2560 × 3(i.e. *x*, ***ln*(*x*)**, 1/***x***) × 3(i.e. the criteria: ***CJ***, ***CF***, ***Sz***) × 2 (i.e. the path criteria: ***Di***, ***De***) = 46080. Note that in most cases, the degeneration induced by

property values and operators lead to a total number of distinct indices around 19,000. In bivariate regression, the first 2^{14}-1=16383 indices recording the best scores in monovariate regression are considered.

Correlating Studies

The mathematical models of a certain property are performed by MLR (Multiple Linear Regression) and/or CNN (Computational Neural Networks).[29-39] In any case, the model is built by using a training set of structures that provides a calibration equation. Next, it is validated by a cross-validation procedure and also by using an external prediction set. In the following, the MLR procedure is presented.

MLR, for *n* observations and *m* independent variables is represented by

$$Y_i = b_0 + \sum_j^m b_{ij} X_{ij} \tag{43}$$

or, in matrix form as

$$\mathbf{Y} = \mathbf{bX} \tag{44}$$

where **Y** is the $\boldsymbol{n}$ x 1 vector of responses, **X** is an $\boldsymbol{n}$ x ($\boldsymbol{m}$ + 1) matrix of independent variables and **b** is the ($\boldsymbol{m}$ + 1) x 1 vector of regression coefficients. The regression coefficients can be determined by the least-squares solution of (44)

$$\mathbf{b} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y} \tag{45}$$

With **b** calculated, eq 44 can be used for estimating the chosen property for other chemical structures.

To avoid the chance correlations, it is recommended that the number of descriptors submitted to regression be less than 60 % of the number of observations in the training set.[40]

Set 1. Substituted 3-(Phthalimidoalkyl)-pyrazolin-5-ones.

We tested the correlating ability of ***FPIF*** on a set of 17 molecular structures from the class of substituted 3-(Phthalimidoalkyl)-pyrazolin-5-ones.

The molecular structure of the selected chemicals is given in Figures 2.(a, b). It was performed by using the MM+ (for 3D-geometries) and semiempirical AM1 (for partial charge calculation) procedures of the HyperChem Program (HyperCube Inc.).

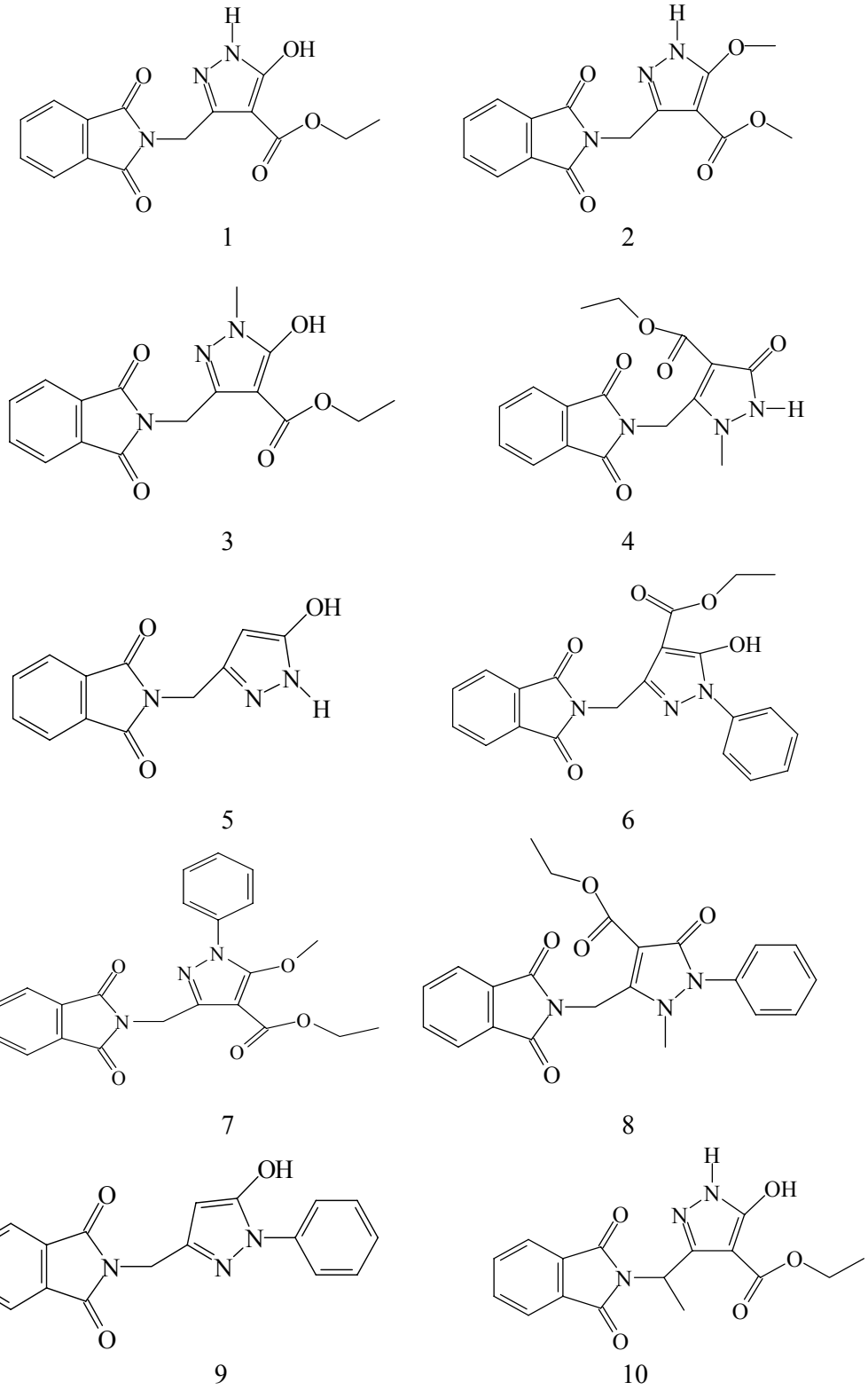
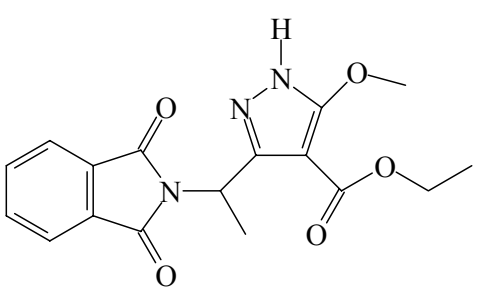
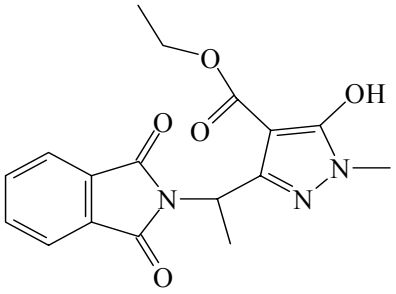


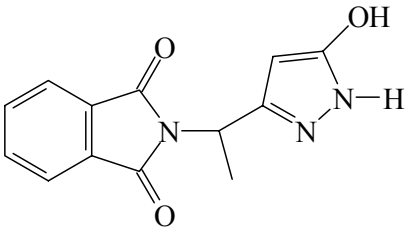
Figure 2.a. Structure of 17 substituted 3-(Phthalimidoalkyl)-pyrazolin-5-ones; molecules 1 to 10.



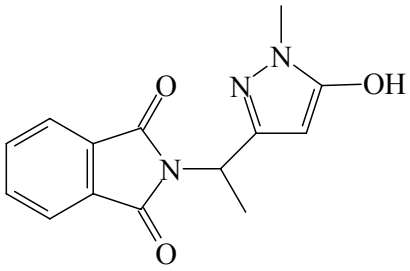
11



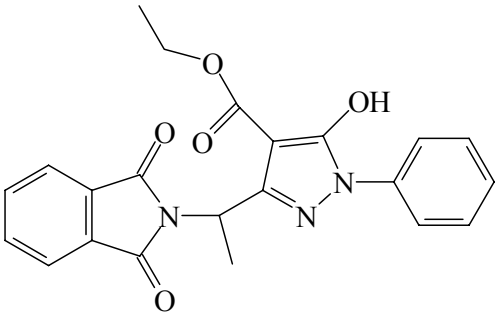
12



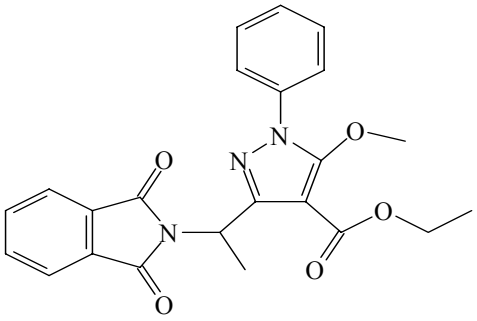
13



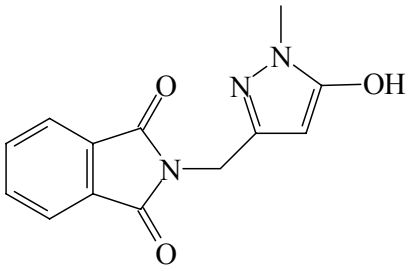
14



15



16



17

Figure 2.b. Structure of 17 substituted 3-(Phthalimidoalkyl)-pyrazolin-5-ones; molecules 11 to 17.

The modeled properties were the sum of one-electron energy calculated at the Extended-Huckel level and the inhibitory activity (in %) of a solution of 0.05 mg/ml pyrazolin-5-one on *Lepidium sativum L*. (Cresson). The data are listed in Table 1.

Table 1. The Sum of One-Electron Energy Calculated at Single Point Semi-Empirical Extended-Huckel and the Inhibitory Activity on *Lepidium sativum L*. (Cresson) for 17 Substituted 3-(Pthalimidoalkyl)-Pyrazolin-5-Ones*

Molecule no.	Energy (kcal/mol)	Inhibition (%)
1	50978.19	28.4
8	64751.09	65.2
7	64752.65	49.4
6	62330.33	68.3
5	38604.68	14.3
4	53416.95	27.7
3	53441.43	30.4
2	51000.36	28
17	41057.46	15.1
16	67104.64	50.6
15	64701.39	71.7
14	43473.37	18.2
13	41020.54	12.2
12	55832.12	32.6
11	55729.99	28.9
10	53424.19	29.3
9	50012.42	46.9

* Values of inhibition are taken from ref. [41]

Monovariate Regression for Energy

For the first six best indices in monovariate correlation, the equation is:

$$\text{Predicted energy} = b_0 + b_1 \cdot \ln Index \tag{46}$$

The indices are listed in Table 2 along with the Pearson correlation index R and the regression coefficients.

Table 2. The Bests Six Correlations of Energy in Monovariate and Divariate Regression

Index No.	Index Name	R	b_0	b_1
1	**lnDGjDeE_p/d2PE_**	**0.99973**	5370	3760
2	DTjDeEp2/d2SE_	0.99964	8802.3	138
3	lnDGjDeE_p/d2PE2	0.99964	1289.5	3671.1
4	DTjDeMp2/d2SE_	0.99939	9188.7	922.34
5	lnDGsDiM_p/d2PE_	0.99938	-59041	28397
6	lnRTjDeE_p*d_PP_	0.99927	3044	1938.6
1 2	lnDGjDeE_p/d2PE_ DTjDeEp2/d2SE_	0.999688	8771.8	-21.648 138.77
1 10175	lnDGjDeE_p/d2PE_ lnDTsDeE_p/d_PP2	0.999871	58810	3678 -6593.1
4 4315	DTjDeMp2/d2SE_ DTfDiE_p/d_AP2	0.999902	13056	1108.8 -95.598
34 5947	RTsDiM_p/d2GP2 DTjDeEp2/d2AP_	0.999969	-1193	1674.3 -41.168
492 1698	RTjDeM_p/d2SP_ 1/RTsDeM_p/d2AE2	0.999974	58267	46.095 -686800
492 1737	**RTjDeM_p/d2SP_ 1/RTsDeM_p/d2AP2**	**0.999981**	56222	47.864 -711240

The best single variable QSPR (boldface in Table 2) was

$$\text{Predicted energy} = 5370 - 3760 * \ln DGjDeE\_p/d^2PE\_ \qquad (47)$$

$$R = 0.99973;\ n = 17$$

This correlation could be satisfactory but usually a molecular property shows more than one dimension dependency. For this reason, we performed the bivariate regression.

Bivariate Regression for Energy

The first 16383 indices, labeled in decreasing order of their score in monovariate regression, are submitted for bivariate correlation. A procedure for finding subsets of optimal even number descriptors was developed. It is a simple, iterative technique that eludes the investigation of all possible descriptor combinations and reduces the time for drawing the best property model. More details will be presented in a future paper.

Here, the bivariate correlation for six pairs of indices is exemplified. The pairs are: (1, 2); (1, 10175); (4, 4315); (34, 5947); (492, 1698) and (492, 1737). The first two pairs are taken to show that the first scored index in monovariate regression does not provide the best

bivariate correlation. Selection of the pairs of indices for bivariate correlation must be done by traversing the whole pool (1...16383). For additional descriptors, our procedure for optimum descriptor selection avoid the mining of all possible index combinations.

The best bivariate score was provided by the pair (492, 1737) (Figure 3):

$$\text{Predicted energy} = 56221.885 + 47.864*\text{RTjDeM\_p/d2SP\_} - 711240.703*1/\text{RTsDeM\_p/d2AP2} \quad (48)$$

$$R = 0.99998;\ s = 57.40;\ n = 17$$

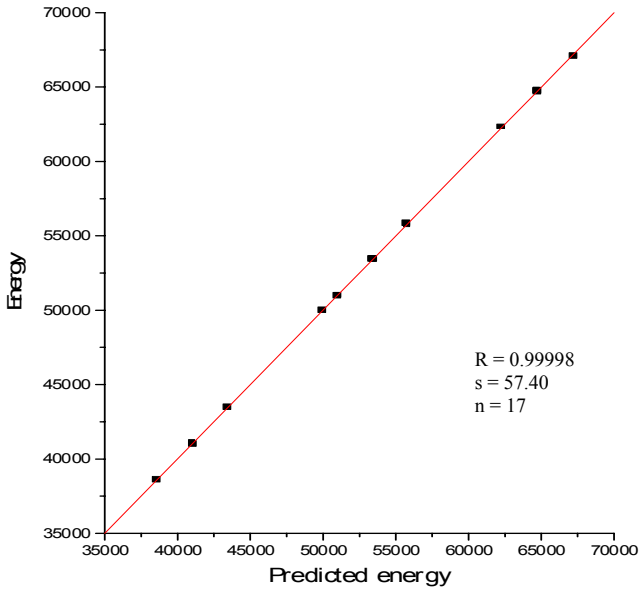


Figure 3. The plot: energy vs predicted energy cf eq 48

An insight in Table 2 reveals that the best models (i.e. those showing R > 0.9999) show a dependency of this energy by the molecular topology (topological models) and the nature of atoms (mass and electronegativity).

Monovariate Regression for Inhibition

For the first six best indices in monovariate regression, the equation of the model is:

$$\text{Predicted inhibition} = b_0 + b_1 \cdot Index \quad (49)$$

for which the indices and statistics are given in Table 3.

Table 3. The Best Six Correlations of Inhibition in Monovariate and Divariate Regression

Index No.	Index Name	R	b_0	b_1
1	lnDGsDeC_1/p_SE_	0.95389	-336.76	96.378
2	1/DGsDeC_1/p_SE_	-0.95231	137.01	-4754.8
3	lnDTjDeE_p*d_HE_	0.95174	135.80	-493.02
4	1/DTjDeE_p*d_HE_	-0.95169	175.92	-6826.3
5	1/DTjDeEp2/d2AE_	-0.95144	186.07	-19546
6	DGsDeC_1/p_SE_	0.95126	-56.53	1.9019
1 2	lnDGsDeC_1/p_SE_ 1/DGsDeC_1/p_SE_	0.955154	-129.47	54.37 -2121.1
1 1369	lnDGsDeC_1/p_SE_ 1/RGsDeMp2/d2SE2	0.96935	-84.14	47.481 -3743200
2 13227	1/DGsDeC_1/p_SE_ lnRGsDeE_p/d_AE2	0.983967	-452.74	-4279.8 118.74
18 16842	DTjDeEp2/d2AE_ RGjDeP_p/d_GP_	0.988316	121.21	1.076 -1.5194
37 11362	DTjDeE_p/d_AE_ lnDGjDeP_p/d_PE2	0.990564	-73.183	2.1644 -4.1769
4304 7649	DTsDiM_p*d_HP_ DGjDeE_p/d2SE2	0.99268	-26.846	1.5619 -1.7043

The best monovariate QSAR was

Predicted inhibition = -336.760 + 96.378* lnDGsDeC_1/p_SE_ (50)

R = 0.9539; n = 17

which is, of course, not satisfactory, despite in ref.[41] a value of R = 0.92 was reported. Thus, we performed the bivariate regression.

Bivariate Regression for Inhibition

Six pairs of indices are considered here for bivariate correlation: (1, 2); (1, 1369); (2, 13227); (18, 16842); (37, 11362) and (4304, 7649).

As in the case of energy, the best scored index in monovariate correlation is not present in the pair of best bivariate correlation.

The best bivariate score was done by the pair (4304, 7649):

Predicted inhibition = -26.846 + 1.562* DTsDiM_p*d_HP_-1.704 * DGjDeE_p/d2SE2

R = 0.9927; s = 2.374; n = 17. (51)

Figure 4. illustrates the plot of inhibition vs predicted inhibition cf eq 51.

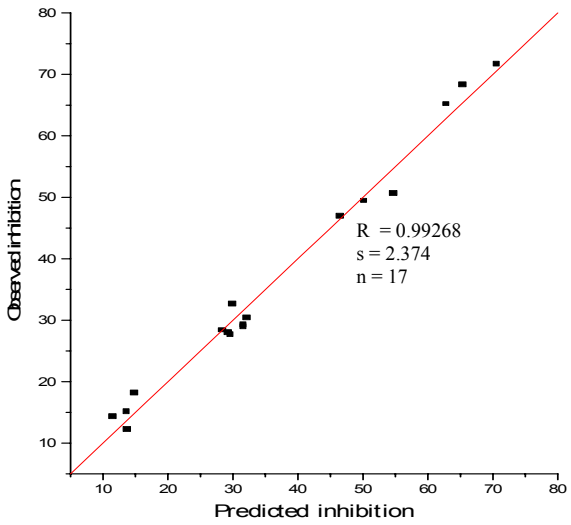


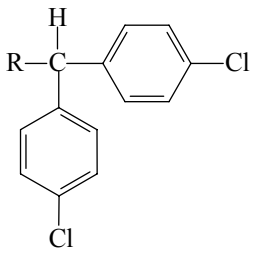
Figure 4. The plot: inhibition vs predicted inhibition cf eq 51

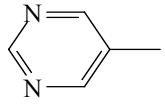
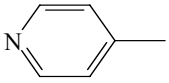
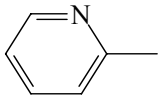
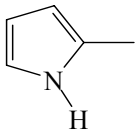
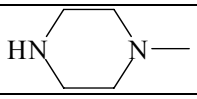
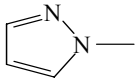
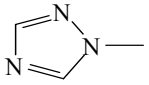
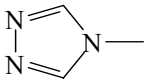
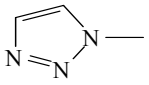
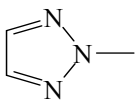
The constant high correlation (see Table 3) between the best indices and the mitodepressive activity on *Lepidium Savitium L. (Cresson)* demonstrate ability of this family of indices to estimate the biological activity of the considered set of chemical structures. The models with R > 0.983 suggest that the mitodepressive activity on *Lepidium Savitium L. (Cresson)* is dependent both on the geometric and topological features of molecules, the nature of atoms (mass and electronegativity) and the electrostatic field of atoms induced by their partial charges.

Set 2. Aromatase Inhibitors.

A set of substituted dichlorodiphenyls (4, 4'-dichlorodiphenyl-methanes) inhibitors of aromatase[42] were considered. Enzymatic aromatization of androgens is involved in the biosynthesis of estrogens, and consequently in the estrogen-dependent diseases.

Table 4. Dichlorodiphenyl Methanes Aromatase Inhibitors.



No.	R	-log EC_{50} obs
1		7.43
2		8.03
3		8.06
4		5.70
5		5.71
6		5.30
7		5.30
8		6.80
9		5.30
10		7.26

For modeling the inhibition, the authors[42] used two dipole moment related descriptors. We modeled the inhibition in monovariate regression but no satisfactory correlation (R^2 around 0.828) was found. In divariate regression, the correlation improved (Figure 5).

Predicted inhibition = 6.177 + 0.513*lnRTjDiP__p__HP2 - 0.071*1/DGjDeP_1/p_SP2

$R^2 = 0.9716$; s = 0.205; n = 10 (52)

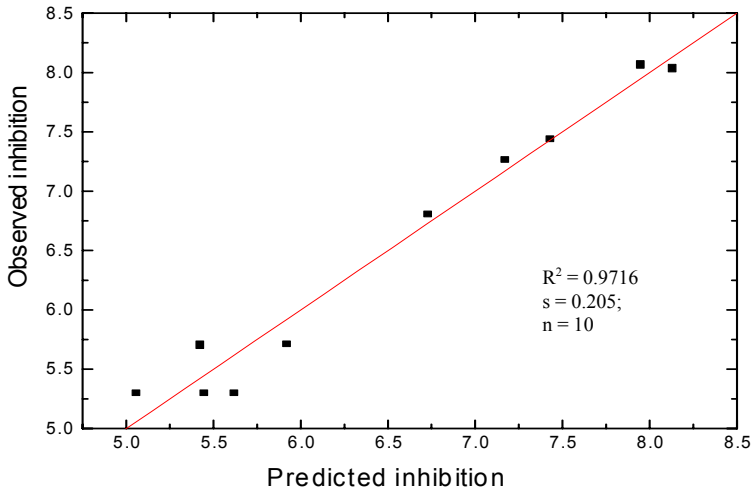


Figure 5. The plot of predicted vs observed inhibition of aromatase.

The best reported[42] correlation for this subset was: R^2 = 0,89; s = 0.44. In our model, both the topology and geometry (see the indices in eq 52) are important in modeling the aromatase inhibition by dichlorodiphenyl methanes.

Set 3. N-containing compounds

A set of 90 N-containing compounds (Table 5) of industrial importance was taken from the paper.[43] The tested property was the normal boiling point, B.P. The authors modeled this property by using four categories of molecular descriptors: topological, geometric, electronic and charged-partial surface area descriptors (CPSA).[44,45] The nitrogen-containig compounds were problematic in modeling a diverse set of organic chemicals, so that the authors excluded such compounds from their initial model.

The best found MLR model involved ten descriptors (1. dipole moment; 2. partial negative surface area; 3. relative negative charge; 4. relative negative charged surface area; 5. number of aromatic bonds; 6. path 2 molecular connectivity index; 7. cluster 3 valence connectivity index; 8. sum of all path weights from heteroatoms; 9. surface area of donatable hydrogens and 10. charge of donatable hydrogens) and showed the following statistics: n = 90

Table 5. Nitrogen-Containing Compounds and Their Boiling Points.

No.	Compound	BP	No.	Compound	BP
1.	2-ethylpyridine	422.2	46.	n-tetradecylamine	564.5
2.	2-ethylpiperidine	416.2	47.	acridine	619.2
3.	1-ethylpiperidine	404.2	48.	tri-n-butylamine	487.2
4.	2,2-dimethyl-1,3-diaminomethane	426.2	49.	n-dodecylamine	532.4
5.	N,N-dimethyl-1,3-diaminomethane	418.2	50.	diamylamine	476.1
6.	3,3-dimethylpiperidine	410.2	51.	tripropylamine	429.7
7.	p-fluorobenzylamine	456.2	52.	n-nonylamine	475.4
8.	cianogene	252	53.	quinoline	510.8
9.	m-bromoaniline	524.2	54.	acetonitrile	354.8
10.	o-bromoaniline	502.2	55.	isoquinoline	516.4
11.	N-ethylbutylamine	381.2	56.	n-octylamine	452.8
12.	triethylamine	362	57.	indole	526.1
13.	N,N-diethylamină	337.2	58.	n-heptylamine	430.1
14.	o-nitrotoluene	498.2	59.	p-nitrotoluene	511.7
15.	nitrocyclopentane	453.2	60.	benzonitrile	464.1
16.	N-alylaniline	492.2	61.	3-nitrobenzotrifluoride	475.9
17.	ethylamine	289.7	62.	di-n-propilamine	382
18.	p-nitrophenole	552.2	63.	nitrohexane	436.8
19.	cyclopentylamine	380.2	64.	phenilhidrazine	516.7
20.	2-methylbutylamine	368.7	65.	methylamine	266.8
21.	N-methylbutylamine	364.2	66.	3-methylpyridine	417.3
22.	benzylamine	457.7	67.	aniline	457.2
23.	p-methoxyaniline	514.7	68.	p-chloroaniline	503.7
24.	m-methoxyaniline	524.2	69.	m-chloroaniline	501.7
25.	o-methoxyaniline	498.2	70.	n-pentylamine	377.6
26.	t-pentylamine	350.2	71.	isobutylamine	340.9
27.	dimethylamine	280	72.	diethylamine	328.6
28.	1-(2-aminoethyl)-piperidine	459.2	73.	tert-butylamine	317.5
29.	1-(2-aminoethyl)-piperidine	493.2	74.	n-butylamine	350.6
30.	9-methyl carbazole	616.8	75.	pirolidine	359.7
31.	carbazole	627.8	76.	nitromethane	374.4
32.	4-methylaniline	473.6	77.	isobutyronitrile	376.8
33.	3-methylaniline	476.5	78.	n-butyronitrile	390.8
34.	2-methylaniline	473.5	79.	cis-crotonitrile	380.6
35.	2-propylamine	304.9	80.	trimethylamine	276
36.	1-naphtylamine	573.8	81.	2-nitropropane	393.4
37.	nitroethane	387	82.	1-nitropropane	404.3
38.	piperidine	376.4	83.	propionitrile	370.5
39.	4-methylpyridine	418.5	84.	acrylonitrile	350.5
40.	2-methylpyridine	402.5	85.	N-methylhexylamine	414.2
41.	pyridine	388.4	86.	n-heptylamine	428.2
42.	pyrole	402.9	87.	N-tert-butylisopropylamine	371.2
43.	2-butylamine	335.9	88.	2-aminoheptane	416.2
44.	triamylamine	516.2	89.	malononitrile	491.5
45.	ethylenimine	329	90.	hydrogen cyanide	298.8

compounds; R = 0.990; s = 10.7 K. The largest pairwise R value of descriptors was 0.83. The modeling was performed by the ADAPT system.[46]

Our aim was to verify the quality of our property descriptors exactly in the same conditions as given in ref.[43] Thus, we extracted from the initial set of 104 N-containing compounds the same subset of 90 structures.

Molecular geometries and partial charges were calculated by the semiempirical AM1 method. The set of 19350 descriptors were reduced to 16383 after the monovariate regression.[19] Our procedure for finding the optimal subset of descriptors led to a subset of 72 descriptors. The best scores in ten variate regression for the set of 90 compounds of Table 5 are listed in Table 6.

Table 6. The Best Multivariate Regressions for the 90 Structures of Table 5.

No	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	R
1	7	4959									0.944600
2	4	4797									0.945140
3	5	4959									0.948670
4	5	4959	14250	14607							0.962734
5	5	4959	10990	9671							0.962809
6	5	4959	10990	7206							0.967113
7	5	4959	10990	9671	16	3320					0.974787
8	5	4959	10990	9671	3320	6422					0.976078
9	5	4959	10990	9671	3528	6422					0.976574
10	5	4959	10990	9671	3528	6422	7256	16148			0.979844
11	5	4959	10990	9671	3528	6422	16148	6895			0.979862
12	5	4959	10990	9671	3528	6422	6895	15789			0.980012
13	5	4959	10990	9671	3528	6422	6895	15789	16158	16225	0.983581
14	5	4959	10990	9671	3528	6422	16158	16225	15060	15789	0.984335
12	6	6895	4	16275	963	841	163	13920	1	4727	0.985431

The best model was:

BP_{calc} = 225.441 - 59.627*lnDTsDiP_p/d2SE2 + 316.627*RTsDiPp2/d2AE_ + (53)
1.124*DGfDePp2/d2PP_ - 1729.562*1/DTsDiE_p*d_HE2 -
0.010*1/DTsDePp2/d2SP2 - 49.623*1/DGsDeP_p*d_HE_ +
8.846*lnDGjDiPp2/d2GP_ - 4.698*1/RGjDeP_p*d_GP_ -
12.188*lnDGjDeP_p/d_HP_ + 33.597*DGjDeE__p__SE2

R = 0.98543; s = 13.149; n = 90

The plot corresponding to eq (53) is given in Figure 6.

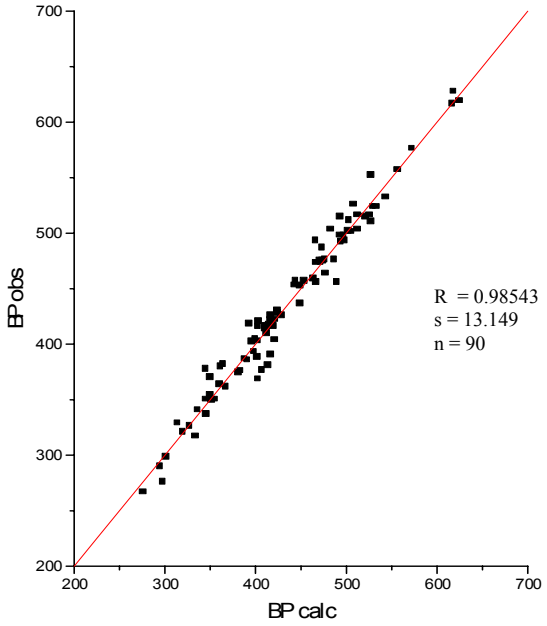


Figure 6. The plot of calculated vs observed normal boiling points
(the set of Table 5)

Our result is slightly lower (s = 13.149 K) than that reported in ref.[43] (s = 10.7 K). It is possible to further improve the model by mining the whole descriptor pool not only within the limits of a heuristic procedure. Another possibility is to use different training subset selection and outlier elimination. Such procedures will be reported in a future paper.

Set 4. Nitrophenols.

A set of 25 nitrophenols[47] showing herbicidal activity (Table 7) was considered for correlation with the Cluj Property indices. Nitrophenols are known to inhibit the electronic flux of photosynthesis.

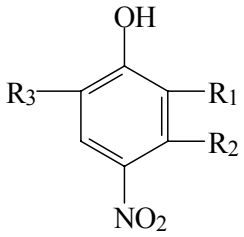


Table 7. Nitrophenols and Their Herbicidal Activity

No	R_1	R_2	R_3	pI_{50}
1	H	methyl	methyl	3.3
2	H	methyl	isopropyl	4.1
3	H	H	t-butyl	5.7
4	H	H	phenyl	4.35
5	H	H	cyclohexyl	4.85
6	Cl	methyl	methyl	4.89
7	Cl	methyl	isopropyl	6.07
8	Cl	H	t-butyl	6.88
9	Cl	H	phenyl	6.45
10	Cl	H	cyclohexyl	6.52
11	Br	methyl	methyl	5.25
12	Br	methyl	isopropyl	6.70
13	Br	H	t-butyl	6.15
14	Br	H	phenyl	6.52
15	Br	H	cyclohexyl	6.75
16	I	methyl	methyl	6.24
17	I	methyl	isopropyl	6.70
18	I	H	t-butyl	7.03
19	I	H	phenyl	6.86
20	I	H	cyclohexyl	6.65
21	NO_2	H	H	3.00
22	NO_2	H	methyl	3.70
23	NO_2	H	s-butyl	5.10
24	NO_2	H	t-butyl	5.79
25	NO_2	H	cyclohexyl	6.05

Table 8 list the best scores of correlation in decreasing order. From this table it can be seen that the monovariate and divariate regression are not satisfactory. Additional variables are needed for good statistics (entries 6-13).

Table 8. Mono- and Multivariate Regression for Nitrophenols

No	X_1	X_2	X_3	X_4	X_5	X_6	R
1	1						0.94991
2	6	155					0.97089
3	7	174					0.96974
4	11	10620					0.96654
5	5	260					0.96617
6	13028	15806	91	15636			0.99078
7	13028	15806	12	15398			0.99013
8	12	15806	13891	15749			0.99067
9	7	13028	382	14214			0.98932
10	13028	15806	12	15398	15228	15865	0.99850
11	12	15806	13891	15749	15648	16378	0.99674
12	7	13028	382	14214	13186	16282	0.99650
13	13028	15806	91	15636	14064	14943	0.99562

The best model is given in eq 54 (see also entry 10, Table 8):

Predicted activity = 8.062 - 0.003*RGsDeM_p/d2PE2 + 0.395*1/DTsDiP_p*d_HE_
- 0.000008*1/RTsDiPp2/d2HE2 - 229.564*1/DGjDeMp2/d2PE2
+ 0.003*RGjDiPp2/d2HP_ + 0.004*DTsDeP_p*d_HP2 (54)

R = 0.9985; s = 0.067; n = 25

The plot of the predicted vs observed herbicidal activity, cf. eq 54, is shown in Figure 7.

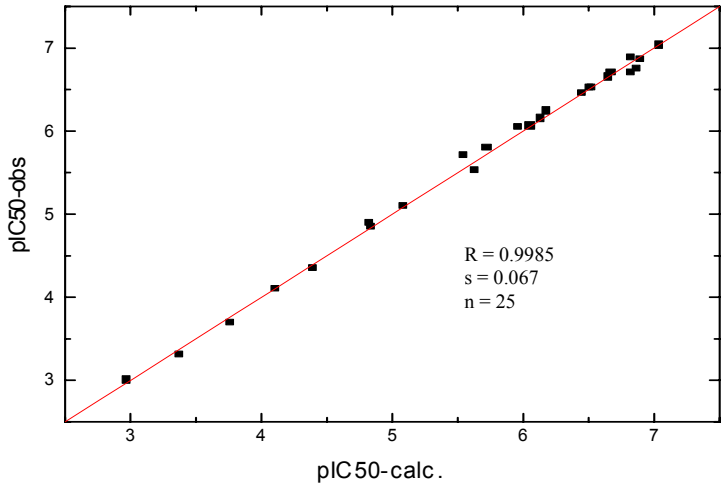


Figure 7. The plot of predicted vs. observed herbicidal activity

The descriptors involved in eq 54 show a rather low inter-correlation (Table 9). The average absolute value of the pairwise correlation coefficients was 0.2200.

Table 9. Intercorrellation of the indices in entry 10, Table 8.

	X_2	X_3	X_4	X_5	X_6
X_1	0.216489	0.697018	0.067367	0.121666	0.171049
X_2		0.137108	0.015049	0.060902	0.264608
X_3			0.201832	0.039716	0.269802
X_4				0.682383	0.22193
X_5					0.133726

Discussion

The fragmental property indices take into account the chemical nature of atoms (mass and electronegativity), various kinds of interactions between the fragments of molecules and the 3D geometry of molecular structures.

There exist an analogy between *CoMFA* and ***FPIF***: both of them calculate the interaction of a chemical structure (or substructure) with a *probe atom* in the 3D space. The property of fragment $\boldsymbol{Fr_{i,j}}$ is viewed as the interaction of atoms forming the fragment $\boldsymbol{Fr_{i,j}}$ with

the atom ***j*** . The major difference is that *CoMFA* uses external probe atoms (with defined chemical identity) whereas ***FPIF*** considers internal probe atoms with no chemical identity. Only the fragments (i.e. substructures) are chemically well defined.

Bivariate correlation with indices belonging to ***FPIF*** offer good quality models for quite diverse molecular properties such as the inhibition of mitodepressive activity on *Lepidium Savitium L.* ($R > 0.99$ - see set 1) and the aromatase inhibition (set 2) as well. The same is true for the sum of one-electron energy calculated at the Extended-Huckel level ($R > 0.9999$).

Multivariate regression provided good models for the boiling points of a very diverse set of N-containing organic molecules (set 3) or for the herbicidal activity (set 4).

Note that there is no causal relationship between descriptors and a chosen property, despite each descriptor encodes some aspects of intra- and/or intermolecular interactions. In case of a highly noncongeneric database, (the case of the set 3) more than one model, with equivalent accuracy are expected. Our results provided different vectorial descriptions that are, in some extent, isomorphic.

The occurrence of a certain descriptor type can be however, indicative, for a possible causal relation between the structure and the investigated property. In the set 4, the best model (eq 54) is described by the mass (M) for X_1 and partial charge (P) for the remainder descriptors. It is just the expected case: the herbicidal activity of nitrophenols (a congeneric set) is controlled by the acidity of the phenol group, which increases with decreased negative partial charge on the oxygen atom.

It appears that, in large ensembles of molecules, the correlation is not strongly dependent of the type of chemical description, which is not the case in the more compact sampling of congeneric sets. At the detailed scale, the observable properties strictly depend on the particular data set and require more specified description. It justifies the conclusion[48] that there exists different mapping behavior of the chemical space at different scales and which one is more suitable in a given problem remains at the latitude of the researcher.

The above presented results demonstrate the good correlating ability of ***FPIF***. It represents a promise for further **QSPR/QSAR** studies.

Acknowledgement. One author (M. D.) gratefully thanks to Professor A. Kerber for fruitful discussions in the period he visited the University of Bayreuth, Germany.

References

1. Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies, *J. Med. Chem*. **1964**, *7*, 395.
2. Gao, C.; Govind, R.; Tabak, H. H. Application of the group contribution method for predicting the toxicity of organic chemicals. *Environmental Toxicol. Chem.* **1992**, *11*, 631-636.
3. Kalivas, J. H.; Sutter, J. M.; Roberts, N. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry, *Anal. Chem.* **1989**, *61*, 2024-2030.
4. Kalivas, J. H. generalized simulated annealing for calibration sample selection from an existing set and orthoginization of undesigned experiments. *J. Chemometrics*, **1991**, *5,* 37-48.
5. Sutters, J. M.; Dixon, S. L.; Jurs, P. C. Automated descriptor selection for Quantitative Structure-Activity Relationships, using generalized simulated annealing, *J. Chem. Inf. Comput. Sci.* **1995**, 35, 77-84.
6. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267.
7. Lucasius, C. B.; Kateman, G. Undersanding and using genetic algorithms. Part 1. Concepts, properties and context. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 1.
8. Diudea, M.V. Cluj matrix $\mathbf{CJ_u}$: source of various graph descriptors, *Commun. Math. Comput. Chem. (MATCH),* **1997**, *35*, 169-183.
9. Diudea, M.V.; Minailiuc, O.; Katona, G.; Gutman, I. Szeged matrices and related numbers, *Commun. Math. Comput. Chem. (MATCH),* **1997**, *35*, 129-143.
10. Diudea, M.V. Cluj matrix invariants, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 300-305.
11. Diudea, M.V.; Pârv, B.; Topan, M.I. Derived Szeged and Cluj indices, *J. Serb. Chem. Soc*. **1997**, 62, 267-276.
12. Diudea, M.V.; Gutman, I. Wiener-type tpological indices, *Croat. Chem. Acta* **1998**, *71*, 21-51.
13. Kiss, A.A.; Katona, G.; Diudea, M.V. Szeged and Cluj matrices within the matrix operator $W_{(M1,M2,M3)}$ *Coll. Sci. Papers Fac. Sci. Kragujevac* **1997,** *19,* 95-107.
14. Gutman, I.; Diudea, M.V. Defining Cluj matrices and Cluj matrix invariants, *J. Serb. Chem. Soc.* **1998**, *63*, 497-504.
15. Diudea, M.V.; Parv, B.; Gutman, I. Detour-Cluj matrix and derived invariants, *J.Chem.Inf.Comput.Sci.* **1997**, *37*, 1101-1108.
16. Diudea, M.V.; Katona, G.; Lukovits, I.; Trinajstić, I. Detour and Cluj-detour indices, *Croat. Chem. Acta* **1998**, *71*, 459-471.

17. Minailiuc, O.; Katona, G.; Diudea, M.V.; Strunje, M., Graovac, A.; Gutman, I. Szeged fragmental indices, *Croat. Chem. Acta* **1998**, *71*, 473-488.
18. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge Univ. Press, Cambridge, **1985**.
19. Diudea, M. V.; Gutman, I.; Jäntschi, L. *Molecular Topology*, Nova Science, Huntington, New York, (in press).
20. Sears, F.W.; Zemansky, M.W.; Young, H.D. *University Physics*, fifth edition, *Addison – Wesley Publishing Company*, USA-Canada, **1976**.
21. Golender, V.; Vesterman, B.; Vorpagel, E. APEX-3D Expert system for drug design. *Network Science*, **1996**, *Jan*, http:\\ www.netsci.org/ Science/ Compchem/ feature09.html.
22. Rose, V.S.; Wood, J. Generalized cluster significance analysis and stepwise cluster significance analysis with conditional probabilities. *Quant. Struct.-Act. Relat.* **1998**, *17*, 348-356.
23. Young, H.D. *Statistical Treatment of Experimental Data*. McGraw-Hill, New York, **1962**.
24. Reif, F. *Fundamentals of Statisticals and Thermal Physics*, McGraw-Hill, Chap.1 New York, **1965**.
25. Diudea, M.V.; Silaghi-Dumitrescu, I.; Valence group electronegativity as a vertex discriminator. *Rev. Roumaine Chim.* **1989**, *34*, 1175-1182.
26. Diudea M.V.; Kacso I.E.; Topan M.I. Molecular topology. 18. A Qspr/Qsar study by using new valence group carbon-related electronegativities. *Rev. Roum. Chim.* **1996**, *41*, 141-157.
27. Crawford, F.S.Jr.; *Waves*, Berkeley Physics Course, Newton, Massachusetts, vol. 3, **1968.**
28. Ivanciuc, O. *3D QSAR models*, in *QSPR/QSAR studies by molecular descriptors*, Ed. Diudea, M.V. Nova Science, Huntington, New York, (in press).
29. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A*. **1982**, *79*, 2554-2558.
30. Zupan, J.; Gasteiger, J. Neural networks: a new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *248*, 1-30.
31. Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem. Int. Ed. Engl*. **1993**, *32*, 503-527.
32. Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH: Weinheim, **1993**.
33. Bulsari, A.B. *Neural Networks for Chemical Engineers*; Elsevier: Amsterdam, **1995**.
34. Devillers, J. *Neural Networks in QSAR and Drug Design*. Academic Press, London, **1996**, p. 279.
35. Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.P. ^{13}C NMR chemical shift prediction of sp^2 carbon atoms in acyclic alkenes using neural networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644-653.
36. Ivanciuc, O. Molecular graph descriptors used in neural network models. In:

Topological Indices and Related Descriptors in QSAR and QSPR; Devillers, J., Balaban, A.T., Eds.; Gordon and Breach Science Publishers: The Netherlands, **1999**, pp. 697-777.

37. Gakh, A.A.; Gakh, E.G.; Sumpter, B.G.; Noid, D.W. Neural network-graph theory approach to the prediction of the physical properties of organic compounds. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 832-839.
38. Baskin, I.I.; Palyulin V.A.; Zefirov, N.S. A Neural device for searching direct correlations between structures and properties of chemical compounds. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 715-721.
39. Kireev, D.B. ChemNet: A Novel neural network based method for graph/property mapping. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 175-180.
40. Topliss, J. G.; Edwards, R. P. Chance factors in in studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238.
41. Nikolić, S.; Medić-Sarić, M.; Matijević-Sosa; J. A QSAR study of 3-(Phtalimidoalkyl)-pyrazolin-5-ones, *Croat. Chem. Acta*, **1993**, *66*, 151-160.
42. Nagy, P.I.; Tokarski, J.; Hopfinger, A. J. Molecular shape and QSAR analysis of a family of substituted dichlorodiphenyl aromatase inhibitors, *J. Chem. Inf. Comput. Chem*. **1994**, *34*, 1190-1197.
43. Wessel, M. D.; Jurs, P. C. Prediction of normal boiling points for a diverse set of industrially important organic compounds from molecular structure, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 841-850.
44. Stanton, D. T.; Jurs, P. C. Development and use of charged partial surface area structural descriptors for Quantitative Structure-Property Relationships studies. *Anal. Chem*. **1990**, *62*, 2323.
46. Goll, E.S.; Jurs, P.C. Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model, *J. Chem. Inf. Comput. Chem*. **1999**, *39*, 974-983.
47. Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-assisted studies of chemical structure and biological function*, Wiley-Interscience: New York, 1979.
47. Trebst, A.; Draber, W. Advan. Pest. Sci., Symp. Papers IV-th Int. Congress Pest. Chem., Zurich, Swiss, 1978, pp. 223.
48. Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani, A. On the equivalence between different descriptions of molecules: value for computational approaches. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 575-578.