Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):2378-2393

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

QSAR study on Indole derivatives

Teodora E. Harsa[a], Alexandra M. Harsa[a], L. Jantschi[b] and Mircea V. Diudea[a*]

[a]Faculty of Chemistry and Chemical Engineering, Babes-Bolyai University, Cluj, Romania
[b]Faculty of Engineering of Materials and of Environment, Technical University, Cluj, Romania

---

ABSTRACT

This paper reviews the use of similarity searching in chemical databases, was performed on a set of 40 indole derivatives. The QSAR models describing log P and LD50 of this set of indole derivatives were validated by leave-one-out procedure, in the external test set and in a new version of prediction by using clusters of similarity.

Keywords: indoles, log P, QSAR, Lethal Dose, hypermolecule.

INTRODUCTION

Indole is an aromatic heterocyclic organic compound. It has a bicyclic structure, consisting of a six-membered benzene ring fused to a five-membered nitrogen-containing pyrole ring. Indole is a common component of fragrances and the precursor to many pharmaceuticals. Compounds that contain an indole ring are called indoles. The amino acid tryptophan is an indole derivative and the precursor of the neurotransmitter serotonin [1].

Indole structures are present in a great number of compounds of biological importance, e.g., the plant growth hormone indoleacetic acid (IAA),2 the pineal gland hormone melatonin, serotonin and tryptophan [2]. An indole is characterized as a benzene ring fused to a nitrogen-containing five-membered heterocylic ring. Substituents may be added anywhere on the above molecule to create an indole derivative. Several thousand indole derivatives appear annually in chemical literature[3].

Serotonin is an important biomolecule in physiological systems, playing a vital role in the regulation of mood, sleep, emesis, sexuality, and appetite. Low levels of serotonin are associated with several disorders, including depression, anxiety, and migraines [4]. Extremely high levels of serotonin can manifest toxicity and potentially fatal effects known as serotonin syndrome [5].

QSARs are mathematical models relating the observed biological activity of molecules (e.g., a indole) to their structural properties [6]. QSARs enable one to predict the biological activity of un-tested or un-synthesized compounds and to investigate the chemo-biological interactions involved in the system under study [7,8].Computational approach to drug design for oxazolidi- nones as antibacterial agents [9] .

Among thousands of topological indices [10], the Cluj indices are used for molecular graph description. They have been defined by Diudea [11, 12] as follows.

The LD50 (Lethal Dose, 50%) value is typically expressed in mg of material per kg of subject-body-weight, and indicates the quantity of material that, if administered to a population of subjects, will cause 50% of the subjects to perish [13].

It is known that compounds with similar physicochemical properties often share similar biological activitiesm [14].

In this article, we propose a new approach that develops clusters of similar structures aimed to be quasi-congeneric subsets in a better prediction of the biological activity.

EXPERIMENTAL SECTION

A set of 40 indoles were taken from PubChem Database [15] (Table 1) and were divided into a training set (25 molecules) and a test set (15 molecules), taken randomly. The property chosen for modeling was log P (calculated) partition coefficient between n-octanol and water (see Table 1) and LD50 (on rat, intraperitoneal route administered).

Table 1. Indole molecular structures (in SMILES code) and their log P and LD50 (taken from PubChem)

Nr. Crt.	**Canonical SMILES**	log P	LD50
1	C1=CC=C2C(=C1)C(=CN2)CCN	1.6	100
2	CNCCC1=CNC2=CC=CC=C21	2.1	158
3	CC1=CC2=C(C=C1)NC=C2CCN	1.9	50
4	C1=CC=C2C(=C1)C(=CN2)CC(C(=O)O)N	-1.1	4800
5	C1=CC=C2C(=C1)C(=CN2)CC(=O)O	1.4	150
6	C1=CC=C2C(=C1)C(=CN2)C=O	1.7	600
7	C1=CC2=C(C=C1O)C(=CN2)CCN	0.2	160
8	CC(=O)C1=CNC2=CC=CC=C21	2.1	300
9	CC(=O)OC1=CNC2=CC=CC=C21	2	600
10	C1=CC=C2C(=C1)C(=CN2)CCO	1.8	351
11	CN(C)CCC1=CNC2=C1C=C(C=C2)O	1.2	290
12	C1=CC=C2C(=C1)C=CN2	2.1	117
13	CN(C)CCC1=CNC2=C1C(=CC=C2)O	2.1	196
14	CN1C(=O)CC2=CC=CC=C21	1	1177
15	CCC(=O)NCCC1=CNC2=CC=CC=C21	2.3	900
16	C1=CC=C2C(=C1)C(=CN2)CCC(=O)O	1.8	100
17	CCNCCC1=CNC2=CC=CC=C21	2.4	562
18	CCOC(=O)C1=CC2=C(N1)C=CC(=C2)OC	3.2	350
19	C1=CC=C2C(=C1)C=C(N2)O	2.4	400
20	CC(=O)C1=CC2=C(C=C1)NC=C2	2	450
21	C1=CC2=C(C=CN2)C=C1C(=O)CO	1.3	600
22	CC(=O)CC1(C2=CC=CC=C2NC1=O)O	0.2	800
23	CN(C)CCC(C1=CNC2=CC=CC=C21)O	1.8	767
24	COC1=CC2=C(C=C1)NC=C2CC(=O)O	1.4	98
25	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	1.1	1125
26	C1=CC2=C(C=CN2)C=C1O	2	1000
27	COC1=CC2=C(C=C1)NC=C2	2.1	370
28	CC(CC1=CNC2=CC=CC=C21)N	2	20
29	CCC(CC1=CNC2=CC=CC=C21)N	2.5	400
30	CN(C)CC1=CNC2=CC=CC=C21	1.8	122
31	C1=CC2=C(C=C1O)C(=CN2)CC(=O)O	1.1	1125
32	C1=CC=C2C(=C1)C(=CN2)C(=O)CO	1.4	700
33	CCC(=O)NCCC1=CNC2=C1C=C(C=C2)OC	1.3	850
34	CC1=C(C2=C(N1)C=CC(=C2)OC)CCN(C)C	2.7	100
35	COC1=CC2=C(C=C1)NC=C2CCN	0.5	176
36	CC(=O)C1=CNC2=C1C(=O)CCC2	0.5	233
37	CC1=C(NC2=C1C(=O)CCC2)C(=O)C	1.2	533
38	CC1=CC2=C(C=C1)NC=C2C(=O)CO	1.8	600
39	COC1=CC2=C(C=C1)NC=C2C(=O)CO	1.1	600
40	COC1=CC2=C(C=C1)C(=CN2)C(=O)CO	1.1	600

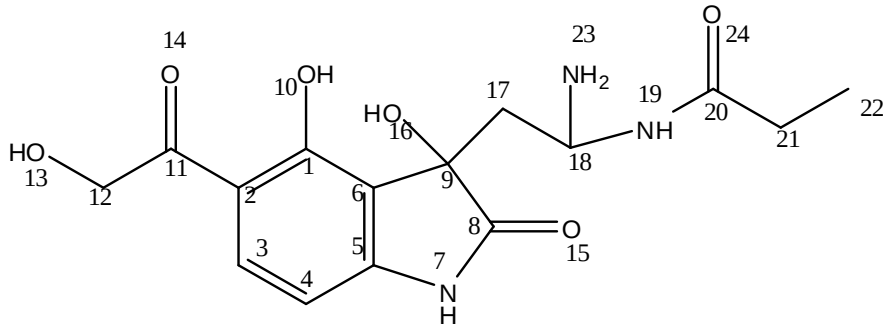

Figure 1. The hypermolecule comprising the common features of the dataset

A hypermolecule (Figure 1) was built up by superposing the all 40 molecules under study. The hypermolecule is considered to mimics the investigated statistical hyperspace[16].

Table 2. Binary vectors, cf. hypermolecule, for the 40 indoles derivatives

Mol.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
3	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
5	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
6	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
8	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
11	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
12	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0
14	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
15	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1
16	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
17	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
18	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0
19	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
21	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
22	1	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
23	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
24	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0
25	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
26	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
28	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
29	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
31	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
32	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	1	1	0	1
33	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
34	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0
35	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
36	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
37	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
38	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0
39	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0
40	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0

RESULTS AND DISCUSSION

3.1. COMPUTATIONAL DETAILS

The structures have been optimized at Hartree-Fock HF (3-21g(,p)) level of theory, in gas phase, by Gaussian 09 [17]. Topological indices have beed computed by TOPOCLUJ software [18]; some of them (Centric index of partial charges shells=Ch, Total adjacency = Adj, Detour = De, Distance = Di, D3D, SD), HOMO (in au) and log P are listed in Table 3 and 4.

Table 3. Topological indices, correlating descriptors, log P for the set of indoles in Table 1

Mol.	log P	Homo	Ch	De	Di	CjDi	CfDi	SD_1	SD_2
1	1.6	-8.418	0.037	477	191	290	312	-14.998	-2.285
2	2.1	-8.411	0.032	576	251	372	395	-14.768	-2.281
3	1.9	-8.351	-0.010	568	238	362	393	-14.890	-2.164
4	-1.1	-8.453	0.279	772	369	534	560	-16.967	-3.814
5	1.4	-8.545	0.322	566	241	361	385	-15.110	-2.416
6	1.7	-8.637	0.265	390	143	220	240	-15.090	-2.537
7	0.2	-8.334	0.088	568	238	362	393	-17.496	-4.825
8	2.1	-8.680	0.186	468	182	280	303	-15.071	-2.386
9	2	-8.539	0.241	566	241	361	385	-14.843	-2.097
10	1.8	-8.435	0.090	477	191	290	312	-15.110	-2.302
11	1.2	-8.282	0.027	790	376	552	586	-15.305	-2.855
12	2.1	-8.355	0.074	248	79	121	129	-14.281	-2.028
13	2.1	-8.178	-0.003	796	368	539	582	-14.776	-2.295
14	1	-8.677	0.277	389	138	212	238	-15.781	-3.171
15	2.3	-8.446	0.251	931	489	678	704	-14.716	-2.075
16	1.8	-8.530	0.250	677	313	457	481	-15.110	-2.312
17	2.4	-8.418	0.014	688	324	468	492	-14.381	-1.711
18	3.2	-8.479	0.280	893	454	671	699	-13.684	-0.907
19	1.16	-8.323	0.142	313	108	167	180	-15.759	-2.814
20	2	-8.614	0.247	462	188	284	305	-14.781	-2.103
21	1.3	-8.644	0.296	550	241	358	381	-15.482	-2.803
22	0.2	-9.209	0.432	765	330	486	542	-16.433	-3.903
23	1.8	-8.412	0.019	897	455	642	670	-14.880	-2.293
24	1.4	-8.387	0.352	782	363	548	583	-15.639	-2.974
25	1.1	-8.421	0.345	667	295	444	477	-15.639	-3.126
26	2	-8.253	0.118	312	108	165	180	-14.810	-2.376
27	2.1	-8.231	0.122	386	147	224	242	-14.810	-2.275
28	2	-8.293	0.029	566	241	361	385	-14.990	-2.351
29	2.5	-8.286	0.023	668	304	447	472	-14.990	-2.339
30	1.8	-8.353	0.006	566	241	361	385	-15.110	-2.119
31	1.1	-8.444	0.325	667	295	444	477	-15.639	-2.742
32	1.3	-8.307	0.300	1226	672	947	987	-15.246	-2.531
33	1.4	-8.730	0.241	558	233	352	377	-15.071	-2.410
34	2.7	-8.215	-0.016	1069	530	781	835	-14.170	-1.357
35	0.5	-8.307	0.073	672	298	453	487	-15.520	-2.633
36	0.5	-9.318	0.380	562	222	341	380	-16.465	-3.715
37	1.2	-9.258	0.408	666	273	415	468	-15.397	-2.763
38	1.8	-8.659	0.186	658	286	434	468	-14.971	-2.391
39	1.1	-8.550	0.267	772	353	537	573	-15.601	-2.890
40	1.1	-8.550	0.267	772	353	537	573	-15.591	-2.890

Table 4. Topological indices, correlating descriptors, LD50 for the set of indoles in Table 1

Mol.	LD50	Homo	SD_3	SD_4
1	100	-8.42	-7551.35	3517.42
2	158	-8.41	-7558.74	3630.25
3	50	-8.35	-7611.50	3524.56
6	600	-8.64	-6997.68	3856.13
7	160	-8.33	-7551.72	3467.45
8	300	-8.68	-7139.59	3827.19
9	600	-8.54	-7111.42	3890.92
10	351	-8.44	-7130.02	3684.37
11	290	-8.28	-7454.12	3536.22
13	196	-8.18	-7530.45	3409.03
15	900	-8.45	-6887.26	4207.44
17	562	-8.42	-7149.71	3869.45
18	350	-8.48	-7359.36	3708.18
19	400	-8.32	-7303.44	3705.60
20	450	-8.61	-7261.72	3785.51
21	600	-8.64	-7111.72	3878.07
22	800	-9.21	-6911.72	4153.83
23	767	-8.53	-7130.02	3998.12
27	370	-8.23	-7352.35	3545.26
28	20	-8.29	-7570.13	3450.61
29	400	-8.29	-7570.13	3441.50
32	700	-8.31	-6936.17	4007.44
33	850	-8.73	-7130.08	3887.46
34	100	-8.21	-7561.90	3401.44
35	176	-8.31	-7503.31	3520.24
36	233	-9.32	-7399.70	3619.89
37	533	-9.26	-7257.73	3812.97
38	600	-8.66	-7161.94	3781.25
39	600	-8.55	-7178.99	3960.78
40	600	-8.55	-7161.53	3960.78

This QSAR study was performed following Diudea's algorithm [19]; it is based on the alignment of molecules over a hypermolecule [20] and a correlation weighting procedure [21, 22] coupled with a predictive validation of the model descriptors within similarity clusters [23] performed for each molecule in the test set. The algorithm can be extended with other powerful statistical tools (e.g. pls or pca) but we limited here to the more common multi linear regression in achieving the best prediction of a chosen property, like log P or LD50.

The models fit abilities were assess by cross-validation leave one analysis [24] using a dedicated software [25,26].

3.2. Mass fragments description (case 1)

3.2.1. Data reduction (for log P)

In the step of data reduction, all the descriptors with the variance Var<20% and those with intercorrelation larger than 0.80 have been discarded.

This new descriptor SD_1, that is a linear combination of the local correlating descriptors for the significant positions in the hypermolecule (i.e. H1, H2, H3, H7, H8, H9, H11, H13, H14, H15, H17, H19, H20, H21, H22, H23).

3.2.2. QSAR models (for log P)

The models were performed on the training set (the first 25 structures in Table 1) and the best results are listed below and in Table 5. The number of descriptors was limited to four, to fulfill the considerations of Topliss and Costello.

(i) Monovariate regression

$$\log P = 14.529 + 0.840 \times SD_1$$

(ii) Bivariate regression

$$\log P = 16.929 + 0.803 \times SD_1 + 0.349 \times HOMO$$

(iii) Three-variate regression

$$\log P = 15.971 + 0.763 \times SD_1 + 0.330 \times HOMO + 0.001 \times D3D$$

(iv) Four-variate regression

$$\log P = 14.846 + 0.868 \times SD_1 - 0.177 \times Ch + 0.007 \times CjDi - 0.006 \times CfDe$$

Table 5. Best models in describing log P in the training set of indoles in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_1**	**0.917**	0.914	0.190	255.182
2	**CjDi**	0.241	0.208	0.577	7.307
3	**Di**	0.250	0.218	0.573	7.686
4	**HOMO**	0.185	0.150	0.598	5.229
5	**SD_1, HOMO**	**0.932**	0.925	0.177	149.937
6	**SD_1, Di**	0.930	0.924	0.179	146.548
7	**SD_1, D3D**	0.930	0.924	0.179	145.925
8	**SD_1, CjDi**	0.930	0.924	0.179	146.047
9	**SD_1, De**	0.929	0.923	0.180	144.419
12	**SD_1, HOMO, D3D**	**0.943**	0.934	0.166	115.071
13	**SD_1, CjDi, HOMO**	0.943	0.934	0.166	114.815
11	**SD_1, De, HOMO**	0.942	0.934	0.167	113.940
15	**SD_1, CfDi, HOMO**	0.942	0.934	0.167	113.594
10	**SD_1, Adj, Ch**	0.939	0.931	0.171	108.180
14	**SD_1, Di, De**	0.932	0.922	0.181	95.569
16	**SD_1, Ch, CfDi, CfDe**	**0.953**	0.944	0.154	101.446
17	**SD_1, HOMO, D3D, De**	0.943	0.932	0.170	82.611
18	**SD_1, C, Adj, Ch**	0.942	0.931	0.170	81.835
19	**SD_1, D3D, Di, De**	0.934	0.920	0.183	70.233

3.2.3. Model Validation (for log P)

(a) Leave-one-out

The performances in leave-one-out analysis related to the models listed as best in Table 5 are presented in Table 6.

Table 6. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_1**	**0.878**	0.039	0.231	166.16
5	**SD_1, HOMO**	**0.897**	0.035	0.212	200.918
12	**SD_1, HOMO, D3D**	**0.916**	0.027	0.192	249.674
16	**SD_1, Ch, CfDi, CfDe**	**0.882**	0.071	0.227	172.585

(b) External Validation

The values log P for the test set of indoles (Table 1),were calculated by using the best equation in Table 5, entry 12. Data are listed in Table 7 and the monovariate correlation: $\log P = 0.941 + 0.605 \times \log P_{calc.}$;n=15; R^2=0.808; s=0.343; F= 54.720 is plotted in Figure 2.

Table 7. Calculated values of log P for the molecules in the test set (Table 1)

Mol.	logP	log $P_{calc.}$
1	1.6	1.89
4	-1.1	0.49
5	1.4	1.79
11	1.2	1.45
12	2.1	2.38
22	0.2	0.62
23	1.8	2.14
25	1.1	1.46
31	1.1	1.45
32	1.3	2.05
33	1.4	1.76
35	0.5	1.59
37	1.2	1.37
39	1.1	1.49
40	1.1	1.5

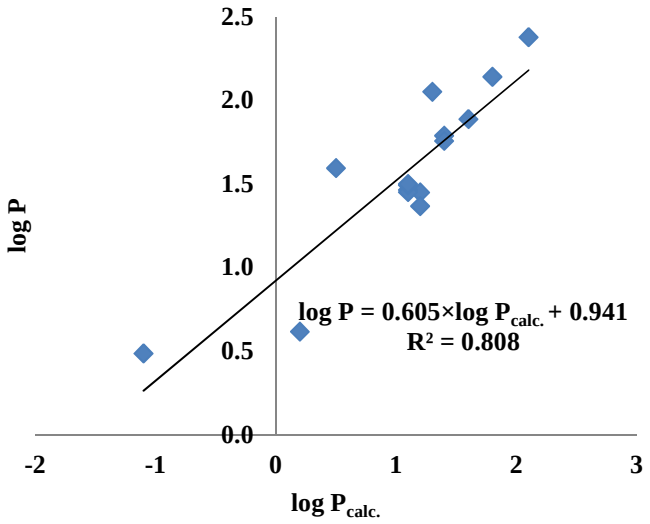

Figure 2. The plot log P vs. log $P_{calc.}$ for the test set (external validation)

(c) Similarity Cluster Validation

Validation can be performed by calculating log P for the molecules in the test set with equations learned on clusters of similarity: each of the 15 molecules is the leader in its own cluster, selected by (2D) similarity among the 25 structures of the initial learning set. The values log $P_{calc.}$ for each of the 15 molecules in the test set were computed by 15 new equations (the leader being left out) with the same descriptors as in eq. 12, Table 5. Data are listed in Table 8 and the monovariate correlation: $\log P = 0.857 \times \log P_{calc.} + 0.301$; n=15; R^2=0.945; s=0.182; F=226.539 is plotted in Figure 3.

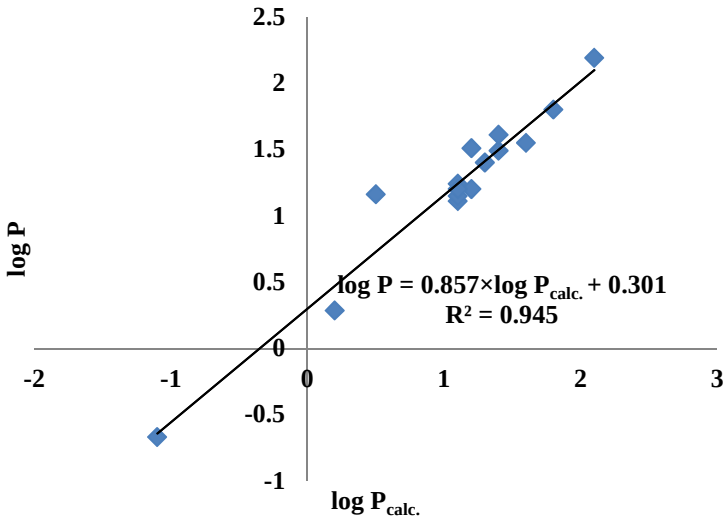

Figure 3. The plot log P vs. log $P_{calc.}$ for the test set (external validation)

Table 8. Calculated values of log P for the molecules in the test set (Table 1)

Mol.	log P	log $P_{calc.}$
1	1.6	1.55
4	-1.1	-0.67
5	1.4	1.49
11	1.2	1.51
12	2.1	2.19
22	0.2	0.29
23	1.8	1.80
25	1.1	1.11
31	1.1	1.24
32	1.3	1.40
33	1.4	1.61
35	0.5	1.16
37	1.2	1.20
39	1.1	1.15
40	1.1	1.20

3.2.4. Data reduction (for LD50)

In data reduction, the same procedure was used as in Section 3.2.1. The local correlating descriptors are summed, to give SD_3 global descriptor, over the following significant positions in the hypermolecule: H1, H2, H3, H7, H11, H12, H13, H14, H16, H17, H18, H19, H20, H21, H23 and H24 and it will be used as the basis of modeling LD50 (see Table 9).

3.2.5. QSAR models (for LD50)

The models were performed on the training set (the 25 structures in Table 1) and the best results are listed below and in Table 9.

(v) Monovariate regression

$$LD50 = 7828.027 + 0.008 \times SD_3$$

(vi) Bivariate regression

$$LD50 = 7734.028 + 1.0002 \times SD_3 + 0.052 \times De$$

(vii)Three-variate regression

$$LD50 = 7657.091 + 0.984 \times SD_3 + 0.881 \times CjDe - 0.464 \times CfDi$$

(viii) Four-variate regression

$$LD50 = 6918.807 + 0.954 \times SD_3 + 4.577 \times De + 3.861 \times CjDi - 8.935 \times CfDi$$

Table 9. Best models in describing LD50 in the training set of indoles in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_3**	**0.852**	0.844	87.268	103.849
2	**CjDe**	0.068	0.016	219.234	1.307
3	**HOMO**	0.290	0.251	191.308	7.355
4	**De**	0.050	0.003	221.284	0.951
5	**SD_3, De**	**0.854**	0.836	89.393	49.562
8	**SD_3, CjDe**	0.854	0.837	89.141	49.891
6	**SD_3, D3D**	0.853	0.836	89.535	49.379
7	**SD_3, Di**	0.853	0.836	89.486	49.442
9	**SD_3, HOMO**	0.853	0.836	89.579	49.322
11	**SD_3, CfDi, CjDe**	**0.859**	0.833	90.413	32.507
13	**SD_3, HOMO, C**	0.856	0.829	91.403	31.691
12	**SD_3, De, CjDi**	0.855	0.828	91.685	31.464
14	**SD_3, HOMO, De**	0.854	0.827	91.912	31.282
16	**SD_3, HOMO, D3D**	0.854	0.827	91.999	31.213
15	**SD_3, De, Di**	0.854	0.827	91.951	31.251
17	**SD_3, De, CjDi, CfDi**	**0.900**	0.854	84.424	28.800
18	**SD_3, HOMO, CjDi, CfDi**	0.885	0.854	84.424	28.800
19	**SD_3, HOMO, D3D, De**	0.855	0.816	94.741	22.097

3.2.6. Model Validation (for LD50)

(a) Leave-one-out

The performances in leave-one-out analysis related to the models listed as best in Table 9 are presented in Table 10.

Table 10. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. Error$_{loo}$	F_{loo}
1	**SD_3**	**0.812**	0.04	98.356	77.926
5	**SD_3, De**	**0.802**	0.052	100.923	73.108
11	**SD_3, CfDi, CjDe**	**0.794**	0.065	103.057	69.374
17	**SD_3, De, CjDi, CfDi**	**0.805**	0.095	100.373	74.108

(b) External Validation

The values LD50 for the test set of indoles (Table 1, 10 structures) were calculated by using the best equation in Table 9, entry 11. Data are listed in Table 11 and the monovariate correlation: $LD50 = 93.628 + 1.183 \times LD50_{calc.}$;n=10; R^2=0.869; s=72.596; F= 78.415 is plotted in Figure 4.

Table 11. Calculated values of LD50 for the molecules in the test set (Table 1)

Mol.	LD50	$LD50_{calc.}$
1	100	214.33
3	50	147.53
6	600	754.57
8	300	615.68
10	351	628.85
28	20	202.52
32	700	877.29
34	100	215.02
35	176	251.75
36	233	354.88

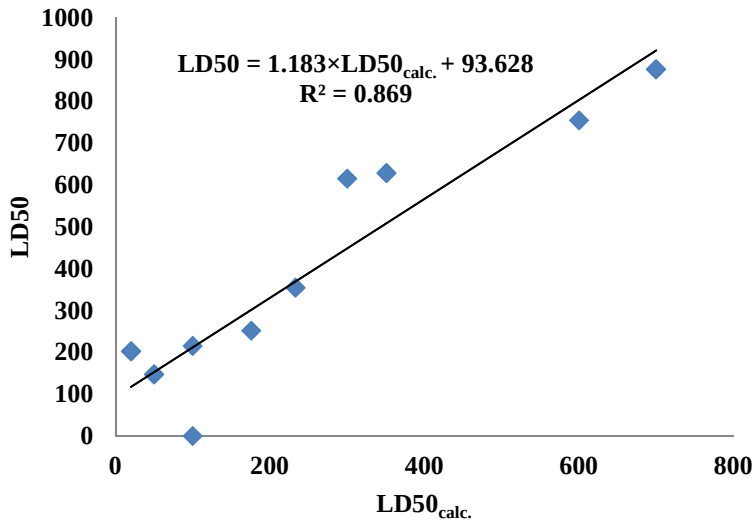

Figure 4. The plot LD50 vs. $LD50_{calc.}$ for the test set (external validation)

(c) Similarity Cluster Validation

Validation was performed by calculating LD50 for the molecules in the test set, similar to the Section 3.2.5. The values $LD50_{calc}$. were computed with the same descriptors as in eq. 11, Table 9.Data are listed in Table 12 and the monovariate correlation: $LD50 = 1.119 \times LD50_{calc.} + 94.282$; n=10; R^2=0.909; s=73.729; F=80.241 is plotted in Figure 5.

Table 12. Calculated values of LD50 by similarity clusters, for the molecules in the test set

Mol	LD50	$LD50_{calc.}$
1	100	166.04
3	50	94.91
6	600	689.75
8	300	592.48
10	351	592.53
28	20	155.90
32	700	866.01
34	100	220.29
35	176	204.30
36	233	304.71

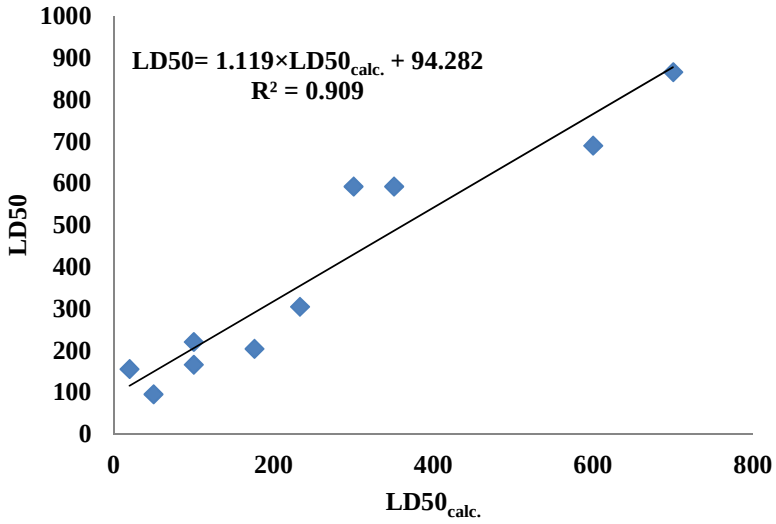

Figure 5. The plot LD50 vs. $LD50_{calc.}$ by similarity clusters

3.3. Partial charges description (case 2)

3.3.1. Data reduction (for log P)

In the step of data reduction, the same procedure was used as in Section 3.2.1. The local correlating descriptors are summed over the following significant positions in the hypermolecule: H2, H3, H4, H5, H6, H8, H10, H12, H13, H14, H16, H17, H21, H23, H24); the resulting SD_2 global descriptor will be used as the basis of modeling log P (see Table 3).

3.3.2. QSAR models (for log P)

The models were performed on the training set (Table 1) and the best results are listed below and in Table 13.

(ix) Monovariate regression

$$\log P = 0.829 \times SD_2 + 0.826$$

(x) Bivariate regression

$$\log P = 7.221 + 0.785 \times SD_2 + 0.411 \times HOMO$$

(xi) Three-variate regression

$$\log P = 6.834 + 0.753 \times SD_2 + 0.397 \times HOMO + 0.0003 \times De$$

(xii) Four-variate regression

$$\log P = 5.929 + 0.761 \times SD_2 + 0.454 \times HOMO - 0.001 \times De + 0.167 \times Adj$$

Table 13. Best models in describing log P in the training set of indoles in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_2**	**0.888**	0.883	0.221	182.766
2	**Di**	0.250	0.218	0.573	7.686
3	**CjDi**	0.241	0.208	0.577	7.307
4	**De**	0.215	0.181	0.587	6.312
5	**SD_2, HOMO**	**0.908**	0.900	0.205	109.025
6	**SD_2, CjDe**	0.899	0.889	0.216	97.561
7	**SD_2, Di**	0.898	0.889	0.216	96.920
8	**SD_2, De**	0.897	0.888	0.217	95.806
9	**SD_2, D3D**	0.898	0.888	0.217	96.608
10	**SD_2, HOMO, De**	**0.916**	0.904	0.201	76.040
11	**SD_2, HOMO, D3D**	0.916	0.904	0.201	76.608
12	**SD_2, Adj, Ch**	0.908	0.895	0.210	69.132
13	**SD_2, Di, De**	0.901	0.887	0.218	63.567
14	**SD_2, De, CjDi**	0.899	0.885	0.220	62.277
15	**SD_2, D3D, De**	0.899	0.884	0.221	62.047
16	**SD_2, HOMO, De, Adj**	**0.919**	0.902	0.202	56.589
17	**SD_2, HOMO, De, Di**	0.916	0.900	0.205	55.193
18	**SD_2, Di, D3D, De**	0.904	0.884	0.221	46.872

3.3.3. Model Validation (for log P)

(a) Leave-one-out

The performances in leave-one-out analysis related to the models listed as best in Table 13 are presented in Table 14.

Table 14. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_2**	**0.85**	0.038	0.256	130.397
5	**SD_2, HOMO**	**0.879**	0.029	0.23	168.056
10	**SD_2, HOMO, De**	**0.884**	0.032	0.225	175.399
16	**SD_2, HOMO, De, Adj**	**0.878**	0.041	0.232	165.098

(b) External Validation

The values log P for the test set of indoles (Table 1),were calculated by using the best equation in Table 13, entry 10. Data are listed in Table 15 and the monovariate correlation: $\log P = 0.401 \times \log P_{calc.} + 1.173$; n=15; R^2=0.736; s=0.403; F=30.605 is plotted in Figure 6.

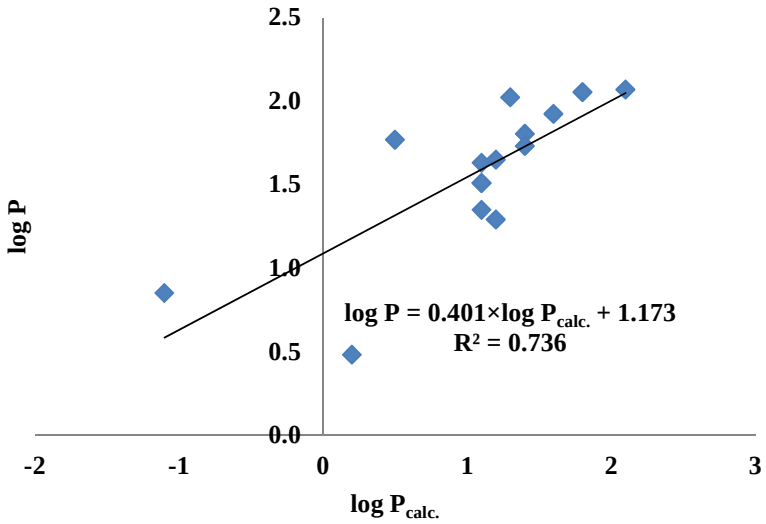

Figure 6. The plot log P vs. log $P_{calc.}$ for the test set (external validation)

Table 15. Calculated values of log P for the molecules in the test set (Table 1)

Mol.	log P	log $P_{calc.}$
1	1.6	1.92
4	-1.1	0.85
5	1.4	1.80
11	1.2	1.65
12	2.1	2.07
22	0.2	0.48
23	1.8	2.05
25	1.1	1.35
31	1.1	1.63
32	1.3	2.02
33	1.4	1.73
35	0.5	1.77
37	1.2	1.29
39	1.1	1.51
40	1.1	1.51

(c) Similarity Cluster Validation

Validation was performed by calculating log P for the molecules in the test set, similar to that in the Section 3.2.3. The values log $P_{calc.}$ were computed with the same descriptors as in eq. 10, Table 13. Data are listed in Table 16 and the monovariate correlation: $\log P = 0.662 \times \log P_{calc.} + 0.597$; n=15; R^2=0.922; s=0.217; F=155.571 is plotted in Figure 7.

Table 16. Calculated values of log P by similarity clusters, for the molecules in the test set

Mol.	log P	log $P_{calc.}$
1	1.6	1.69
4	-1.1	-0.05
5	1.4	1.48
11	1.2	1.44
12	2.1	2.09
22	0.2	0.53
23	1.8	1.96
25	1.1	1.27
31	1.1	1.32
32	1.3	1.51
33	1.4	1.43
35	0.5	1.28
37	1.2	1.25
39	1.1	1.18
40	1.1	1.18

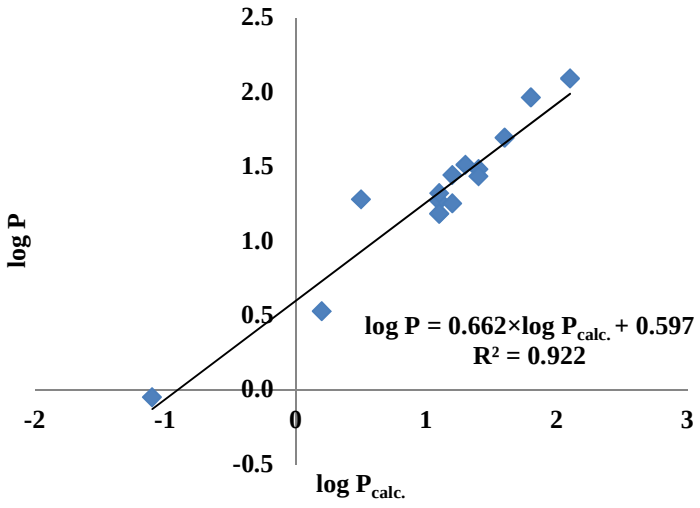

Figure 7. The plot Log P vs. log $P_{calc.}$ by similarity clusters

3.3.4. Data reduction (for LD50)

In the step of data reduction, the same procedure was used as in Section 3.2.4. The local correlating descriptors are summed over the following significant positions in the hypermolecule: H1, H2, H3, H4, H6, H7, H11, H14, H17, H19, H20, H21, H22, H23, H24); the resulting SD_4 global descriptor will be used as the basis of modeling LD50 (see Table 4).

3.3.5. QSAR models (for LD50)

The models were performed on the training set (Table 1) and the best results are listed below and in Table 17.

(xiii) Monovariate regression

$$LD50 = 0.889 \times SD_4 - 2848.41$$

(xiv) Bivariate regression

$$LD50 = -2894.12 + 0.884 \times SD_4 - 6.922 \times HOMO$$

(xv) Three-variate regression

$$LD50 = -2724.73 + 0.872 \times SD_4 + 0.586 \times Di - 0.367 \times De$$

(xvi) Four-variate regression

$$LD50 = -3060.6 + 0.803 \times SD_4 + 3.803 \times De + 3.713 \times CjDi - 7.918 \times CfDi$$

Table 17. Best models in describing LD50 in the training set of indoles in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_4**	**0.896**	0.890	71.605	154.431
2	**Di**	0.024	0.031	219.000	0.434
3	**HOMO**	0.203	0.159	197.817	4.593
4	**De**	0.011	0.044	220.404	0.200
5	**SD_4, HOMO**	**0.896**	0.883	73.665	72.963
6	**SD_4, Di**	0.896	0.884	73.621	73.059
7	**SD_4, CjDe**	0.896	0.883	73.675	72.939
8	**SD_4, D3D**	0.896	0.884	73.607	73.090
9	**SD_4, De**	0.896	0.884	73.584	73.141
10	**SD_4, Di, CjDi**	**0.905**	0.888	72.339	50.985
11	**SD_4, Di, De**	0.897	0.878	75.380	46.532
12	**SD_4, HOMO, D3D**	0.896	0.876	75.872	45.862
13	**SD_4, HOMO, Di**	0.896	0.876	75.886	45.843
14	**SD_4, D3D, De**	0.896	0.877	75.649	46.164
15	**SD_4, Adj, Ch**	0.896	0.876	75.917	45.801
16	**SD_4, D3D, De**	0.896	0.877	75.649	46.164
17	**SD_4, De, CjDi, CfDi**	**0.921**	0.899	68.402	43.490
18	**SD_4, C, Di, De, CjDi**	0.911	0.888	72.295	38.540
19	**SD_4, C, D3D, De**	0.906	0.881	74.433	36.145
20	**SD_4, De, D3D, Di**	0.899	0.872	77.225	33.313

3.3.6. Model Validation (for LD50)

(a) Leave-one-out

The performances in leave-one-out analysis related to the models listed as best in Table 17 are presented in Table 18.

Table 18. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_4**	**0.868**	0.028	80.594	118.112
5	**SD_4, HOMO**	**0.865**	0.031	81.381	115.492
10	**SD_4, Di, De**	**0.864**	0.041	81.522	115.033
17	**SD_4, De, CjDi, CfDi**	**0.821**	0.1	93.722	82.653

(b) External Validation

The values LD50 for the test set of caffeine (Table 1),were calculated by using the best equation in Table 17, entry 10. Data are listed in Table 19 and the monovariate correlation: $LD50 = 0.544 \times LD50_{calc.} + 259.11$; n=10; R^2=0.811; s=143.121; F=29.989 is plotted in Figure 8.

Table 19. Calculated values of LD50, for the molecules in the test set

Mol.	LD50	$LD50_{calc.}$
1	100	287.58
3	50	384.34
6	600	278.48
8	300	541.77
10	351	410.44
28	20	812.83
32	700	227.69
34	100	589.17
35	176	355.87
36	233	-2603.45

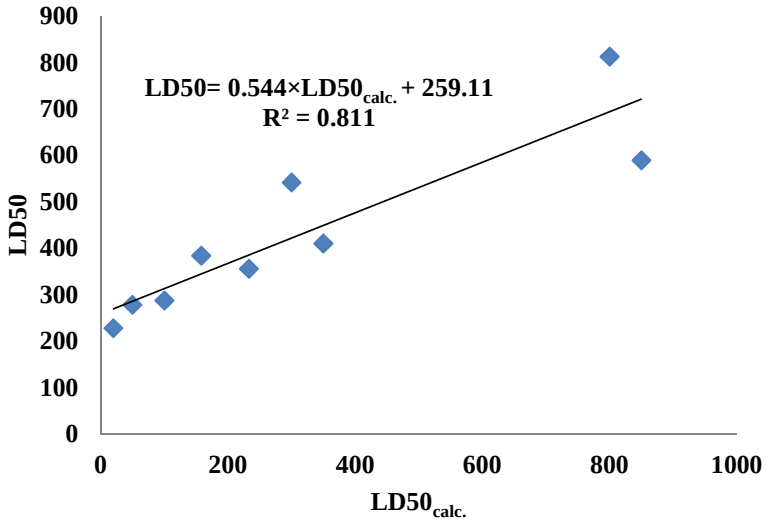

Figure 8. The plot LD50 vs. $LD50_{calc.}$ (external validation)

(c) Similarity Cluster Validation

Validation was performed by calculating LD50 for the molecules in the test set, similar to that in the Section 3.2.6. The values $LD50_{calc.}$ were computed with the same descriptors as in eq. 10, Table 17. Data are listed in Table 20 and the monovariate correlation: $LD50 = 192.63 + 0.723 \times LD50_{calc.}$;n=10; R^2=0.934; s=82.445; F= 114.006 is plotted in Figure 9.

Table 20. Calculated values of LD50 for the molecules, similarity cluster in the test set (Table 1)

Mol.	LD50	$LD50_{calc.}$
1	100	231.22
2	158	340.6
3	50	239.14
8	300	527.67
18	350	378.29
22	800	712.21
28	20	207.31
33	850	861.32
36	233	304.74
39	600	627.27

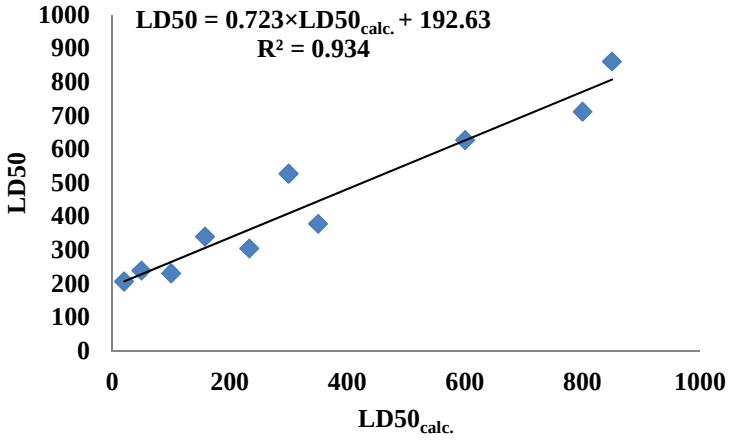

Figure 9. The plot LD50 vs. LD50 $_{calc.}$ for the test set (similarity cluster validation)

CONCLUSION

A set of 40 indoles, downloaded from the PubChem database, was submitted to a QSAR study, involving the hypermolecule concept, in a procedure similar to that of the „alignment" of drug molecules to the biological receptors. A hypermolecule of a dataset is generated by aligning successive pairs of molecules and combining features of the same type that lie in close proximity whilst retaining the individuality of the atoms themselves.

The set was split into a learning set and a test set, the last one being used for the validation of the models, in the so-called external set validation. Also, the validation was made by a new version of prediction by using similarity clusters.

Acknowledgements
The work was supported: POSDRU/159/1.5/S/137750

REFERENCES

[1] DL Nelson; MM Cox; Principles of Biochemistry (4th ed.), New York: W. H. Freeman, ISBN 0-7167-4339-6, **2005**.
[2] N Sugiama; M Akutagama; and H Yamamoto; *Bull. Chem. Soc. Japan*, **1968**, 41**,** 937–941.
[3] PR Brodfuehrer; B Chen; TR Sattelberg; PR Smith; JP Reddy; DR Stark; SL Quinlan; JG Reid; *J. Org. Chem.*, **1997**, 62, 9192-9202.
[4] MT Lin; HJ Tsay; WH Su; FY Chueh; *Am. J. Physiol. Integr. Comp. Physiol*,. **1998**, 274,1260–1267.
[5] L Imeri; M Mancia; S Bianchi; MR Opp, *Neuroscience*, **1999**, 95, 445–452.
[6] F De Benedetti and PG De Benedetti; F Fanelli, *Drug Discovery*, **2010**, 15, 859–866.
[7] O Deeb; B Hemmateenejad, *Chem. Biol. Drug Des*., **2007**, 70, 19–29.
[8] Y Li; DQ Wei; WN Gao; H Gao; BN Liu; CJ Huang; WR Xu; DK Liu; HF Chen; KC Chou, *Med Chem.*, **2007,** 6(3), 576-82.
[9] GM Maggiora; C Zhang; C.T. K. C. C, D. W. Elrod, In in Neural Networks in QSAR and Drug Design. In: Devillers, J. (Ed.), *Academic Press*, London, **1996,** 3, 576–582.
[10] M Randić, *J. Chem. Inf. Comput. Sci*., **1995**, 35,373-382.
[11] MV Diudea, *MATCH Commun. Math. Comput. Chem*., **1997**, 35, 169-183.
[12] MV Diudea , *J. Chem. Inf. Compu. Sci*., **1997**, 37, 300-305.
[13] AD DeWeese, and TW Schultz, *Environ. Toxicol*. **2001**, 16,54–60.
[14] M Dunkel; S Günther; J Ahmed; B Wittig; R Preissner; *Nucleic Acids Res*., **2008**, 36,55–59.
[15] PubChem database.
[16] AT Balaban; A Chiriac; I Motoc; and Z Simon, *Steric Fit in QSAR* (*Lectures Notes in Chemisry*, Vol. 15), *Springer, Berlin*, **1980**.

[17] **Gaussian 09**, Gaussian Inc Wallingford CT, Revision A.1 MJ Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, V Barone, B Mennucci, GA Petersson, H Nakatsuji, M Caricato, X Li, HP Hratchian, AF Izmaylov, J Bloino, G Zheng, JL Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, JA Montgomery, JE Peralta, F Ogliaro, M Bearpark, JJ Heyd, E Brothers, KN Kudin, VN Staroverov, R Kobayashi, J Normand, K Raghavachari, A Rendell, JC Burant, SS Iyengar, J Tomasi, M Cossi, N Rega, NJ Millam, M Klene, JE Knox, JB Cross, V Bakken, C Adamo, J Jaramillo, R Gomperts, RE Stratmann, O Yazyev, AJ Austin, R Cammi, C Pomelli, JW Ochterski, RL Martin, K Morokuma, VG Zakrzewski, GA Voth, P Salvador, JJ Dannenberg, S Dapprich, AD Daniels, Ö Farkas, JB Foresman, JV Ortiz, J Cioslowski, DJ Fox, **2009**.
[18] O Ursu; MV Diudea; "TOPOCLUJ software program", Babes-Bolyai University,Cluj, **2005**.
[19] CD Moldovan; A Costescu; G Katona; and MV Diudea; *MATCH Commun. Math. Comput. Chem.*, **2008**, 60, 977-984.
[20] NJ Richmond; CA Abrams; PR Wolohan; E Abrahamian; P Willett; RD Clark; *J Comput Aided Mol Des*, **2006**, 20, 567–587.
[21] AA Toropov; AP Toropova; *Internet El. J.Molec.Design*, **2002**, 1,108-114.
[22] AA Toropov; AP Toropova; *J. Mol. Struct.* (*Theochem*) **2001**, 538,287–293.
[23] P Willett; *J. Chem. Inf. Model.* **2013**, 53, 1−10.
[24] DM Hawkins; SC Basak; and D Mills; *J. Chem. Inf. Comput. Sci.*, **2003**, 43,579-586.
[25] SD Bolboacă; L Jäntschi; and MV Diudea; *Current Computer-Aided Drug Design*, **2013**, 9(2), 195-205.
[26] L Jäntschi; LOO Analysis (LOO: leave one out), Academic Direct Library of software, **2005**, Available at: http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/loo/