Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):2281-2295

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

QSAR study on Testosterone derivatives

Alexandra M. Harsa[a], Teodora E. 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

Steroid hormones are important in physiological processes of humans, as maturation, reproduction and even mortality. We report here a QSAR study performed on a set of 40 testosterone derivatives, downloaded from the PubChem database and aligned over a hypermolecule that mimics the investigated correlational space. The best models describing log P and LD50 of these testosterone derivatives were validated by leave-one-out procedure, in the external test set and in a new version of prediction based on clusters of similarity.

Keywords: testosterone, QSAR, LD50, log P, steroid, leave-one-out, similarity cluster

---

INTRODUCTION

Steroid hormones regulate important physiological processes in humans including maturation, reproduction, development of gonads, maintenance of blood volume and electrolyte concentration, and synthesis of bone and muscle [1].

Anabolic–androgenic steroids (AAS) are synthetic derivatives of testosterone with four-ring structures consisting of 19 carbon atoms. There are approximately 60 different AAS available, and these molecules vary in their chemical structure and, as a result, in their metabolic and physiological effects [2, 3].

The evidence is mixed with regard to testosterone levels and cardiovascular disease risk [4, 5]. Some studies have found an association between low testosterone and increased risk [6-8], while others have not [9]. A study of 208 men with heart failure found that deficiencies in serum testosterone were associated with increased mortality [10]. Testosterone promotes growth of many prostate cancers, and therefore, reducing circulating testosterone to very low (castration) levels is often the treatment goal in the management of men with advanced prostate cancer [11].

Human studies also demonstrate decline in testosterone at the transition to fatherhood, and lower testosterone levels were found in fathers who were more involved in child care [12]. In this respect, a very important role may be played by computer-added drug discovery techniques based on Quantitative-Structure–Activity-Relationship (QSAR) models [13].

The logarithm of the partition coefficient between n-octanol and water, also referred to as logP, has been widely used in QSAR studies as a key parameter for characterizing lipophilicity [14].

The LD50 for a particular substance is the amount that can cause death in half (i.e. 50%) of a group of some particular animal species, usually rats or mice, when entering the animal's body by a particular route [15].

The concept of similarity, enabling one to group molecules according to their structure, or biological effects or physico-chemical properties, has found extensive use in drug discovery. Similarity methods have found particular favor in the pharmaceutical industry. Indeed, medicinal chemistry relies heavily on the concept of bioisosterism in which similar substructures may be interchanged whilst maintaining some degree of activity [16,17].

The non-symmetric Cluj matrix, UCJ, has been introduced by Diudea [18-20]. It is defined by using either the distance or the detour concept. The non-diagonal entries, [UM]ij, M=CJD (Cluj-Distance) or CJΔ (Cluj-Detour), are defined as: $[\mathrm{UM}]_{i,j} = \underset{k=1,2,\ldots}{max} \left| V(i, j, p(i, j)_k \right|$ [18].

EXPERIMENTAL SECTION

A set of 40 testosterones were taken from PubChem Database [21] (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 mouse, oral route administered).

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

Nr. Crt.	**Canonical SMILES**	log P	LD50
1	CC12CCC3C(C1CCC2O)CCC4=CC(=O)CCC34C	3.3	5000
2	CCC(=O)OC1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	4.4	1000
3	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4(C)O)C	3.4	2500
4	CC12CCC(CC1CCC3C2CCC4(C3CCC4=O)C)O	3.7	980
5	CCCCCCC(=O)OC1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	6.3	1000
6	CC12CCC3C(C1CCC2OC(=O)CCC4=CC=CC=C4)CCC5=CC(=O)CCC35	5.1	595
7	CCCCC(=O)OC1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	5.3	980
8	CC(=O)OC1CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C	3.9	980
9	CC12CCC3C(C1CCC2(C#C)O)CCC4=CC(=O)CCC34	3	2000
10	CC12CCC3C(C1CCC2=O)CC(=C)C4=CC(=O)C=CC34C	3.1	980
11	CC(=O)OC1(CCC2C1(CCC3C2CCC4=CC(=O)CCC34)C)C#C	3.5	980
12	CC1CC2C(CCC3(C2CCC3(C(=O)C)OC(=O)C)C)C4(C1=CC(=O)CC4)C	4.1	6400
13	CC(=O)OC1CCC2C1(CCC3C2CCC4=C(C(=O)CCC34C)Cl)C	4.7	980
14	CCC12CCC3C(C1CCC2(C#C)O)CCC4=CC(=O)CCC34	3.3	5010
15	CC12CCC3C(C1CCC2OC(=O)C4=CC=CC=C4)CCC5=CC(=O)CCC35C	5.6	980
16	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4(C#C)O)C	3.5	980
17	CCC12CC(=C)C3C(C1CCC2(C#C)O)CCC4=CC(=O)CCC34	3.3	980
18	CC12CCC3C(C1CCC2O)CCC4=CC(=O)C=CC34C	3.5	980
19	CC1CC2C3CCC(C3(CC(C2C4(C1=CC(=O)CC4)C)O)C)C(=O)C	2.7	980
20	CC12CC(C3C(C1CCC2(C(=O)CO)O)CCC4=CC(=O)C=CC34C)O	1.6	250
21	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)CO)O)C)O	1.6	5000
22	CCC(=O)C1(C(CC2C1(CC(C3C2CCC4=CC(=O)C=CC34C)O)C)C)C	3.5	980
23	CC1CC2C3CCC(C3(CC(C2C4(C1=CC(=O)C=C4)C)O)C)(C(=O)CO)O	1.9	4000
24	CC1CC2C3CCC(C3(CC(C2C4(C1=CC(=O)C=C4)C)O)C)(C(=O)COC(=O)C)O	2.7	10000
25	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4(C#C)O)C	3.5	980
26	CC12CCC3C(C1CCC2(C)O)CCC4C3(COC(=O)C4)C	3.7	10000
27	CC12C=CC3=C4CCC(=O)C=C4CCC3C1CCC2(CC=C)O	2.8	980
28	CC12CCC3C(C1CCC2OC(=O)C4=CC=CC=C4)CCC5=CC(=O)CCC35C	5.6	980
29	CC12CCC(=O)C=C1CCC3C2C(CC4(C3CCC4(C(=O)O)O)C)O	1.6	980
30	CC12CCC3C(C1CCC2(C)O)CCC4=CC(=O)C=CC34C	3.6	1000
31	CC1CC(=O)CC2C1(C3CCC4(C(C3CC2)CCC4O)C)C	4.1	980
32	CC1=CC2C(CCC3(C2CCC3(C(=O)C)OC(=O)C)C)C4(C1=CC(=O)CC4)C	3.1	980
33	CC12CCC3C(C1CCC2C(=O)COC(=O)C(C)(C)C)CCC4=CC(=O)CCC34C	4.5	980
34	CCCCCC(=O)OC1(CCC2C1(CCC3C2CCC4=CC(=O)CCC34C)C)C(=O)C	5.7	980
35	CC1CC2(C(CCC3C2CCC4(C3CCC4O)C)CC1=O)C	4.2	980
36	CC12CCC(=O)C=C1CCC3C2CCC4(C3CCC4(C#C)O)C	3.5	2000
37	CC12CCC3C(C1CCC2(C#C)O)CCC4=C3CCC(=O)C4	2.1	980
38	CCC(=O)OC1CCC2C1(CCC3C2CCC4C3(CC(C(=O)C4)C)C)C	5.3	980
39	CC1=CC(=O)CC2C1(C3CCC4(C(C3CC2)CCC4OC(=O)C)C)C	4.4	4000
40	CC12CCC(=O)C(=C1CCC3C2CCC4(C3CCC4(C)O)C)O	3.2	980

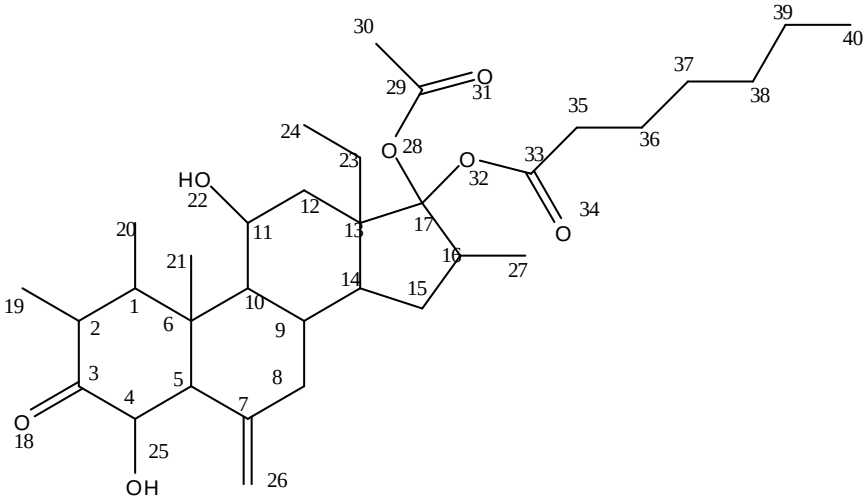

Figure 1. Hypermolecule

A hypermolecule is a representation of a dataset that seeks to maximize the degree of molecular overlap between structures while preserving the geometry and molecular connectivity of each molecule in the dataset [22]. The hypermolecule in Figure 1 was built up by superposing all the 40 molecules under the study (see [23] for the algorithm of superposing). According to the numbering of the hypermolecule positions, *binary vectors* were constructed, with 1 when in the current molecule there exists a corresponding atom and zero, otherwise.

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 [24]. Topological indices have beed computed by TOPOCLUJ software [25]; some of them (Centric index of partial charge shells=Cen, Total adjacency = Adj, Detour = De, Distance = Di, 3D-distance D3D, Connectivity = C, SD), HOMO (in au) and log P are listed in Table 2.

The models fit abilities were assessed by leave-one-out analysis [26] using a dedicated software [27, 28].

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 smaller than 20% and those with intercorrelation larger than 0.80 have been discarded.

Correlation weighting was performed on all the positions in the hypermolecule: the correlating coefficients of the statistically significant positions of the hypermolecule (i.e. the best multilinear regression) were used to multiply the local descriptors, e.g. the hydride mass fragments, thus resulting new weighted vectors CD_{ij}. Next, the local correlating descriptors are summed to give a global descriptor, $SD_i = \sum_j CD_{ij}$. This new descriptor SD_1 is a linear combination of the local correlating descriptors for the significant positions in the hypermolecule (i.e. H11, H17, H22, H28, H32, H34, H35, H36, H37, H38, H39).

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 3. The number of descriptors was limited to four, to fulfill the considerations of Topliss and Costello [29], see also ref. [30].

(i) Monovariate regression

$$\log P = -0.787 + 0.963 \times SD_1$$

(ii) Bivariate regression

$$\log P = -0.967 + 0.938 \times SD_1 + 0.00006 \times De$$

(iii) Three-variate regression

$$\log P = -1.702 + 0.931 \times SD_1 + 0.193 \times Adj. - 0.136 \times C$$

(iv) Four-variate regression

$$\log P = -2.501 + 1.115 \times SD_1 - 0.062 \times CjDe. - 0.0003 \times CfDi + 0.062 \times CfDe$$

Table 2. Topological indices, correlating descriptors, log P and LD50 for the set of testosterone in Table 1

Mol.	log P	LD50	Homo	Di	De	CjDi	CfDi	SD_1	SD_2	SD_3	SD_4
1	3.3	5000	-9.92	802	2969	1505	1739	4.74	18144.21	13913.02	2.40
2	4.4	1000	-9.84	1392	4375	2512	2809	5.79	17592.12	12672.72	3.14
3	3.4	2500	-9.91	907	3278	1701	1964	4.44	18144.21	14664.59	2.11
4	3.7	980	-10.47	802	2969	1505	1739	4.47	17442.47	11639.19	1.98
5	6.3	1000	-9.57	2340	6139	3901	4237	6.79	17592.12	12377.47	4.56
6	5.1	595	-9.68	2637	6850	4480	4766	6.67	17592.12	12136.68	3.40
7	5.3	980	-9.99	1812	5203	3151	3469	5.92	17592.12	12067.06	3.93
8	3.9	980	-9.85	1220	3999	2233	2517	5.17	17695.45	12666.99	2.64
9	3	2000	-9.94	941	3286	1755	1977	4.14	18209.20	13397.45	1.51
10	3.1	980	-9.73	892	3270	1667	1971	4.47	17442.43	11090.60	1.68
11	3.5	980	-9.85	1358	4282	2460	2728	4.96	17540.46	11880.55	2.44
12	4.1	6400	-9.41	1759	5415	3179	3617	4.19	22577.06	17378.92	2.56
13	4.7	980	-9.85	1220	3999	2233	2517	5.26	17540.46	12664.34	2.64
14	3.3	5010	-9.74	1053	3609	1960	2214	4.14	21472.44	14441.05	1.50
15	5.6	980	-8.26	2194	6305	3937	4284	6.61	17695.45	11948.77	4.08
16	3.5	980	-9.94	1034	3609	1929	2211	4.14	18209.20	12672.89	1.84
17	3.3	980	-9.94	1150	3953	2134	2469	3.93	17442.44	11474.94	1.61
18	3.5	980	-10.08	802	2969	1505	1739	4.74	18144.21	11558.16	1.98
19	2.7	980	-9.42	1149	3971	2130	2485	3.84	18721.20	12190.84	0.84
20	1.6	250	-9.99	1425	4677	2612	3018	2.71	18226.38	12618.41	0.08
21	1.6	5000	-9.61	1425	4677	2612	3018	2.71	18226.38	13930.92	0.56
22	3.5	980	-9.96	1556	5083	2849	3326	4.00	17689.01	13908.03	1.65
23	1.9	4000	-10.00	1425	4677	2612	3018	2.71	22221.26	13575.52	0.92
24	2.7	10000	-10.04	2199	6381	3858	4398	3.76	24988.97	20930.96	1.21
25	3.5	980	-9.94	1034	3609	1929	2211	4.14	18209.20	12672.80	1.84
26	3.7	10000	-10.39	907	3278	1701	1964	4.44	26462.43	21022.87	2.24
27	2.8	980	-8.96	1082	3620	1993	2229	4.46	17442.39	12305.90	1.09
28	5.6	980	-8.23	2194	6305	3937	4284	6.61	17695.45	11949.46	4.08
29	1.6	980	-9.65	1265	4302	2342	2730	2.68	17353.76	12018.76	-0.06
30	3.6	1000	-10.05	822	2974	1543	1746	4.44	18144.21	12015.91	2.42
31	4.1	980	-10.40	895	3283	1677	1967	4.74	19254.35	12696.49	3.04
32	3.1	980	-9.54	1620	5011	2926	3277	4.19	17442.40	12043.57	1.69
33	4.5	980	-10.11	2288	5984	3824	4104	5.83	15936.30	11831.31	3.04
34	5.7	980	-9.77	3027	6201	4528	4843	6.55	17592.12	10816.74	4.19
35	4.2	980	-10.31	908	3272	1703	1973	4.74	17718.53	11343.95	2.83
36	3.5	2000	-9.94	1034	3609	1929	2211	4.14	18209.20	12673.23	1.84
37	2.1	980	-9.58	941	3286	1755	1977	4.14	18209.20	13015.55	1.10
38	5.3	980	-10.13	1542	4750	2788	3121	5.79	17166.44	12593.93	3.83
39	4.4	4000	-9.96	1342	4369	2458	2801	5.26	18650.61	14305.09	2.62
40	3.2	980	-9.40	929	3296	1747	1992	4.44	18144.21	11803.26	1.48

Table 3. Alternatives and best models in describing log P in the training set of testosterone in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_1**	**0.896**	0.892	0.381	198.45
2	**CjDi**	0.216	0.181	1.046	6.32
3	**Di**	0.242	0.209	1.029	7.33
4	**SD_1, De**	**0.9**	0.891	0.382	98.93
5	**SD_1, D3D**	0.9	0.891	0.382	98.91
6	**SD_1, HOMO**	0.896	0.887	0.389	95.17
7	**SD_1, CjDi**	0.899	0.89	0.383	98.28
8	**SD_1, Di**	0.899	0.89	0.383	98.16
9	**SD_1, Adj, C**	**0.911**	0.899	0.368	71.93
10	**SD_1, De, HOMO**	0.9	0.886	0.391	63.01
11	**SD_1, HOMO, D3D**	0.9	0.886	0.391	62.94
12	**SD_1, CjDe, HOMO**	0.899	0.884	0.394	62.09
13	**SD_1, Di, HOMO**	0.899	0.885	0.392	62.48
14	**SD_1, CfDi, HOMO**	0.899	0.885	0.392	62.56
15	**SD_1, CfDi, CfDe, CjDe**	**0.924**	0.909	0.349	60.70
16	**SD_1, C, Adj, CjDi**	0.918	0.901	0.363	55.9
17	**SD_1, C, Adj, HOMO**	0.916	0.9	0.366	54.81

3.2.3. Model Validation (for log P)

(a) Leave-one-out

The performances in leave-one-out analysis related to the best models listed in Table 3 are presented in Table 4.

Table 4. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_1**	**0.876**	0.02	0.416	162.46
4	**SD_1, De**	**0.864**	0.036	0.435	146.59
9	**SD_1, Adj, C**	**0.872**	0.039	0.422	157.09
15	**SD_1, CfDi, CfDe, CjDe**	**0.873**	0.069	0.422	157.55

(b) External Validation

The values log P for the test set of testosterone (the last 15 structures in Table 1), were calculated by using the best trivariate equation in Table 3, entry 9. Data are listed in Table 5 and the monovariate correlation: $\log P = 0.845 \times \log P_{calc.} + 2.306$;n=15; R^2=0.722; s=0.652; F= 33.8 is plotted in Figure 2.

Table 5. Calculated values log P for the molecules in the test set (Table 1)

Mol.	log P	log $P_{calc.}$
26	3.7	3.7
27	2.8	5.86
28	5.6	7.39
29	1.6	3.53
30	3.6	5.08
31	4.1	5.68
32	3.1	4.9
33	4.5	6.68
34	5.7	7.33
35	4.2	5.68
36	3.5	4.77
37	2.1	4.44
38	5.3	6.75
39	4.4	6.06
40	3.2	5.27

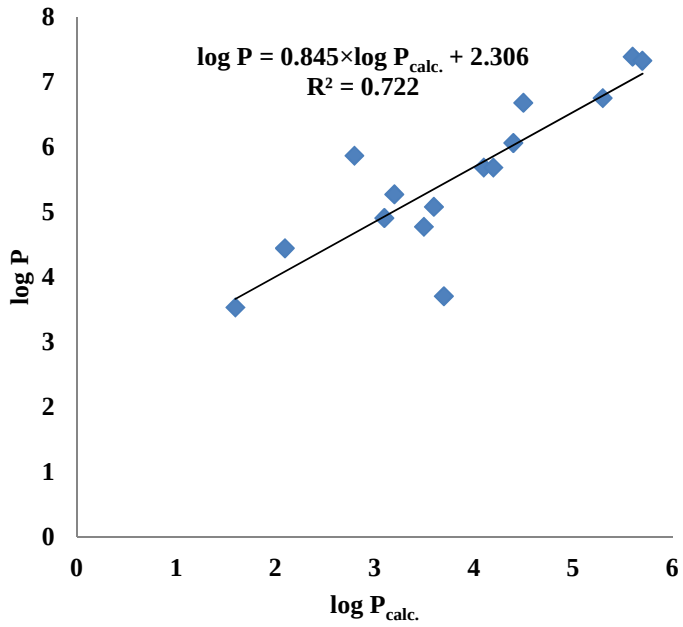

Figure 2. The plot log P vs. log $P_{calc.}$ for the test set (external validation)

(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. 9, Table 3. Data are listed in Table 6 and the monovariate correlation: : $\log P = 0.943 \times \log P_{calc.} + 0.376$; n=15; R^2=0.951; s=0.273; F=254.62 is plotted in Figure 3.

Table 6. Calculated values of log P by similarity clusters, for the molecules in the test set (Table 1)

Mol.	log P	log $P_{calc.}$
26	3.7	3.70
27	2.8	2.96
28	5.6	5.43
29	1.6	1.60
30	3.6	3.52
31	4.1	4.34
32	3.1	3.72
33	4.5	4.97
34	5.7	5.86
35	4.2	4.42
36	3.5	3.28
37	2.1	2.45
38	5.3	5.14
39	4.4	4.62
40	3.2	3.75

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 the SD_2 global descriptor, over the following significant positions in the hypermolecule: H3, H5, H7, H8, H11, H12, H19, H20, H22, H24, H26, H28, H32, H33, H34 and H35 and it will be used as the basis of modeling LD50 (see Table 2).

3.2.5. QSAR models (for LD50)

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 7.

(v) Monovariate regression

$$LD50 = -15924 + 0.972 \times SD_2$$

(vi) Bivariate regression

$$LD50 = -1682 + 0.989 \times SD_2 + 0.149 \times De$$

(vii) Three-variate regression

$$LD50 = -19850 + 0.985 \times SD_2 - 291.01 \times HOMO + 0.199 \times De$$

(viii) Four-variate regression $LD50 = -1819 + 1.001 \times SD_2 - 3.138 \times Di + 2.785 \times D3D + 0.498 \times De$

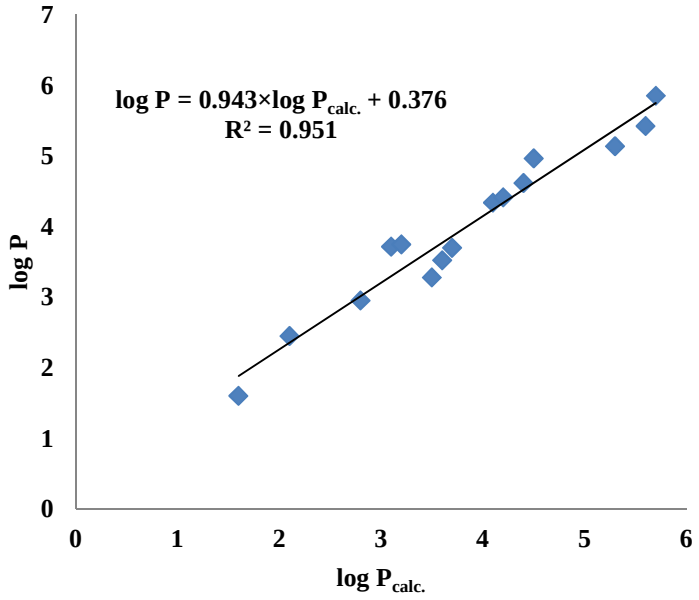

Figure 3. The plot log P vs. log $P_{calc.}$ by similarity clusters

Table 7. Best models in describing LD50 in the training set of testosterone in Table 1.

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_2**	**0.905**	0.900	691.882	217.88
2	**CjDe**	0.030	0.012	2205.542	0.70
3	**HOMO**	0.028	0.014	2207.211	0.70
4	**De**	0.019	-0.023	2217.499	0.45
5	**SD_2, De**	**0.910**	0.901	688.044	110.79
6	**SD_2, D3D**	0.910	0.902	685.537	111.68
7	**SD_2, Di**	0.908	0.899	694.960	108.38
8	**SD_2, CjDe**	0.907	0.898	699.782	106.74
9	**SD_2, HOMO**	0.905	0.896	705.211	104.93
14	**SD_2, HOMO, De**	**0.913**	0.901	691.310	73.43
13	**SD_2, HOMO, Adj.**	0.913	0.900	693.118	73.01
11	**SD_2, CJdi, CjDe**	0.912	0.899	697.003	72.12
16	**SD_2, HOMO, D3D**	0.912	0.900	694.689	72.65
15	**SD_2, De, D3D**	0.910	0.898	701.668	71.07
12	**SD_2, Di, C**	0.909	0.895	708.750	69.52
17	**SD_2, De, D3D, Di**	**0.924**	0.909	660.919	60.99
19	**SD_2, CjDi, D3D, De**	0.923	0.908	666.553	59.88
18	**SD_2, D3D, Di, C**	0.921	0.905	673.864	58.48

3.2.6. Model Validation (for LD50)

(a) Leave-one-out

The performances in leave-one-out analysis related to the best models in Table 7 are presented in Table 8.

Table 8. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_2**	**0.891**	0.013	738.243	188.58
5	**SD_2, De**	**0.889**	0.021	747.599	183.31
14	**SD_2, De, HOMO**	**0.885**	0.028	760.642	176.3
17	**SD_2, De, D3D, Di**	**0.897**	0.027	719.338	199.84

(b) External Validation

The values LD50 for the test set of testosterone (Table 1, last 15 structures) were calculated by using the best equation in Table 7, entry 14. Data are listed in Table 9 and the monovariate correlation: $LD50 = 883.48 + 0.655 \times LD50_{calc.}$;n=15; R^2=0.658; s=1618.663; F= 24.96 is plotted in Figure 4.

Table 9. Calculated values of LD50 for the molecules in the test set (Table 1)

Mol.	LD50	$LD50_{calc.}$
1	5000	1497.73
6	595	1660.89
8	980	1240.10
15	980	1240.63
16	980	1694.42
17	980	1009.40
18	980	1544.38
19	980	2120.44
20	250	1940.07
21	5000	1829.83
22	980	1483.04
23	4000	5876.90
24	10000	8954.71
25	980	1694.42
30	1000	1536.21

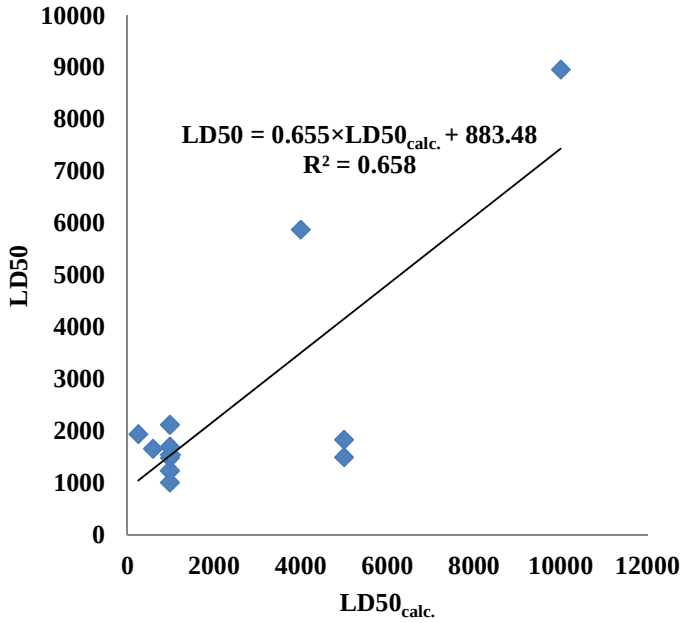

Figure 4. The plot LD50 vs. $LD50_{calc.}$ for the test set (external validation)

(c) Similarity Cluster Validation

Validation was performed by calculating LD50 for the molecules in the test set, similar to the Section 3.2.3. The values $LD50_{calc}$. were computed with the same descriptors as in eq. 14, Table 7. Data are listed in Table 10 and the

monovariate correlation: $LD50 = 0.877 \times LD50_{calc.} + 142.53$; n=15; R^2=0.935; s=706.157; F=186.44 is plotted in Figure 5.

Table 10. Calculated values of LD50 by similarity clusters, for the molecules in the test set

Mol.	LD50	$LD50_{calc.}$
1	5000	2507.05
6	595	881.54
8	980	1194.66
15	980	1082.21
16	980	954.23
17	980	985.43
18	980	922.69
19	980	922.69
20	250	612.35
21	5000	4839.34
22	980	944.58
23	4000	4498.85
24	10000	9413.14
25	980	954.23
30	1000	952.63

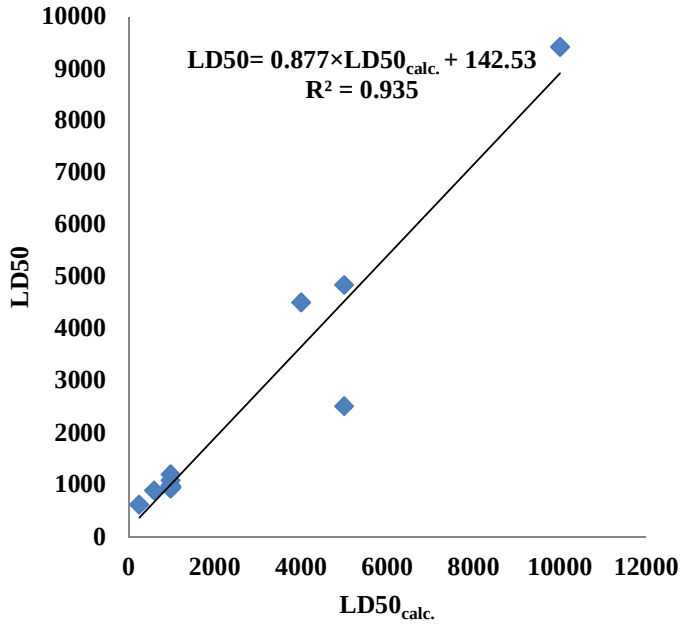

Figure 5. The plot LD50 vs. $LD50_{calc.}$ by similarity clusters

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 3.1.1. The local correlating descriptors, actually the partial charges (computed at HF level of theory) are summed over the following significant positions in the hypermolecule: H1, H3, H5, H8, H9, H11, H15, H18, H21, H21, H28, H29, H32, H33, H35, H36, H37); the resulting SD_3 global descriptor will be used as the basis of modeling log P (see Table 2).

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 11.

(ix) Monovariate regression

$$\log P = 1.002 \times SD_3 - 10.806$$

(x) Bivariate regression

$$\log P = 2.938 + 0.930 \times SD_3 + 0.130 \times HOMO$$

(xi) Three-variate regression

$$\log P = 2.561 + 0.907 \times SD_3 + 0.099 \times HOMO + 0.00005 \times CjDi$$

(xii) Four-variate regression

$$\log P = 2.379 + 0.863 \times SD_3 - 0.002 \times Di + 0.003 \times CjDi - 0.001 \times De$$

Table 11. Best models in describing log P in the training set of testosterone in Table 1.

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_3**	**0.955**	0.953	0.223	482.57
3	**Di**	0.370	0.343	0.829	13.53
2	**CjDi**	0.355	0.327	0.839	12.66
4	**De**	0.298	0.267	0.875	9.75
5	**SD_3, HOMO**	**0.959**	0.955	0.216	256.80
6	**SD_3, CjDi**	0.958	0.954	0.219	251.77
7	**SD_3, Di**	0.958	0.954	0.219	250.76
8	**SD_3, De**	0.958	0.954	0.219	249.49
9	**SD_3, D3D**	0.957	0.953	0.222	242.89
10	**SD_3, CjDe, HOMO**	**0.961**	0.955	0.216	171.69
11	**SD_3, Di, D3D**	0.961	0.955	0.217	170.97
12	**SD_3, HOMO, D3D**	0.960	0.955	0.218	169.04
13	**SD_3, Adj, C**	0.959	0.953	0.222	162.38
14	**SD_3, De, CjDi**	0.958	0.952	0.223	161.11
15	**SD_3, D3D, De**	0.957	0.953	0.222	242.89
16	**SD_3, Di, CjDi, De**	**0.962**	0.954	0.220	124.85
17	**SD_3, HOMO, D3D, De**	0.960	0.952	0.223	120.87
18	**SD_3, C, Adj, Di**	0.959	0.951	0.227	116.50

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 11 are presented in Table 12.

Table 12. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. $Error_{loo}$	F_{loo}
1	**SD_3**	**0.949**	0.006	0.235	429.64
5	**SD_3, HOMO**	**0.953**	0.006	0.227	462.23
10	**SD_3, CjDe, HOMO**	**0.952**	0.009	0.227	461.89
16	**SD_3, Di, CjDi, De**	**0.951**	0.011	0.23	451.26

Table 13. Calculated values of log P for the molecules in the test set (Table 1)

Mol.	log P	log $P_{calc.}$
1	3.3	3.83
2	4.4	4.55
3	3.4	3.58
4	3.7	3.40
5	6.3	5.94
18	3.5	3.44
19	2.7	2.51
20	1.6	1.78
21	1.6	2.26
22	3.5	3.20
23	1.9	2.54
37	2.1	2.71
38	5.3	5.15
39	4.4	4.06
40	3.2	3.07

(b) External Validation

The values log P for the test set of testosterone (Table 1),were calculated by using the best equation in Table 11, entry 10. Data are listed in Table 13 and the monovariate correlation: $\log P = 0.809 \times \log P_{calc.} + 0.722$; n=15; R^2=0.938; s=0.347; F=195.43 is plotted in Figure 6.

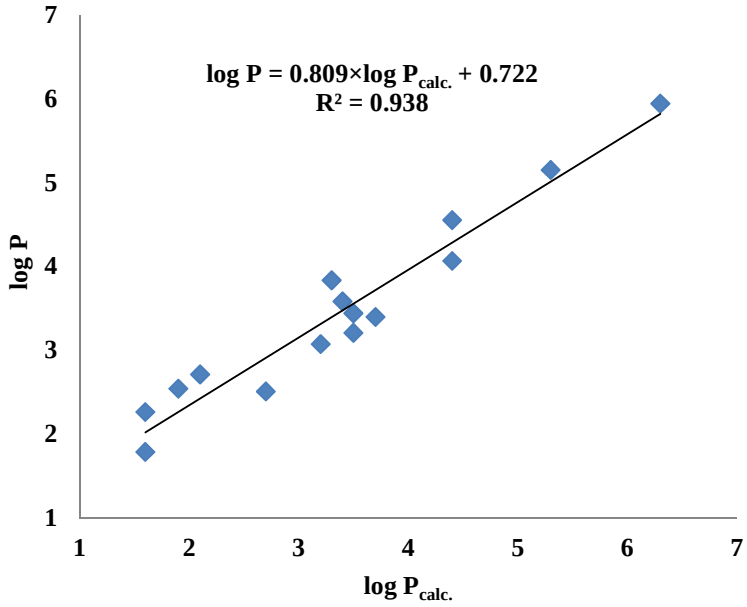

Figure 6. The plot log P vs. log P $_{calc.}$ for the test set (external validation)

(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 11. Data are listed in Table 14 and the monovariate correlation: $\log P = 0.900 \times \log P_{calc.} + 0.443$; n=15; R^2=0.97; s=0.240; F=420.87 is plotted in Figure 7.

3.3.4. Data reduction (for LD50)

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, H5, H6, H7, H10, H12, H14, H16, H19, H23, H29, H32, H34, H38); the resulting SD_4 global descriptor will be used as the basis of modeling LD50 (see Table 2).

Table 14. Calculated values of log P by similarity clusters, for the molecules in the test set

Mol.	log P	log $P_{calc.}$
1	3.3	3.65
2	4.4	4.42
3	3.4	3.62
4	3.7	3.71
5	6.3	6.13
18	3.5	3.48
19	2.7	2.52
20	1.6	1.59
21	1.6	2.12
22	3.5	3.31
23	1.9	2.41
37	2.1	2.56
38	5.3	5.45
39	4.4	4.27
40	3.2	3.23

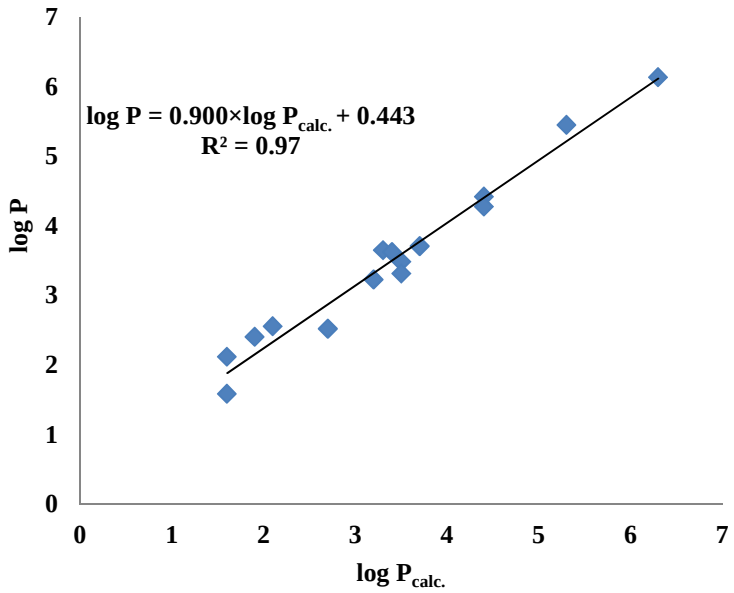

Figure 7. The plot log P vs. log $P_{calc.}$ by similarity clusters

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 15.

(xiii) Monovariate regression

$$LD50 = 1.001 \times SD_4 - 10834$$

(xiv) Bivariate regression

$$LD50 = -7307.5 + 1.008 \times SD_4 + 371.294 \times HOMO$$

(xv) Three-variate regression

$$LD50 = -7056.5 + 1.022 \times SD_4 + 464.718 \times HOMO + 0.351 \times D3D$$

(xvi) Four-variate regression

$$LD50 = -6384.02 + 1.034 \times SD_4 + 434.794 \times HOMO - 0.882 \times De + 1.282 \times CjDi$$

Table 15. Best models in describing LD50 in the training set of testosterone in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD_4**	**0.898**	0.893	752.054	201.84
2	**Di**	0.010	0.033	2339.239	0.24
3	**De**	0.004	0.039	2346.122	0.10
4	**HOMO**	0.001	0.043	2350.891	0.01
5	**SD_4, HOMO**	**0.906**	0.897	737.604	105.87
6	**SD_4, Di**	0.900	0.890	761.958	98.52
7	**SD_4, CjDi**	0.899	0.889	765.237	97.58
8	**SD_4, D3D**	0.899	0.890	762.955	98.23
9	**SD_4, De**	0.898	0.889	767.709	96.88
10	**SD_4, HOMO, D3D**	**0.911**	0.898	734.533	71.56
11	**SD_4, CjDi, HOMO**	0.910	0.897	738.119	70.80
12	**SD_4, De, CjDi**	0.905	0.891	759.934	66.40
13	**SD_4, D3D, De**	0.902	0.887	772.154	64.09
14	**SD_4, Di, D3D**	0.900	0.885	779.449	62.77
15	**SD_4, Adj, C**	0.898	0.884	784.732	61.83
16	**SD_4, HOMO, De, CjDi**	**0.915**	0.897	737.081	53.52
17	**SD_4, HOMO, D3D, Di**	0.911	0.893	751.261	51.33
18	**SD_4, C, Adj, De**	0.903	0.884	784.998	46.59

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 15 are presented in Table 16.

Table 16. Leave-one-out analysis for best log P models

	Descriptors	Q^2	R^2-Q^2	St. Error$_{loo}$	F_{loo}
1	**SD_4**	**0.885**	0.013	797.377	177.01
5	**SD_4, HOMO**	**0.886**	0.02	793.365	179.03
10	**SD_4, HOMO, D3D**	**0.886**	0.025	792.731	179.36
16	**SD_4, HOMO, De, CjDi**	**0.882**	0.032	805.922	172.79

(a) External Validation The values LD50 for the test set of testosterone (Table 1),were calculated by using the best equation in Table 15, entry 10. Data are listed in Table 17 and the monovariate correlation: $LD50 = 0.830 \times LD50_{calc.} + 839.36$; n=15; R^2=0.840; s=1039.906; F=68.31 is plotted in Figure 8.

Table 17. Calculated values of LD50 for the molecules in the test set

Mol.	LD50	LD50$_{calc.}$
1	5000	2967.99
2	1000	1745.51
3	2500	3699.88
4	980	530.19
5	1000	1834.74
6	595	1619.14
7	980	1563.68
8	980	1751.13
9	2000	2467.30
22	980	3051.90
23	4000	2819.61
24	10000	10376.48
25	980	1666.63
37	980	1973.52
38	980	1886.35

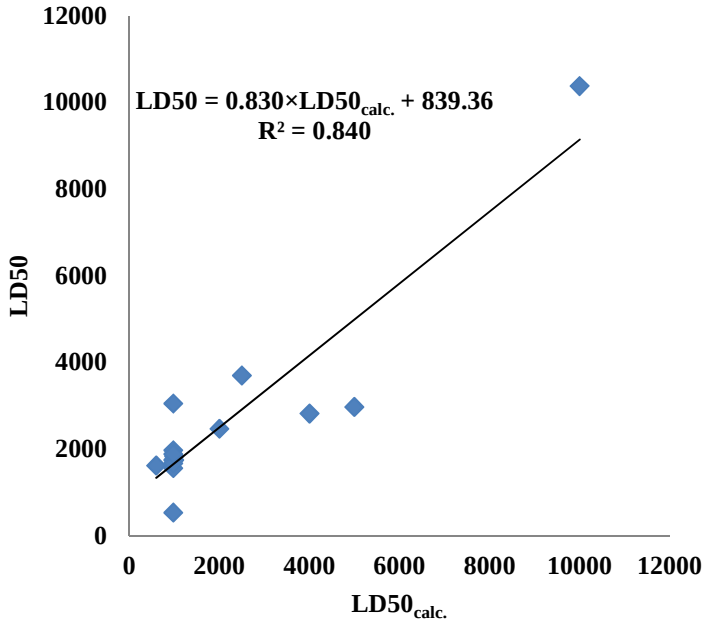

Figure 8. The plot LD50 vs. LD50$_{calc.}$ (external validation)

(b)Similarity Cluster Validation

Validation was performed by calculating LD50 for the molecules in the test set, similar to that in the Section 3.2.3. The values $LD50_{calc}$. were computed with the same descriptors as in eq. 10, Table 15. Data are listed in Table 18 and the monovariate correlation: $LD50 = 464.29 + 0.832 \times LD50_{calc.}$;n=15; R^2=0.943; s=622.801; F= 213.68 is plotted in Figure 9.

Table 18. Calculated values of LD50 by similarity clusters for the molecules in the test set (Table 1)

Mol.	LD50	$LD50_{calc.}$
1	5000	3632.01
2	1000	1382.42
3	2500	3207.06
4	980	1225.48
5	1000	860.77
6	595	350.34
7	980	1372.07
8	980	1282.10
9	2000	2463.65
22	980	1554.78
23	4000	2913.17
24	10000	9339.85
25	980	1722.29
37	980	1303.88
38	980	1758.71

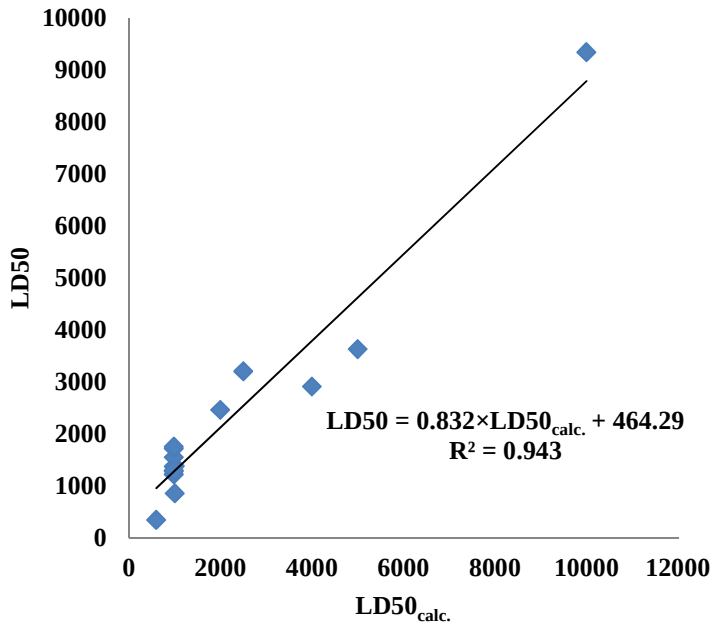

Figure 9. The plot LD50 vs. $LD50_{calc.}$ for the test set by similarity clusters

CONCLUSION

A set of 40 testosterone, 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.

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. This last case provided accurate predictions, that originate in the „quasi-congeneric" state of the clustered structures.

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

REFERENCES

[1] KJ Yang; SC Lin; SJ Huang; WM Ching; CH Hung; DLM Tzou; *Steroids*, **2014**, 80, 64–70.
[2] RC Kammerer, Drug testing and anabolic steroids. In: Yesalis CE, editor. 2, **2000**, 415–59.
[3] NT Shahidi; *Clin Ther*, **2001**, 23(9), 1355–90.
[4] PY Liu; AK Death; DJ Handelsman; *Endocr. Rev.*, **2003**, 24, 313—340
[5] K Choong; S Basaria; *Aging Male*, **2010**, 13, 1—9.
[6] AM Traish; MM Miner; A Morgentaler; M Zitzmann; *Am. J. Med*., **2011**, 124, 578—587.
[7] D Pereg; J Chan; E Russell; T Berlin; M Mosseri; JA Seabrook; G Koren; S Van Uum; *Psychoneuroendocrinology*, **2013**, 38, 2875—2882.
[8] J Arnlo¨v; MJ Pencina; S Amin; BH Nam; EJ Benjamin; JM Murabito; TJ Wang; PE Knapp; RB D'Agostino Sr.; S. Bhasin; RS Vasan; *Ann. Intern. Med.*, **2006,** 145, 176-184.
[9] EA Jankowska; B Biel; J Majda; A Szklarska; M Lopuszanska; M Medras; SD Anker; W Banasiak; PA Poole-Wilson; P Poni-Kowski; *Circulation*, **2006**, 114, 1829—1837.
[10] T Nishiyama, Urologic Oncology:Seminars and Original Investigations, **2014**, 38(32), 17–38.
[11] O Weisman; O Zagoory-Sharon, R Feldman, *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, **2014**, 49, 47–52.
[12] O Prado-Prado; V Martinez; E Uriarte; FM Ubeira; KC Chou; H González Díaz; *Bioorg. Med. Chem.* **2008**, doi:10.1016/j.bmc.**2008**.11.075.
[13] R Wang; Y Gao; L Lai; *Perspect. Drug Discov. Des*., **2000**, 19, 47-66.
[14] A Hayes; *Taylor and Francis*, 4th Ed, 18, **2001**.
[15] RP Sheridan; and SK Kearsley; *Drug Discovery Today*, 7, **2002**, 903.
[16] HJ Bohm, and G Schneider, Wiley-VCH, Weinheim, **2000**.
[17] A Bender; and RC Glen; *Org. Biomol.Chem*., **2004**, 2, 3204-3218.
[18] MV Diudea; *MATCH Commun. Math. Comput. Chem.*, **1997**, 35, 169–183.
[19] MV Diudea; I Gutman; *Croat. Chem. Acta.*, **1998**, 71, 21–51.
[20] N Dorosti; A Iranmanesh; and MV Diudea; *MATCH Commun. Math. Comput. Chem.*, **2009**, 62, 389-395.
[21] PubChem database, accessed 23.05. **2013**.
[22] NJ Richmond; CA Abrams; PR Wolohan; E Abrahamian; P Willett; RD Clark; *J Comput Aided Mol Des*, **2006**, 20, 567–587.
[23] AT Balaban, A Chiriac, I Motoc, and Z Simon, *Steric Fit in QSAR*, *Lectures Notes in Chemistry*, *Springer*, Berlin, **1980**, 15.
[24] **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**.
[25] O Ursu; MV Diudea; TopoCluj software program. Babes-Bolyai University Cluj, **2005**.
[26] DM Hawkins; SC Basak; and D Mills; *J. Chem. Inf. Comput. Sci.*, **2003**, 43, 579-586.
[27] SD Bolboacă; L Jäntschi; and MV Diudea; *Current Computer-Aided Drug Design*, **2013**, 9(2), 195-205.
[28] 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/
[29] JG Topliss; RJ Costello; *J. Med. Chem*., **1972**, 15, 1066-1068.
[30] TE Harsa; AM Harsa; B Szefler; *Cent. Eur. J. Chem*., **2014**, 12, 365-376.