Supporting Information

Is your QSAR/QSPR descriptor real or trash?

*Rudolf Kiralj, Márcia M. C. Ferreira**

Laboratory of Theoretical and Applied Chemometrics, Institute of Chemistry, University of Campinas, 13083-740 Campinas SP, Brazil

CONTENTS

Table S1. Statistics for the sign change problem in correlation and regression vectors.*

Data[a]	Model	Split	Ref.	Descriptor	r_c [a]	r_t [b]	r_e [c]	β_c [d]	β_t [e]	F_1 [f]	F_2 [g]	F_3 [h]	F_4 [i]
1	MLR	75/78	[1]	LogK_{ow}	0.8807	0.8897	0.8710	0.9340	0.9491	0.8852	0.8758	0.9070	0.9143
	QSAR			pK_a	0.0249	0.0819	-0.0469	0.1185	0.1316	0.0452	-0.0342	0.0543	0.0573
				E_{LUMO}	-0.0998	-0.1061	-0.0928	-0.3331	-0.2855	0.1029	0.0962	0.1823	0.1688
				E_{HOMO}	-0.0058	0.0210	-0.0387	0.0328	0.0080	-0.0110	0.0150	-0.0138	-0.0068
				N_{hdon}	-0.4100	-0.4121	-0.4077	0.0391	0.0182	0.4110	0.4088	-0.1266	-0.0864
2	PLS	56/30	[1]	CYP51-g	-0.7223	-0.7206	-0.7285	-0.3380	-0.3355	0.7214	0.7254	0.4941	0.4923
	QGSAR			CYP51-e	-0.7263	-0.7264	-0.7288	-0.3719	-0.3901	0.7264	0.7275	0.5197	0.5323
				PMR1-t	-0.5023	-0.4612	-0.5867	-0.2800	-0.2720	0.4813	0.5429	0.3750	0.3696
				CYP51-e*Npi	-0.6238	-0.6306	-0.6094	0.4894	0.5012	0.6272	0.6166	-0.5525	-0.5592
				PCR*Npi	-0.5558	-0.5678	-0.5288	-0.3883	-0.4131	0.5618	0.5421	0.4646	0.4792
				PMR1-e*Lpi	-0.6775	-0.6899	-0.6558	-0.1113	-0.1419	0.6836	0.6665	0.2746	0.3101
				CYP51-e*Lpi	-0.6337	-0.6612	-0.5792	0.4173	0.3981	0.6473	0.6059	-0.5142	-0.5023
				PCR*Lpi	-0.5635	-0.5898	-0.5101	-0.3037	-0.2465	0.5765	0.5362	0.4137	0.3727
3	PLS	40/10	[1]	E_e	-0.8561	-0.8445	-0.9166	-0.2401	-0.2481	0.8503	0.8858	0.4534	0.4609
	QSPR			E_{CC}	-0.8920	-0.8842	-0.9589	-0.1475	-0.1267	0.8881	0.9249	0.3627	0.3362
				Q_{Omul}	0.9282	0.9176	0.9753	0.2833	0.2429	0.9229	0.9515	0.5128	0.4748
				Δ_{HL}	-0.8267	-0.8435	-0.7463	-0.5277	-0.4888	0.8351	0.7854	0.6605	0.6357
				σ_b	0.8619	0.8580	0.8984	0.0863	0.0848	0.8600	0.8800	0.2727	0.2704
				σ_r	-0.8905	-0.8989	-0.8526	-0.6669	-0.7204	0.8947	0.8714	0.7706	0.8010
				D_{CC}	0.9069	0.8976	0.9769	0.2044	0.1733	0.9022	0.9412	0.4305	0.3964
				Q_{C2mul}	0.8915	0.8863	0.9172	0.2607	0.2613	0.8889	0.9043	0.4821	0.4827
4	MLR	13/2	[1]	ClogP	0.6323	0.6977	(-)	0.7856	0.8072	0.6642	-	0.7048	0.7144
	QSAR			MgVol	-0.1070	-0.0412	(-)	-0.5450	-0.5068	0.0664	-	0.2415	0.2329
				$B1_{X.2}$	-0.4509	-0.4485	(x)	-0.2929	-0.3027	0.4497	-	0.3634	0.3695
5	PLS	13/2	[1]	E_e	-0.8159	-0.8135	(-)	-0.2900	-0.3288	0.8147	-	0.4864	0.5180
	QSPR			E_{CC}	-0.8587	-0.8565	(-)	-0.0475	-0.0617	0.8576	-	0.2020	0.2302
				Q_{Omul}	0.9204	0.9179	(-)	0.3191	0.3322	0.9192	-	0.5420	0.5530
				Δ_{HL}	-0.8266	-0.8257	(-)	-0.4814	-0.5493	0.8261	-	0.6308	0.6738
				σ_b	0.8459	0.8411	(+)	0.0378	0.0558	0.8435	-	0.1788	0.2173
				σ_r	-0.8550	-0.8444	(-)	-0.7026	-0.6179	0.8496	-	0.7750	0.7268
				D_{CC}	0.8718	0.8680	(+)	0.0933	0.1021	0.8699	-	0.2852	0.2983
				Q_{C2mul}	0.8943	0.8885	(+)	0.2760	0.2840	0.8914	-	0.4968	0.5040
6	MLR	20/20	[2]	X3A	0.6456	0.7673	0.4883	0.6055	0.6385	0.7038	0.5615	0.6252	0.6420
	QSAR			BEHv2	0.3023	0.0175	0.5984	0.6263	0.5597	0.0727	0.4253	0.4351	0.4113
				R7v	-0.5008	-0.6369	-0.3267	-0.4910	-0.5282	0.5648	0.4045	0.4959	0.5143
7	MLR	29/7[j]	[3]	HE	-0.4734	-0.4567	-0.5675	-0.0686	-0.0689	0.4650	0.5183	0.1802	0.1806

	QSAR			DM_z	0.0103	0.0990	-0.3219	-0.0777	-0.0724	0.0319	-0.0575	-0.0282	-0.0273
				DM_t	0.0243	-0.0176	0.1884	0.0566	0.0549	-0.0206	0.0676	0.0371	0.0365
				Q_{mean}	0.6287	0.6842	0.3280	0.0742	0.0764	0.6559	0.4541	0.2160	0.2192
				SSC	-0.3238	-0.3587	-0.0901	-0.3026	-0.3189	0.3408	0.1708	0.3130	0.3214
				X5	-0.2754	-0.3206	0.0059	0.7521	0.7464	0.2972	-0.0403	-0.4551	-0.4534
				S0K	-0.3317	-0.3713	-0.0784	-0.5607	-0.5581	0.3509	0.1613	0.4312	0.4302
				PW2	-0.1359	-0.0980	-0.4165	-0.0948	-0.1044	0.1154	0.2379	0.1135	0.1191
8	PLS	56/30	[4]	PMR1-g	-0.4906	-0.4475	-0.5797	-0.4129	-0.3823	0.4686	0.5333	0.4501	0.4331
	QGAR			PMR1-e	-0.6942	-0.7097	-0.6629	-0.2424	-0.2973	0.7019	0.6784	0.4102	0.4543
				CYP51-g	-0.7223	-0.7206	-0.7285	-0.4269	-0.4294	0.7214	0.7254	0.5553	0.5569
				CYP51-e	-0.7263	-0.7264	-0.7288	-0.4291	-0.4331	0.7264	0.7275	0.5582	0.5608
				PCR	-0.7221	-0.7203	-0.7285	-0.4270	-0.4295	0.7212	0.7253	0.5553	0.5569
				PMR1-t	-0.5023	-0.4612	-0.5867	-0.4713	-0.4571	0.4813	0.5429	0.4865	0.4791
9	MLR	87/43	[5]	1/SIC2	-0.6733	-0.6932	-0.5387	-0.7327	-0.7606	0.6832	0.6023	0.7024	0.7156
	QSAR			1/DPSA3	-0.4665	-0.5081	-0.2630	0.2393	0.2658	0.4869	0.3503	-0.3341	-0.3521
				1/HPCSA	-0.6506	-0.5610	-0.8036	-0.4806	-0.4325	0.6041	0.7231	0.5592	0.5305
				DPSA1	0.2978	0.2572	0.2893	-0.4182	-0.4046	0.2768	0.2935	-0.3529	-0.3471
10	MLR	44/20[k]	[6]	E	0.7525	0.7503	0.7607	0.3002	0.2922	0.7514	0.7566	0.4753	0.4689
	LFER			S	0.5294	0.5236	0.5426	-0.1558	-0.1505	0.5265	0.5360	-0.2872	-0.2823
				A	0.0286	0.0721	-0.0662	0.1541	0.1468	0.0454	-0.0435	0.0663	0.0647
				B	-0.0550	-0.0725	-0.0177	-0.4473	-0.4545	0.0632	0.0313	0.1569	0.1581
				V	0.8549	0.8610	0.8441	0.8135	0.8148	0.8579	0.8495	0.8339	0.8346
11	PLS	16/4	[7]	HBD/N	-0.7674	-0.8693	(-)	-0.4989	-0.5283	0.8168	-	0.6188	0.6367
	QSPR			Mor06u	0.5583	0.6738	(+)	0.3630	0.4095	0.6134	-	0.4502	0.4782
				Qcnpa	0.7772	0.7345	(+)	0.5053	0.4463	0.7555	-	0.6267	0.5890
				Ar	0.5797	0.6249	(+)	0.3769	0.3798	0.6019	-	0.4674	0.4692
				QNUnpa	-0.7248	-0.7539	(-)	-0.4712	-0.4581	0.7392	-	0.5844	0.5762
12	MLR	40/12	[8]	ACIC1	0.3642	0.3123	0.4644	-0.3495	-0.3460	0.3372	0.4113	-0.3568	-0.3550
	QSAR			MIA	-0.8135	-0.7809	-0.9140	-0.6981	-0.6826	0.7970	0.8623	0.7536	0.7452
				FNSA3	0.3290	0.3212	0.2343	0.1644	0.1669	0.3251	0.2776	0.2326	0.2343
				RPCS	-0.8188	-0.8144	-0.8308	-0.3211	-0.3150	0.8166	0.8248	0.5128	0.5079
				APMIA	-0.7472	-0.7224	-0.8314	0.5103	0.5360	0.7347	0.7882	-0.6175	-0.6329
13	MLR	31/11	[9]	S_aaCH	-0.2422	-0.3301	-0.0580	-0.3539	-0.2555	0.2828	0.1185	0.2928	0.2488
	QSAR			Shad_XYfrac	-0.1335	-0.2490	0.1095	-0.1399	-0.2594	0.1823	-0.1209	0.1367	0.1861
				Hbond_Acc	0.5257	0.6809	0.1066	0.8536	0.8222	0.5983	0.2367	0.6698	0.6574
				LUMO	0.1328	0.1684	0.0895	0.3557	0.4375	0.1495	0.1090	0.2173	0.2410
14	PLS	18/9	[10]	DE	0.9525	0.9587	0.9483	0.6348	0.6259	0.9556	0.9504	0.7776	0.7721
	QSAR			M_w	-0.9502	-0.9549	-0.9687	-0.6333	-0.6234	0.9525	0.9594	0.7757	0.7696
				E_{HOMO}	-0.6640	-0.7180	-0.6186	-0.4426	-0.4687	0.6905	0.6409	0.5421	0.5579
15	PLS[l]	16/16	[11]	TE	0.9443	0.9399	0.9488	0.6961	0.6010	0.9421	0.9465	0.8108	0.7533
	QSPR			R_e	-0.9308	-0.9213	-0.9410	-0.6851	-0.5107	0.9261	0.9359	0.7986	0.6895

				E_{LUMO}-E_{HOMO}	0.7589	0.9049	0.5868	0.0750	0.1541	0.8287	0.6673	0.2386	0.3420
				E_{LUMO}	0.7628	0.8785	0.6387	0.0643	0.0780	0.8186	0.6980	0.2215	0.2439
				$(E_{LUMO}$-$E_{HOMO})^2$	0.7486	0.8709	0.6016	0.0520	0.0346	0.8075	0.6711	0.1973	0.1609
				E_{HOMO}	-0.6653	-0.8321	-0.4818	-0.0825	-0.2594	0.7441	0.5662	0.2343	0.4154
				E_{LUMO}+E_{HOMO}	0.5252	0.5837	0.4876	0.0172	-0.1274	0.5537	0.5061	0.0950	-0.2587
				L_{CC}	0.4004	0.4973	0.2717	-0.0106	-0.1680	0.4462	0.3298	-0.0651	-0.2594
				Q_H^+	0.3920	0.4552	0.3621	0.1625	0.4851	0.4224	0.3768	0.2524	0.4361
16	MLR	18/4	[12]	ClogP	-0.5564	-0.4562	(-)	0.4451	0.4879	0.5038	-	-0.4976	-0.5210
	QSAR			CMR	-0.8559	-0.8078	(-)	-0.8916	-0.8682	0.8315	-	0.8736	0.8620
				Q_{C28}	-0.2335	-0.2265	(-)	-0.0837	-0.0902	0.2299	-	0.1398	0.1451
17	MLR	14/3[m]	[13]	SAS	0.6073	0.6290	(+)	0.5545	0.3659	0.6181	-	0.5803	0.4714
	QSRR			HOMO	0.8099	0.8565	(+)	0.7014	0.5446	0.8329	-	0.7537	0.6641
				Charge	-0.8951	-0.9559	(-)	-0.4478	-0.7547	0.9250	-	0.6331	0.8219
18	PLS	20/8	[14]	SASA	0.9369	0.9336	0.9524	0.9900	0.7094	0.9352	0.9446	0.9631	0.8152
	QSAR			ADDD	0.8844	0.8730	0.9014	-0.0914	0.6634	0.8787	0.8929	-0.2843	0.7660
				L/Bw	0.2722	0.3130	0.1402	0.1079	0.2378	0.2918	0.1953	0.1714	0.2544
19	MLR	68/55[n]	[15]	R_2	0.2792	0.2666	0.2910	0.0634	0.0452	0.2728	0.2851	0.1330	0.1123
	QSAR			$\Sigma\alpha_2^H$	-0.0069	0.0736	-0.1214	-0.0055	0.0059	-0.0225	0.0289	0.0061	-0.0064
				$\Sigma\beta_2^O$	-0.1016	-0.0792	-0.1567	-0.1856	-0.1425	0.0897	0.1262	0.1373	0.1203
				V_x	0.6653	0.6233	0.7311	0.1840	0.1002	0.6439	0.6974	0.3499	0.2582
				W	0.5055	0.4731	0.5685	0.6378	0.6521	0.4890	0.5361	0.5678	0.5742
				$^1\chi$	0.5826	0.5529	0.6228	0.1704	0.1693	0.5676	0.6024	0.3151	0.3141
				Log(RB)	0.4935	0.4537	0.5757	-0.7013	-0.7168	0.4732	0.5330	-0.5883	-0.5948
20	MLR	64/21	[16]	GATS1e	-0.3723	-0.3901	-0.3437	-0.1061	-0.1163	0.3811	0.3577	0.1987	0.2081
	QSAR			EEig08x	0.4664	0.4056	0.7166	-0.6313	-0.6198	0.4349	0.5781	-0.5426	-0.5376
				EEig07d	0.6113	0.5731	0.7692	0.7311	0.7375	0.5919	0.6857	0.6685	0.6715
				GGI6	0.6110	0.5547	0.8033	0.1448	0.1402	0.5822	0.7006	0.2975	0.2927
				R6v+	0.0635	0.1872	-0.4952	0.1300	0.1501	0.1091	-0.1774	0.0909	0.0977
				H-051	-0.4290	-0.3757	-0.6217	-0.1336	-0.1272	0.4015	0.5165	0.2394	0.2336
21	MLR	54/30[o]	[17]	S_{av}	0.6876	0.6594	0.7426	0.4422	0.4762	0.6734	0.7146	0.5514	0.5722
	QSAR			π_{R1}	0.4626	0.4209	0.5315	0.2111	0.2700	0.4413	0.4959	0.3125	0.3534
				I_1	0.5392	0.5543	0.5111	0.6191	0.6283	0.5467	0.5249	0.5778	0.5820
				I_2	0.4967	0.4655	0.5550	0.4044	0.3811	0.4808	0.5250	0.4482	0.4351
				I_{OH}	-0.5429	-0.5050	-0.6138	-0.4616	-0.4005	0.5237	0.5773	0.5006	0.4663
22	MLR	36/9[p]	[18]	log*P*	0.3231	0.3588	0.8180	0.7460	0.6141	0.3405	0.5141	0.4910	0.4455
	QSAR			*n*	-0.2772	-0.6384	0.1333	-0.6660	-0.7893	0.4206	-0.1922	0.4296	0.4677
23	MLR	50/33[q]	[19]	R.No.Cat	0.4987	0.5093	0.4595	0.5882	0.6225	0.5039	0.4787	0.5416	0.5572
	QSAR			HBdonCSA	-0.4713	-0.5451	-0.3465	-0.3871	-0.4042	0.5068	0.4041	0.4271	0.4365
				Av.v.Hat	-0.0588	-0.0367	-0.0643	0.4376	0.4432	0.0464	0.0615	-0.1604	-0.1614
				RNCh	-0.0277	0.0993	-0.2110	-0.4791	-0.4134	-0.0524	0.0764	0.1151	0.1069
				$(\log P)^2$	0.0115	0.0556	-0.0906	-0.1688	-0.1934	0.0253	-0.0322	-0.0440	-0.0471

				Av.e.r.Cat	-0.1439	-0.0556	-0.2759	0.2336	0.2108	0.0894	0.1992	-0.1833	-0.1741
24	MLR QSAR	234/226	[20][r]	HOMO	-0.6987	-0.6986	-0.7153	-0.7962	-0.7753	0.6986	0.7069	0.7459	0.7360
				nX	0.6160	0.6399	0.5092	0.3904	0.4183	0.6278	0.5600	0.4904	0.5076
				CIC0	-0.5786	-0.6045	-0.4900	-0.3285	-0.3547	0.5914	0.5325	0.4360	0.4530
				nCaH	-0.1122	-0.1598	-0.0693	0.3252	0.3133	0.1339	0.0882	-0.1910	-0.1875
25	MLR QSAR	15/4[s]	[21]	$(E_s)^2$	-0.5592	-0.5398	(-)	0.4830	0.4824	0.5494	-	-0.5197	-0.5194
				E_s	0.6375	0.6054	(+)	0.7542	0.7506	0.6213	-	0.6934	0.6918
				H–N–πS	0.3872	0.4133	(x)	0.2260	0.2368	0.4000	-	0.2958	0.3028
				–C=O electr char	0.8210	0.8055	(+)	0.3832	0.3845	0.8132	-	0.5609	0.5619
26	MLR QSAR	20/6[t]	[22][u]	α	0.9430	0.9235	(+)	0.9747	0.9557	0.9332	-	0.9587	0.9493
				E_{HOMO}	0.0743	0.1572	(x)	0.1810	0.2064	0.1081	-	0.1160	0.1238
				qH^+	-0.6168	-0.6583	(-)	-0.1309	-0.2099	0.6372	-	0.2841	0.3598
27	MLR QSAR	37/13[v]	[23]	$lgEnr_M$	0.8444	0.8400	0.8646	0.8857	0.8935	0.8422	0.8544	0.8648	0.8686
				$GAP_{h1\text{-}M}$	-0.2712	-0.2644	-0.2936	-0.3304	-0.3299	0.2678	0.2822	0.2993	0.2991
				$GAPV_{mM}$	-0.1497	-0.1226	-0.1419	-0.2753	-0.2628	0.1355	0.1457	0.2030	0.1984
				μ_M	0.0451	0.0261	0.1430	0.1747	0.1541	0.0343	0.0803	0.0888	0.0834
28	MLR QSAR	106/27	[24]	ALFA	0.9371	0.9416	0.9267	0.9677	0.9713	0.9393	0.9319	0.9523	0.9540
				MVC	-0.1846	-0.2264	0.0428	0.2402	0.2281	0.2044	-0.0889	-0.2106	-0.2052
				FPSA	-0.5021	-0.4558	-0.7380	0.0763	0.0683	0.4784	0.6087	-0.1957	-0.1852
29	MLR QSPR	184/47[w]	[25]	$\mu_1\mu_2^{Std}$	0.4336	0.4367	0.4249	-0.2929	-0.2934	0.4351	0.4292	-0.3564	-0.3567
				μ_{10}^{Std}	0.3974	0.3953	0.4074	0.4264	0.4262	0.3963	0.4023	0.4116	0.4115
				$\mu_5^{Ab\text{-}R2}$	0.4132	0.4107	0.4249	-0.7535	-0.7558	0.4120	0.4190	-0.5580	-0.5589
				μ_1^{Hyd}	0.4624	0.4827	0.3900	0.0417	0.0430	0.4724	0.4247	0.1389	0.1410
				μ_1^{Dip2}	0.2578	0.2659	0.2250	-0.0716	-0.0700	0.2618	0.2409	-0.1359	-0.1343
				μ_3^{Van}	0.5188	0.5228	0.5039	0.1802	0.1738	0.5208	0.5112	0.3057	0.3003
				$\mu_1\mu_4^{Dip2}$	0.3181	0.3216	0.3250	0.0678	0.0706	0.3199	0.3215	0.1469	0.1499
				$\mu_4^{Ab\text{-}logL16}$	0.4678	0.4688	0.4645	0.1943	0.2102	0.4683	0.4662	0.3015	0.3136
				$\mu_4^{Ab\text{-}\Sigma\beta 2o}$	0.4547	0.4544	0.4568	0.2879	0.2735	0.4545	0.4557	0.3618	0.3526
				μ_4^{Pols}	0.0943	0.0779	0.1485	0.0081	0.0116	0.0857	0.1183	0.0276	0.0331
30	MLR QSPR	15/4[x]	[26]	RDF020u	0.2787	0.2087	(+)	-0.3265	-0.3576	0.2412	-	-0.3016	-0.3157
				Mor28e	-0.0944	-0.0120	(+)	0.1528	0.1761	0.0336	-	-0.1201	-0.1290
				Mor07p	0.9273	0.9168	(+)	0.9328	0.9171	0.9221	-	0.9300	0.9222
31	MLR QSAR	40/79	[27]	RB	0.6409	0.6778	0.6224	0.4458	0.3902	0.6591	0.6316	0.5345	0.5001
				HBA	0.3827	0.3263	0.4149	-0.4460	-0.4385	0.3534	0.3985	-0.4131	-0.4096
				CHI	0.6731	0.7514	0.6324	0.7761	0.8096	0.7112	0.6525	0.7228	0.7382
32	MLR QSAR	46/10[y]	[28]	DTsDeP1/dGP2	-0.3841	-0.3792	-0.3991	0.0166	-0.3328	0.3816	0.3915	-0.0798	0.3575
				lnDGsDiE1/pGE	0.3294	0.4180	0.1606	-0.0635	0.4318	0.3711	0.2300	-0.1446	0.3771
				DTjDeMp/d2GP	0.6772	0.6845	0.6647	0.7842	0.5253	0.6808	0.6709	0.7287	0.5964
				lnDTjDeEp2/d2AE	0.7422	0.7258	0.8420	-0.5539	0.5472	0.7339	0.7905	-0.6412	0.6373
				DTsDeP1/dGP2	-0.0837	-0.0728	-0.1048	-0.0886	0.1212	0.0781	0.0937	0.0861	-0.1007

				LnRGsDeMp2/d2AE	-0.6239	-0.5898	-0.8428	-0.2571	-0.3358	0.6066	0.7251	0.4005	0.4577
33	MLR QSAR	28/9	[29]	X1	-0.1495	-0.1681	-0.0038	-0.2158	-0.3705	0.1585	0.0239	0.1796	0.2353
				X2	-0.3116	-0.3506	-0.1314	-0.1107	-0.1396	0.3305	0.2023	0.1857	0.2086
				X3	0.4283	0.5493	-0.5945	0.2405	0.3533	0.4851	-0.5046	0.3210	0.3890
				X4	0.3872	0.4389	0.5639	0.7113	0.6729	0.4122	0.4673	0.5248	0.5104
				X5	0.6253	0.6811	-0.0024	0.2350	0.1665	0.6526	-0.0389	0.3833	0.3227
				X6	0.0888	0.0311	0.5906	-0.5676	-0.4879	0.0525	0.2290	-0.2245	-0.2081
34	MLR QSAR	14/2[z1]	[30]	Dip	0.6328	0.6088	(+)	0.9477	0.9505	0.6207	-	0.7744	0.7755
				IP	0.1335	0.0224	(x)	0.2000	0.1622	0.0547	-	0.1634	0.1472
				Polar	-0.1662	-0.1979	(x)	-0.2489	-0.2652	0.1814	-	0.2034	0.2099
35	MLR MI-QSAR	46/16[z2]	[31]	FH20	0.4514	0.4318	0.5197	0.2439	0.2687	0.4415	0.4843	0.3318	0.3483
				Dipole	0.1283	0.1294	0.1254	-0.1335	-0.1281	0.1289	0.1268	-0.1309	-0.1282
				$\Delta\Sigma h(r)$	0.8203	0.8004	0.8914	0.9606	0.9547	0.8103	0.8551	0.8877	0.8849
36	MLR QSPR	80/80[z3]	[32]	SX_{1CH}	-0.9738	-0.9699	-0.9787	0.3574	0.3248	0.9718	0.9762	-0.5899	-0.5624
				SX_{1CC}	0.9207	0.9167	0.9263	-0.2403	-0.2415	0.9187	0.9235	-0.4704	-0.4715
				SV_{ij}	0.5176	0.5016	0.5377	0.0961	0.0841	0.5096	0.5275	0.2230	0.2086
				OEI	0.9607	0.9547	0.9685	0.1946	0.1602	0.9577	0.9646	0.4324	0.3923
				$N^{2/3}$	0.9910	0.9910	0.9912	0.8760	0.8964	0.9910	0.9911	0.9317	0.9425
37	MLR QSAR	22/8[z4]	[33]	ASMmVQt	0.5171	0.5351	0.5214	0.4832	0.5344	0.5260	0.5193	0.4999	0.5257
				lfDdOQg	-0.5599	-0.4503	-0.8974	-0.7862	-0.7553	0.5021	0.7088	0.6635	0.6503
				InMrLQg	-0.1839	-0.1877	-0.3048	0.2683	0.2576	0.1858	0.2368	-0.2221	-0.2177
				LsDMpQg	0.1422	0.0048	0.5864	-0.2766	-0.2786	0.0261	0.2888	-0.1984	-0.1991
38	MLR QSAR	23/6	[34]	ΔV	0.2472	0.2736	(x)	0.1580	0.1780	0.2601	-	0.1976	0.2098
				MR2	-0.5139	-0.6454	(x)	-0.1136	-0.1653	0.5759	-	0.2416	0.2914
				ΔE_1	-0.2482	-0.2744	(x)	-0.6421	-0.6530	0.2610	-	0.3992	0.4026
				$(\Delta E_1)^2$	0.2036	0.2278	(x)	-0.7005	-0.6765	0.2154	-	-0.3776	-0.3711
				ΔE_2	-0.4247	-0.4907	(-)	-0.2431	-0.2388	0.4566	-	0.3213	0.3185
39	MLR QSAR	20/8	[35]	MR	0.6417	0.7267	0.4586	0.3170	0.3192	0.6829	0.5425	0.4510	0.4526
				DM	0.4237	0.4886	0.2422	0.0289	0.0201	0.4550	0.3203	0.1107	0.0923
				SASA	0.7350	0.8234	0.5289	0.5832	0.5940	0.7779	0.6235	0.6547	0.6607
				Polrz	0.6390	0.7265	0.4444	-0.7296	-0.6693	0.6813	0.5329	-0.6828	-0.6540
				LogPo/w	-0.4484	-0.5103	-0.2744	-0.1003	-0.2362	0.4784	0.3508	0.2121	0.3254
				LogS	-0.1628	-0.2813	-0.0477	-0.1272	-0.2026	0.2140	0.0881	0.1439	0.1816
40	MLR QSAR	25/9[z5]	[36]	β_{xxx}	0.2353	0.1393	0.2420	0.2579	0.1168	0.1810	0.2386	0.2463	0.1658
				β_{xyy}	-0.2370	-0.0362	-0.3758	-0.1795	-0.0048	0.0927	0.2985	0.2063	0.0337
				Ω_{xyz}	0.3254	0.5188	-0.0675	0.1837	0.2373	0.4109	-0.1482	0.2445	0.2779
				Ω_{zzz}	0.0518	0.0959	-0.0931	-0.1799	-0.1191	0.0705	-0.0695	-0.0966	-0.0786
				$\Delta\alpha$	0.8645	0.8970	0.8463	0.9139	0.9570	0.8806	0.8553	0.8889	0.9096
41	MLR QSAR	18/6	[37]	x_2	-0.3397	-0.3584	(-)	-0.1222	-0.1371	0.3489	-	0.2037	0.2158
				x_9	-0.4627	-0.5122	(-)	-0.6901	-0.7406	0.4868	-	0.5651	0.5854
				x_{21}	0.0146	-0.0422	(+)	0.1052	0.1623	-0.0248	-	0.0392	0.0487

				x_{62}	0.5638	0.4745	(+)	0.7055	0.6375	0.5172	-	0.6307	0.5995
42	MLR QSAR	20/6[z6]	[38]	HOMO	-0.8265	-0.7854	(-)	-0.9448	-0.9219	0.8057	-	0.8837	0.8729
				TOE	-0.1535	-0.2709	(-)	-0.3278	-0.3875	0.2039	-	0.2243	0.2439
43	MLR QSAR	23/9[z7]	[39]	qC_{13}	0.1768	0.1326	0.3621	0.3538	0.3169	0.1531	0.2530	0.2501	0.2367
				GAP	0.2194	0.1618	0.4688	0.5264	0.5042	0.1884	0.3207	0.3399	0.3326
				SA	-0.3407	-0.2687	-0.5814	-0.2695	-0.2635	0.3025	0.4450	0.3030	0.2996
				nHAcc	-0.4166	-0.4035	-0.5050	-0.7246	-0.7589	0.4100	0.4587	0.5495	0.5623
44	MLR QSAR	28/8[z8]	[40]	CRI	0.4006	0.3977	0.3985	0.4753	0.4976	0.3992	0.3995	0.4364	0.4465
				E_{LUMO}	-0.8088	-0.7719	-0.8818	-0.8798	-0.8674	0.7902	0.8445	0.8436	0.8376
45	MLR QSPR	40/18[z9]	[41]	$^{1}\chi^{\mathrm{f}}(x,y)$	0.9935	0.9947	0.9889	0.9823	0.9838	0.9941	0.9912	0.9879	0.9887
				$^{m}EM_2$	0.9311	0.9290	0.9440	0.1871	0.1790	0.9300	0.9375	0.4174	0.4082
46	MLR ADME	57/49	[42]	V^2	-0.1967	-0.4514	0.2028	-0.5854	-0.6117	0.2980	-0.1997	0.3394	0.3469
				V	-0.2004	-0.4605	0.2557	0.6852	0.6928	0.3038	-0.2264	-0.3706	-0.3726
				PSA	-0.7767	-0.8561	-0.6886	-0.4333	-0.3819	0.8154	0.7313	0.5801	0.5446
47	MLR QSAR	40/18[z10]	[43]	x_5	0.5505	0.4746	0.7224	0.2999	0.2588	0.5111	0.6306	0.4063	0.3775
				x_{21}	0.6323	0.6100	0.7137	0.4994	0.4957	0.6211	0.6718	0.5619	0.5599
				x_{26}	0.5940	0.5891	0.6104	0.6869	0.6959	0.5916	0.6021	0.6388	0.6429
				x_{32}	0.2951	0.2641	0.3559	0.4029	0.4063	0.2792	0.3241	0.3448	0.3463
				x_{51}	0.2118	0.2983	-0.1562	0.1628	0.1949	0.2514	-0.1819	0.1857	0.2032
48	MLR QSAAR	18/5[z11]	[44]	Human Liver	0.7308	0.6675	(+)	0.3154	0.2991	0.6984	-	0.4801	0.4675
				LUMO	-0.8367	-0.7880	(-)	-0.9111	-0.8961	0.8120	-	0.8731	0.8659
				N_{O}	-0.0429	-0.1603	(x)	-0.2654	-0.3280	0.0829	-	0.1067	0.1186
49	MLR QSAR	20/6[z12]	[45]	α	0.4061	0.5739	(+)	0.7527	0.7604	0.4828	-	0.5529	0.5557
				α^2	0.3330	0.4869	(+)	-0.6443	-0.6433	0.4027	-	-0.4632	-0.4628
				F^{N}_{C*}	-0.4939	-0.2166	(-)	-0.1331	-0.0857	0.3271	-	0.2564	0.2057
				Q_{N**}	0.3333	0.3059	(+)	0.0239	0.0269	0.3193	-	0.0892	0.0947
50	MLR QSPR	53/30	[46]	$^{2}\Omega_{p}^{\mathrm{C}}(q)$	-0.8798	-0.7368	-0.8671	-0.8537	-0.7527	0.8052	0.8734	0.8667	0.8138
				$^{6}\varepsilon_{\mathrm{Ch}}(\rho)$	-0.6902	-0.6520	-0.7495	-0.5207	-0.6584	0.6708	0.7192	0.5995	0.6741

[*]PLS and MLR models from diverse studies: QSAR – quantitative structure-activity relationship; QGAR – quantitative genome-activity relationship; QGSAR – quantitative genome/structure-activity relationship; QSAAR – quantitative structure/activity-activity relationship; ADME – absorption, distribution, metabolism, excretion; LFER – linear free energy relationship; MI-QSAR – membrane-interaction QSAR; QSSR – quantitative structure-structure relationship; and QSPR – quantitative structure-property relationship.

[a]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the complete dataset.

[b]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the training set after data split.

[c]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the external validation set after data split. The correlation coefficient was not calculated for external sets with less than seven samples, but a qualitative parameter for correlation was determined from scatterplots: (+) – positive correlation, (-) – negative correlation, and (x) – direction of correlation could not be determined. This qualitative parameter was not used in classification of variables according to criterion II (Table S2).

[d]Normalized regression vector for the complete dataset.

[e]Normalized regression vector for the training set after data split.

[f]Function $F_1 = \mathrm{sign}(r_c r_t)\sqrt{|r_c r_t|}$.

[g]Function $F_2 = \mathrm{sign}(r_\mathrm{c} r_\mathrm{e})\sqrt{|r_\mathrm{c} r_\mathrm{e}|}$.

[h]Function $F_3 = \mathrm{sign}(r_\mathrm{c} \beta_\mathrm{c})\sqrt{|r_\mathrm{c} \beta_\mathrm{c}|}$.

[i]Function $F_4 = \mathrm{sign}(r_\mathrm{c} \beta_\mathrm{t})\sqrt{r_\mathrm{c} \beta_\mathrm{t}}$

[j]A new data split had to be applied in this work because the original reference did not contain sufficient or any information about the external validation samples. Based on HCA clustering at similarity index 0.45, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 6, 9, 23, 29, 31 and 34.

[k]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.65, the following samples were selected for the external validation set along the whole range of values of **y**: 3, 6, 9, 14-17, 19, 23, 25, 27, 30, 33, 36, 38, 40, 42, 46, 55, 61 and 62.

[l]The dependent variable **y** was molar solubility S_m.

[m]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.65, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 4, 9 and 16.

[n]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.80, the following samples were selected for the external validation set along the whole range of values of **y**: 1, 5, 6, 8, 11, 12, 16, 17, 19-22, 25, 28, 30, 32, 33, 36, 38-42, 49, 51, 52, 54, 55, 58, 60, 62, 63, 65, 67, 70, 75, 76, 78, 81, 83, 84, 87, 88, 89, 91, 92, 95, 97, 102, 108, 110, 115, 117, 119 and 122.

[o]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.70, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 7, 9, 13, 15, 18, 25, 29, 30, 32, 35-38, 42, 44, 50, 52, 54, 55, 58, 65, 69-71, 75, 76, 79, 81 and 84.

[p]The dependent variable was anabolic activity log(1/LA).

[q]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.70, the following samples were selected for the external validation set along the whole range of values of **y**: 3-5, 8, 9, 15, 16, 18, 22-24, 30, 31, 33, 35, 36, 38, 40, 44, 49, 53-55, 57, 60, 63, 64, 67, 69, 70, 75, 80 and 82.

[r]Data split was based on D-optimal design from the original work.

[s]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.60, the following samples were selected for the external validation set along the whole range of values of **y**: 1, 6, 10 and 17.

[t]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.55, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 10, 13, 18, 21 and 28.

[u]The dependent variable **y** was –logLC_{50}.

[v]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.30, the following samples were selected for the external validation set along the whole range of values of **y**: 3, 9, 11, 12, 19, 27, 33, 34, 38, 40, 42, 47 and 48.

[w]One sample was missing from the complete dataset of 133 samples, *i.e.*, from the external validation set.

[x]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.50, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 9, 12 and 18.

[y]Two outliers had to be removed from the complete dataset (11 and 12) in this work. After that, the split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.40, the following samples were selected for the external validation set along the whole range of values of **y**: 8, 20, 22, 23, 34, 36, 37, 45, 53 and 58.

[z1]The investigated dataset was cluster III from the original reference. The split for this dataset was applied in this work, since no split was made in the original reference. Based on HCA analysis showing two completely distinct clusters, the following samples were selected for the external validation set along the whole range of values of **y**: 3 and 16.

[z2]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.40, the following samples were selected for the external validation set along the whole range of values of **y**: 3, 4, 8, 15, 17, 18, 22, 26, 33, 37, 38, 45, 49, 51, 54 and 61.

[z3]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.50 and distribution of **y**, the following samples were selected for the external validation set along the whole range of values of **y**:2, 5, 7-9, 11, 12 and all odd samples starting with 15 and ending with 159.

[z4]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.30, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 7, 9, 11,14, 16, 25 and 28.
[z5]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.40, the following samples were selected for the external validation set along the whole range of values of **y**: 23, 4, 10, 11, 17, 27, 32 and 33. This numeration of samples is according to their position in the dataset which was used to build the model.
[z6]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.40, the following samples were selected for the external validation set along the whole range of values of **y**: 3, 4, 7, 17, 23 and 24.
[z7]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.40, the following samples were selected for the external validation set along the whole range of values of **y**: 10-12, 15, 20, 21, 28, 30 and 36.
[z8]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.60, the following samples were selected for the external validation set along the whole range of values of **y**: 1, 3, 4, 15, 19, 23, 25, 33 and 36.
[z9]The split applied was defined in this work, since no split was made in the original reference. Based on HCA clustering at similarity index 0.60, the following samples were selected for the external validation set along the whole range of values of **y**: 6, 11, 16, 18-20, 24, 25, 27, 28, 30, 31, 34, 35, 42, 47, 49 and 55.
[z10]A new split applied was defined in this work for 58 peptides, since no sufficient information about data splits was provided in the original reference. Based on HCA clustering at similarity index 0.60, the following samples were selected for the external validation set along the whole range of values of **y**: 2, 6, 8, 10, 17, 20, 21, 31, 33, 34, 35, 37, 40, 43, 45, 47, 51 and 54.
[z11]A new split applied was defined in this work, since no sufficient information about data splits was provided in the original reference. Based on HCA clustering at similarity index 0.60, the following samples were selected for the external validation set along the whole range of values of **y**: 8, 31, 36, 38 and 39.
[z12]The split applied was defined in this work, since no sufficient information about data splits was provided in the original reference. Based on HCA clustering at similarity index 0.85, the following samples were selected for the external validation set along the whole range of values of **y**: 2c, 2g, 2j, 2q, 2r and 2t.

References

(1) Kiralj, R.; Ferreira, M. M. C. Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Application. *J. Braz. Chem. Soc.* **2009**, *20* 770-787.
(2) Rasulev, B. F.; Toropov, A. A.; Hamme, A. T. II.; Leszcynski, J. Multiple Linear Regression Analysis and Optimal Descriptors: Predicting the Cholesteryl Ester Transfer Protein Inhibition Activity. *QSAR Comb. Sci.* **2008**, *27*, 595-606.
(3) Deeb, O.; Youssef, K. M.; Hemmateenejad, B. QSAR of Novel Hydroxyphenyureas as Antioxidant Agents. *QSAR Comb. Sci.* **2008**, *27*, 417-424.
(4) Kiralj, R.; Ferreira, M. M. C. Extensive Chemometric Investigations of the Multidrug Resistance in Strains of the Phytopathogenic Fungus *Penicillium digitatum*. *QSAR Comb. Sci.* **2008**, *27*, 289-301.
(5) Takkis, K.; Sild, S.. QSAR Modeling of HIV-1 Protease Inhibition on Six- and Seven-membered Cyclic Ureas. *QSAR Comb. Sci.* **2009**, *28*, 52-58.
(6) Sprunger, L. M.; Gibbs, J.; Acree, W. E. Jr.; Abraham, M. H. Linear Free Energy Relationship Correlation of The Distribution of Solutes Between Water And Cetytrimethylammonium Bromide (CTAB) Micelles. *QSAR Comb. Sci.* **2009**, *28*, 72-88.
(7) Teófilo, R. F.; Kiralj, R.; Ceraglioli, H. J.; Peterlevitz, A. C.; Baranauskas, V.; Kubota, L. T.; Ferreira, M. M. C. QSPR study of passivation by phenolic compounds at platinum and boron-doped diamond electrodes. *J. Electrochem. Soc.* **2008**, *155*, D640-D650.
(8) Chang, J.; Lei, B.-L.; Li, J.-Z.; Li, S.-Y.; Shen, Y.-L.; Yao, X.-J.. Accurate and Validated Quantitative Structure – Activity Relationship Model of Caspase-mediated Apoptosis-inducing Activity of Phenolic Compounds Using Density Functional Theory Calculation and Genetic Algorithm – Multiple Linear Regression. *QSAR Comb. Sci.* **2008**, *27*, 1318-1325.

(9) Li, Z. G.; Chen, K.-X.; Xie, H.-Y.; Gao, J.-R. Quantitative Structure-Activity Relationship Analysis of Some Thiourea Derivatives with Activities Against HIV-1 (IIIB). *QSAR Comb. Sci.* **2009**, *28*, 89-97.
(10) Wu, D.; Liu, X.-H.; Wang, L.; Wang, L.; Xu, M.-Z.; Sun, T.; Yang, Z.-F.; Zhou, J.-L. QSARs on the Depuration Rate Constants of Polycyclic Aromatic Hydrocarbons in *Elliptio complanata*. *QSAR Comb. Sci.* **2009**, *28*, 537-541.
(11) Lu, G.-N.; Dang, Z.; Tao, X.-Q.; Yang, C.; Yi, X.-Y.. Estimation of Water Solubility of Polycyclic Aromatic Hydrocarbons Using Quantum Chemical Descriptors and Partial Least Squares. *QSAR Comb. Sci.* **2008**, *27*, 618-626.
(12) Liao, S. Y.; Qian, L.; Lu, H. L.; Shen, Y.; Zheng, K. C. A Combined 2D- and 3D-QSAR Study on Analogues of ARC-111 with Antitumor Activity. *QSAR Comb. Sci.* **2008**, *27*, 740-749.
(13) Filipic, S.; Nikolic, K.; Krizman, M.; Agbaba, D. The Quantitative Structure-Retention Relationship (QSRR) Analysis of Some Centrally Acting Antihypertensives and Diuretics. *QSAR Comb. Sci.* **2008**, *27*, 1036-1044.
(14) Camargo, A. B.; Marchevsky, E.; Luco, J. M.. QSAR Study for the Soybean 15-Lipoxygenase Inhibitory Activity of Organosulfur Compounds Derived from the Essential Oil of Garlic. *J. Agric. Food Chem.* **2007**, *55*, 3096-3103.
(15) Agrawal, V. K.; Chaturvedi, S.; Abraham, M. H.; Khadikar, P. V.. QSAR Study on Tadpole Narcosis. *Bioorg. Med. Chem.* **2003**, *11*, 4523-4533.
(16) Liu, H. X.; Gramatica, P.. QSAR study of selective ligands for the thyroid hormone receptor β. *Bioorg. Med. Chem.* **2007**, *15*, 5251-5261.
(17) Gayen, S.; Debnath, B.; Samanta, S.; Jha, T.. QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters. *Bioorg. Med. Chem.* **2004**, *12*, 1493-1503.
(18) Alvarez-Ginarte, Y. M.; Crespo-Otero, R.; Marrero-Ponce, Y.; Noheda-Marin, P.; de la Vega, J. M. G.; Montero-Cabrera, L. A.; García, J. A. R.; Caldera-Luzardo, J. A.; Alvarado, Y. J. Chemometric and chemoinformatic analyses of anabolic and androgenic activities of testosterone and dihydrotestosterone analogues. *Bioorg. Med. Chem.* **2008**, *16*, 6448-6459.
(19) Katritzky, A. R.; Slavov, S. H.; Dobchev, D. A.; Karelson, M. QSAR modeling of the antifungal activity against *Candida albicans* for a diverse set of organic compounds. *Bioorg. Med. Chem.* **2008**, *16*, 7055-7069.
(20) Gramatica, P.; Pilutti, P.; Papa, E.. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-Test sets and Consensus Modeling. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794-1802.
(21) Juranić, I. O.; Drakulić, B. J.; Petrović, S. D.; Mijin, D. Ž.; Stanković, M. V. A QSAR study of acute toxicity of *N*-substituted fluoroacetamides to rats. *Chemosphere* **2006**, *62*, 641-649.
(22) Wang, Z.-Y.; Zhai, Z.-C.; Wang, L.-S. Quantitative Structure-activity Relationship of Toxicity of Alkyl(1-phenylsulfonyl) Cycloalkane-carboxylates Using MLSER Model and Ab initio. *QSAR Comb. Sci.* **2005**, *24*, 211-217.
(23) Zhang, L.; Zhou, P.-J.; Yang, F.; Wang, Z.-D. Computer-based QSARs for predicting mixture toxicity of benzene and its derivatives. *Chemosphere* **2007**, *67*, 396-401.
(24) Lu, W.-J.; Chen, Y.-L.; Liu, M.-C.; Chen, X.-G.; Hu, Z. QSPR prediction of *n*-octanol/water partition coefficient for polychlorinated biphenyls. *Chemosphere* **2007**, *69*, 469-478.
(25) Pérez-Garrido, A.; Helguera, A. M.; Cordeiro, M. N. D. S.; Escudero, A. G. QSPR Modelling With the Topological Substructural Molecular Design Approach: β-Cyclodextrin Complexation. *J. Pharm. Sci.* **2009**, *98*, 4557-4576.
(26) Turabekova, M. A.; Rasulev, B. F. A QSAR Toxicity Study of a Series of Alkaloids with the Lycoctonine Skeleton. *Molecules* **2004**, *9*, 1194-1207.
(27) Maccari, L.; Magnani, M.; Strappaghetti, G.; Corelli, F.; Botta, M.; Manetti, F. A Genetic-Function-Approximation-Based QSAR Model for the Affinity of Arylpiperazines toward α_1 Adrenoceptors. *J. Chem. Inf. Model.* **2006**, *46*, 1466-1478.
(28) Ursu, O.; Don, M.; Katona, G.; Jäntschi, L.; Diudea, M.. QSAR Study On Dipeptide Ace Inhibitors. *Carpathian J. Math.* **2004**, *20*, 275-280.
(29) Dessalew, N. Investigation of the structural requirement for inhibiting HIV integrase: QSAR study. *Acta Pharm.* **2009**, *59*, 31-43.

(30) Pillai, A. D.; Rani, S.; Rathod, P. D.; Xavier, F. P.; Vasu, K. K.; Padh, H.; Sudarsanam, V. QSAR studies on some thiophene analogs as anti-inflammatory agents: enhancement of activity by electronic parameters and its utilization for chemical lead optimization. *Bioorg. Med. Chem.* **2005**, *13*, 1275-1283.
(31) Zheng, T.; Hopfinger, A. J.; Esposito, E. X.; Liu, J.-Z.; Tseng, Y.-F. J. Membrane-Interaction Quantitative Structure-Activity Relationship (MI-QSAR) Analyses of Skin Penetration Enhancers. *J. Chem. Inf. Model.* **2008**, *48*, 1238-1256.
(32) Cao, C.-Z.; Jiang, L.-H.; Yuan, H. Eigenvalues of the Bond Adjacency Matrix Extended to Application in Physicochemical Properties of Alkanes. *Internet Electron. J. Mol. Des.* **2003**, *2*, 621-641.
(33) Jäntschi, L.; Popescu, V.; Bolboacă, S. D. Toxicity caused by para-substituted phenols on *Tetrahymena pyriformis*: The structure-activity relationships. *Electron. J. Biotechnol.* **2008**, *11*, issue-3-fulltext-9.
(34) Fan, F.; Cheng, J.-G.; Li, Z.; Xu, X.-Y.; Qian, X.-H. Novel Dimer Based Descriptors with Solvational Computation for QSAR Study of Oxadiazoylbenzoyl-ureas as Novel Insect-growth Regulators. *J. Comput. Chem.* **2010**, *31*, 586-591.
(35) Kumari, K. M.; Kanth, S. S.; Vijjulatha, M. Docking and QSAR Studies for Inhibitors of Thymidylate Synthase. *Internet Electron. J. Mol. Des.* **2008**, *7*, 131-141.
(36) Gu, C.-G.; Jiang, X.; Ju, X.-H.; Yu, G.-F.; Bian, Y.-R. QSARs for the toxicity of polychlorinated dibenzofurans through DFT-calculated descriptors of polarizabilities, hyperpolarizabilities and hyper-order electric moments. *Chemosphere* **2007**, *67*, 1325-1334.
(37) Liu, S.-S.; Liu, H.-L.; Shi, Y.-Y.; Wang, L.-S. QSAR of Cyclooxygenase–2 (COX–2) Inhibition by 2,3-Diarylcyclopentenones Based on MEDV–13. *Internet Electron. J. Mol. Des.* **2002**, *1*, 310-318.
(38) Jaiswal, D.; Karthikeyan, C.; Shrivastava, S. K.; Trivedi, P. QSAR Modeling of Sulfonamide Inhibitors of Histone Deacetylase. *Internet Electron. J. Mol. Des.* **2006**, *5*, 345-354.
(39) Panda, P.; Samanta, S.; Alam, Sk. M.; Basu, S.; Jha, T. QSAR for Analogs of 1,5–*N*,*N'*–Disubstituted–2 (substitutedbenzenesulphonyl) Glutamamides as Antitumor Agents. *Internet Electron. J. Mol. Des.* **2007**, *6*, 280-301.
(40) Saçan, M. T.; Özkul, M.; Erdem, S. S. QSPR analysis of the toxicity of aromatic compounds to the algae (*Scenedesmus obliquus*). *Chemosphere* **2007**, *68*, 695-702.
(41) Janežič, D.; Lučić, B.; Nikolić, S.; Miličević, A.; Trinajstić, N. Boiling Points of Alcohols – A Comparative QSPR Study. *Internet Electron. J. Mol. Des.* **2006**, *5*, 192-200.
(42) X.-C. Fu, Z.-F. Song, W.-Q. Liang. A Predictive Model for Blood-Brain Barrier Penetration. Internet Electron. J. Mol. Des., 4 (2005) 737-750.
(43) Liu, S.-S.; Yin, C.-S.; Wang, L.-S. Combined MEDV-GA-MLR Method for QSAR of Three Panels of Steroids, Dipeptides, and COX-2 Inhibitors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749-756.
(44) Lessigiarska, I.; Worth, A. P.; Netzeva, T. I.; Dearden, J. C.; Cronin, M. T. D.. Quantitative structure-activity-activity and quantitative structure-activity investigations of human and rodent toxicity. *Chemosphere* **2006**, *65*, 1878-1887.
(45) Wan, J.; Zhang, L.; Yang, G.-F.; Zhan, C.-G. Quantitative Structure-Activity Relationship for Cyclic Imide Derivatives of Protoporphyrinogen Oxidase Inhibitors: A Study of Quantum Chemical Descriptors from Density Functional Theory. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2099-2105.
(46) Estrada, E.; Delgado, E. J.; Alderete, J. B.; Jaña, G. A. Quantum-Connectivity Descriptors in Modeling Solubility of Environmentally Important Organic Compounds. *J. Comput. Chem.* **2004**, *25*, 1787-1796.

Table S2. Descriptors characterization according to criteria I, II, III and IV with respect to the sign change problem.

Data[a]	Model	Descriptor[a]	r_c[b]	r_t[c]	r_e[d]	β_c[e]	β_t[f]	Type[g]	Characterization and visual diagnostics[h]
1	MLR QSAR	LogK_{ow}	0.8807	0.8897	0.8710	0.9340	0.9491	Real descriptor	Good
		pK_a	**0.0249**	**0.0819**	**-0.0469**	**0.1185**	**0.1316**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		E_{LUMO}	-0.0998	-0.1061	-0.0928	-0.3331	-0.2855	Hidden noise	Not acceptable; serious problems with distinct groups and dispersion
		E_{HOMO}	**-0.0058**	**0.0210**	**-0.0387**	**0.0328**	**0.0080**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		N_{hdon}	**-0.4100**	**-0.4121**	**-0.4077**	**0.0391**	**0.0182**	Anti descriptor	Acceptable; moderate distribution problems
2	PLS QGSAR	CYP51-g	-0.7223	-0.7206	-0.7285	-0.3380	-0.3355	Real descriptor	Acceptable; moderate distribution problems
		CYP51-e	-0.7263	-0.7264	-0.7288	-0.3719	-0.3901	Real descriptor	Acceptable; moderate distribution problems
		PMR1-t	-0.5023	-0.4612	-0.5867	-0.2800	-0.2720	Real descriptor	Acceptable; moderate distribution problems
		CYP51-e*Npi	**-0.6238**	**-0.6306**	**-0.6094**	**0.4894**	**0.5012**	Anti descriptor	Acceptable after linearization
		PCR*Npi	-0.5558	-0.5678	-0.5288	-0.3883	-0.4131	Real descriptor	Acceptable after linearization
		PMR1-e*Lpi	-0.6775	-0.6899	-0.6558	-0.1113	-0.1419	Real descriptor	Acceptable after linearization
		CYP51-e*Lpi	**-0.6337**	**-0.6612**	**-0.5792**	**0.4173**	**0.3981**	Anti descriptor	Acceptable after linearization
		PCR*Lpi	-0.5635	-0.5898	-0.5101	-0.3037	-0.2465	Real descriptor	Acceptable after linearization
3	PLS QSPR	E_e	-0.8561	-0.8445	-0.9166	-0.2401	-0.2481	Real descriptor	Acceptable; moderate distribution problems
		E_{CC}	-0.8920	-0.8842	-0.9589	-0.1475	-0.1267	Real descriptor	Acceptable; moderate distribution problems
		Q_{Omul}	0.9282	0.9176	0.9753	0.2833	0.2429	Real descriptor	Acceptable; moderate distribution problems
		Δ_{HL}	-0.8267	-0.8435	-0.7463	-0.5277	-0.4888	Real descriptor	Accept; pronounced dispersion
		σ_b	0.8619	0.8580	0.8984	0.0863	0.0848	Real descriptor	Acceptable; moderate distribution problems
		σ_r	-0.8905	-0.8989	-0.8526	-0.6669	-0.7204	Real descriptor	Accept; pronounced dispersion
		D_{CC}	0.9069	0.8976	0.9769	0.2044	0.1733	Real descriptor	Acceptable; moderate distribution problems
		Q_{C2mul}	0.8915	0.8863	0.9172	0.2607	0.2613	Real descriptor	Accept; pronounced dispersion
4	MLR QSAR	ClogP	0.6323	0.6977	(-)	0.7856	0.8072	Real descriptor	Acceptable; moderate distribution problems
		MgVol	-0.1070	-0.0412	(-)	-0.5450	-0.5068	Hidden noise	Not acceptable; serious problems with distinct groups
		$B1_{X.2}$	-0.4509	-0.4485	(x)	-0.2929	-0.3027	Real descriptor	Not acceptable; serious distribution problems
5	PLS QSPR	E_e	-0.8159	-0.8135	(-)	-0.2900	-0.3288	Real descriptor	Acceptable; moderate distribution problems
		E_{CC}	-0.8587	-0.8565	(-)	-0.0475	-0.0617	Real descriptor	Acceptable; moderate distribution problems
		Q_{Omul}	0.9204	0.9179	(-)	0.3191	0.3322	Real descriptor	Acceptable; moderate distribution problems
		Δ_{HL}	-0.8266	-0.8257	(-)	-0.4814	-0.5493	Real descriptor	Acceptable; moderate distribution problems
		σ_b	0.8459	0.8411	(+)	0.0378	0.0558	Real descriptor	Acceptable; moderate distribution problems
		σ_r	-0.8550	-0.8444	(-)	-0.7026	-0.6179	Real descriptor	Acceptable; moderate distribution problems
		D_{CC}	0.8718	0.8680	(+)	0.0933	0.1021	Real descriptor	Acceptable; moderate distribution problems
		Q_{C2mul}	0.8943	0.8885	(+)	0.2760	0.2840	Real descriptor	Acceptable; moderate distribution problems
6	MLR QSAR	X3A	0.6456	0.7673	0.4883	0.6055	0.6385	Real descriptor	Accept; pronounced dispersion
		BEHv2	0.3023	0.0175	0.5984	0.6263	0.5597	Hidden noise	Acceptable after linearization
		R7v	-0.5008	-0.6369	-0.3267	-0.4910	-0.5282	Real descriptor	Acceptable; moderate distribution problems
7	MLR QSAR	HE	-0.4734	-0.4567	-0.5675	-0.0686	-0.0689	Real descriptor	Accept; pronounced dispersion
		DM_z	**0.0103**	**0.0990**	**-0.3219**	**-0.0777**	**-0.0724**	Unstable noise	Not acceptable; serious problems with distinct groups

		DM_t	**0.0243**	**-0.0176**	**0.1884**	**0.0566**	**0.0549**	Unstable noise	Not acceptable; serious problems with distinct groups
		Q_{mean}	0.6287	0.6842	0.3280	0.0742	0.0764	Real descriptor	Acceptable; moderate distribution problems
		SSC	-0.3238	-0.3587	-0.0901	-0.3026	-0.3189	Quasi descriptor	Not acceptable; serious problems with distinct groups and dispersion
		X5	**-0.2754**	**-0.3206**	**0.0059**	**0.7521**	**0.7464**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		S0K	-0.3317	-0.3713	-0.0784	-0.5607	-0.5581	Quasi descriptor	Not acceptable; serious problems with distinct groups and dispersion
		PW2	-0.1359	-0.0980	-0.4165	-0.0948	-0.1044	Real noise	Not acceptable; problematic distinct groups and distribution
8	PLS QGAR	PMR1-g	-0.4906	-0.4475	-0.5797	-0.4129	-0.3823	Real descriptor	Acceptable; moderate distribution problems
		PMR1-e	-0.6942	-0.7097	-0.6629	-0.2424	-0.2973	Real descriptor	Acceptable; moderate distribution problems
		CYP51-g	-0.7223	-0.7206	-0.7285	-0.4269	-0.4294	Real descriptor	Acceptable; moderate distribution problems
		CYP51-e	-0.7263	-0.7264	-0.7288	-0.4291	-0.4331	Real descriptor	Acceptable; moderate distribution problems
		PCR	-0.7221	-0.7203	-0.7285	-0.4270	-0.4295	Real descriptor	Acceptable; moderate distribution problems
		PMR1-t	-0.5023	-0.4612	-0.5867	-0.4713	-0.4571	Real descriptor	Acceptable; moderate distribution problems
9	MLR QSAR	1/SIC2	-0.6733	-0.6932	-0.5387	-0.7327	-0.7606	Real descriptor	Accept; pronounced dispersion
		1/DPSA3	**-0.4665**	**-0.5081**	**-0.2630**	**0.2393**	**0.2658**	Anti descriptor	Accept; pronounced dispersion
		1/HPCSA	-0.6506	-0.5610	-0.8036	-0.4806	-0.4325	Real descriptor	Accept after removing the outliers; pronounced dispersion
		DPSA1	**0.2978**	**0.2572**	**0.2893**	**-0.4182**	**-0.4046**	Unstable noise	Not acceptable; serious distribution problems, especially dispersion
10	MLR LFER	E	0.7525	0.7503	0.7607	0.3002	0.2922	Real descriptor	Acceptable; moderate distribution problems
		S	**0.5294**	**0.5236**	**0.5426**	**-0.1558**	**-0.1505**	Anti descriptor	Acceptable after linearization
		A	**0.0286**	**0.0721**	**-0.0662**	**0.1541**	**0.1468**	Unstable noise	Not acceptable; serious distribution problems, especially dispersion
		B	-0.0550	-0.0725	-0.0177	-0.4473	-0.4545	Hidden noise	Acceptable after linearization
		V	0.8549	0.8610	0.8441	0.8135	0.8148	Real descriptor	Good
11	PLS QSPR	HBD/N	-0.7674	-0.8693	(-)	-0.4989	-0.5283	Real descriptor	Accept; pronounced dispersion
		Mor06u	0.5583	0.6738	(+)	0.3630	0.4095	Real descriptor	Acceptable; moderate distribution problems
		Qcnpa	0.7772	0.7345	(+)	0.5053	0.4463	Real descriptor	Acceptable; moderate distribution problems
		Ar	0.5797	0.6249	(+)	0.3769	0.3798	Real descriptor	Acceptable; moderate distribution problems
		QNUnpa	-0.7248	-0.7539	(-)	-0.4712	-0.4581	Real descriptor	Acceptable; moderate distribution problems
12	MLR QSAR	ACIC1	**0.3642**	**0.3123**	**0.4644**	**-0.3495**	**-0.3460**	Anti descriptor	Not acceptable; serious problems with distinct groups; non-linearity
		MIA	-0.8135	-0.7809	-0.9140	-0.6981	-0.6826	Real descriptor	Acceptable after linearization; modest distribution problems
		FNSA3	0.3290	0.3212	0.2343	0.1644	0.1669	Quasi descriptor	Not acceptable; serious problems with distinct groups and dispersion
		RPCS	-0.8188	-0.8144	-0.8308	-0.3211	-0.3150	Real descriptor	Acceptable; moderate distribution problems
		APMIA	**-0.7472**	**-0.7224**	**-0.8314**	**0.5103**	**0.5360**	Anti descriptor	Acceptable; moderate distribution problems, including dispersion
13	MLR QSAR	S_aaCH	-0.2422	-0.3301	-0.0580	-0.3539	-0.2555	Hidden noise	Not acceptable; serious problems with distinct groups
		Shad_XYfrac	**-0.1335**	**-0.2490**	**0.1095**	**-0.1399**	**-0.2594**	Unstable noise	Not acceptable; pronounced dispersion
		Hbond_Acc	0.5257	0.6809	0.1066	0.8536	0.8222	Quasi descriptor	Acceptable; moderate distribution problems
		LUMO	0.1328	0.1684	0.0895	0.3557	0.4375	Hidden noise	Not acceptable; serious problems with distinct groups
14	PLS QSAR	DE	0.9525	0.9587	0.9483	0.6348	0.6259	Real descriptor	Good
		M_w	-0.9502	-0.9549	-0.9687	-0.6333	-0.6234	Real descriptor	Good
		E_{HOMO}	-0.6640	-0.7180	-0.6186	-0.4426	-0.4687	Real descriptor	Accept; pronounced dispersion
15	PLS[1] QSPR	TE	0.9443	0.9399	0.9488	0.6961	0.6010	Real descriptor	Good
		R_e	-0.9308	-0.9213	-0.9410	-0.6851	-0.5107	Real descriptor	Good
		E_{LUMO}-E_{HOMO}	0.7589	0.9049	0.5868	0.0750	0.1541	Real descriptor	Acceptable; moderate distribution problems

		E_{LUMO}	0.7628	0.8785	0.6387	0.0643	0.0780	Real descriptor	Acceptable; moderate distribution problems
		$(E_{LUMO}-E_{HOMO})^2$	0.7486	0.8709	0.6016	0.0520	0.0346	Real descriptor	Acceptable; moderate distribution problems
		E_{HOMO}	-0.6653	-0.8321	-0.4818	-0.0825	-0.2594	Real descriptor	Acceptable after linearization; pronounced dispersion
		$E_{LUMO}+E_{HOMO}$	**0.5252**	**0.5837**	**0.4876**	**0.0172**	**-0.1274**	Anti descriptor	Acceptable after linearization and removing the outliers
		L_{CC}	**0.4004**	**0.4973**	**0.2717**	**-0.0106**	**-0.1680**	Anti descriptor	Acceptable after removing distinct groups; pronounced dispersion
		Q_H^+	0.3920	0.4552	0.3621	0.1625	0.4851	Real descriptor	Not acceptable; serious problems with distinct groups and dispersion
16	MLR	ClogP	**-0.5564**	**-0.4562**	**(-)**	**0.4451**	**0.4879**	Anti descriptor	Accept after removing the outliers
	QSAR	CMR	-0.8559	-0.8078	(-)	-0.8916	-0.8682	Real descriptor	Acceptable; moderate distribution problems
		Q_{C28}	-0.2335	-0.2265	(-)	-0.0837	-0.0902	Real descriptor	Accept after removing the outliers
17	MLR	SAS	0.6073	0.6290	(+)	0.5545	0.3659	Real descriptor	Acceptable; moderate distribution problems
	QSRR	HOMO	0.8099	0.8565	(+)	0.7014	0.5446	Real descriptor	Acceptable; moderate distribution problems
		Charge	-0.8951	-0.9559	(-)	-0.4478	-0.7547	Real descriptor	Acceptable after linearization
18	PLS	SASA	0.9369	0.9336	0.9524	0.9900	0.7094	Real descriptor	Good
	QSAR	ADDD	**0.8844**	**0.8730**	**0.9014**	**-0.0914**	**0.6634**	Anti descriptor	Acceptable; moderate distribution problems
		L/Bw	0.2722	0.3130	0.1402	0.1079	0.2378	Real noise	Not acceptable; serious problems with distinct groups and dispersion
19	MLR	R_2	0.2792	0.2666	0.2910	0.0634	0.0452	Real noise	Not acceptable; serious problems with distinct groups and dispersion
	QSAR	$\Sigma\alpha_2^H$	**-0.0069**	**0.0736**	**-0.1214**	**-0.0055**	**0.0059**	Unstable noise	Not acceptable; pronounced dispersion
		$\Sigma\beta_2^O$	-0.1016	-0.0792	-0.1567	-0.1856	-0.1425	Real noise	Not acceptable; problematic distinct groups and distribution
		V_x	0.6653	0.6233	0.7311	0.1840	0.1002	Real descriptor	Accept after removing the outliers
		W	0.5055	0.4731	0.5685	0.6378	0.6521	Real descriptor	Accept after removing the outliers
		$^1\chi$	0.5826	0.5529	0.6228	0.1704	0.1693	Real descriptor	Accept after removing the outliers; pronounced dispersion
		Log(RB)	**0.4935**	**0.4537**	**0.5757**	**-0.7013**	**-0.7168**	Anti descriptor	Accept after removing the outliers
20	MLR	GATS1e	-0.3723	-0.3901	-0.3437	-0.1061	-0.1163	Real descriptor	Accept; pronounced dispersion
	QSAR	EEig08x	**0.4664**	**0.4056**	**0.7166**	**-0.6313**	**-0.6198**	Anti descriptor	Not acceptable; serious problems with distinct groups and dispersion
		EEig07d	0.6113	0.5731	0.7692	0.7311	0.7375	Real descriptor	Acceptable; moderate distribution problems
		GGI6	0.6110	0.5547	0.8033	0.1448	0.1402	Real descriptor	Acceptable; moderate distribution problems
		R6v+	**0.0635**	**0.1872**	**-0.4952**	**0.1300**	**0.1501**	Unstable noise	Not acceptable; problematic distinct groups and distribution
		H-051	-0.4290	-0.3757	-0.6217	-0.1336	-0.1272	Real descriptor	Acceptable; moderate distribution problems
21	MLR	S_{av}	0.6876	0.6594	0.7426	0.4422	0.4762	Real descriptor	Acceptable; moderate distribution problems
	QSAR	π_{R1}	0.4626	0.4209	0.5315	0.2111	0.2700	Real descriptor	Acceptable; moderate distribution problems
		I_1	0.5392	0.5543	0.5111	0.6191	0.6283	Real descriptor	Not acceptable; only two distinct values of indicator variable
		I_2	0.4967	0.4655	0.5550	0.4044	0.3811	Real descriptor	Not acceptable; only two distinct values of indicator variable
		I_{OH}	-0.5429	-0.5050	-0.6138	-0.4616	-0.4005	Real descriptor	Not acceptable; only two distinct values of indicator variable
22	MLR	log*P*	0.3231	0.3588	0.8180	0.7460	0.6141	Real descriptor	Acceptable after removing distinct groups; distribution problems
	QSAR	*n*	**-0.2772**	**-0.6384**	**0.1333**	**-0.6660**	**-0.7893**	Unstable noise	Not acceptable; problematic outliers and large dispersion
23	MLR	R.No.Cat	0.4987	0.5093	0.4595	0.5882	0.6225	Real descriptor	Accept; pronounced dispersion
	QSAR	HBdonCSA	-0.4713	-0.5451	-0.3465	-0.3871	-0.4042	Real descriptor	Acceptable after linearization; pronounced dispersion
		Av.v.Hat	**-0.0588**	**-0.0367**	**-0.0643**	**0.4376**	**0.4432**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		RNCh	**-0.0277**	**0.0993**	**-0.2110**	**-0.4791**	**-0.4134**	Unstable noise	Acceptable after linearization
		$(\log P)^2$	**0.0115**	**0.0556**	**-0.0906**	**-0.1688**	**-0.1934**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		Av.e.r.Cat	**-0.1439**	**-0.0556**	**-0.2759**	**0.2336**	**0.2108**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion

24	MLR	HOMO	-0.6987	-0.6986	-0.7153	-0.7962	-0.7753	Real descriptor	Accept; pronounced dispersion
	QSAR	nX	0.6160	0.6399	0.5092	0.3904	0.4183	Real descriptor	Acceptable after linearization; pronounced dispersion
		CIC0	-0.5786	-0.6045	-0.4900	-0.3285	-0.3547	Real descriptor	Acceptable after linearization; pronounced dispersion
		nCaH	**-0.1122**	**-0.1598**	**-0.0693**	**0.3252**	**0.3133**	Unstable noise	Not acceptable; serious distribution problems, especially dispersion
25	MLR	$(E_s)^2$	**-0.5592**	**-0.5398**	**(-)**	**0.4830**	**0.4824**	Anti descriptor	Acceptable after linearization; distribution problems
	QSAR	E_s	0.6375	0.6054	(+)	0.7542	0.7506	Real descriptor	Acceptable after linearization; distribution problems
		H–N–πS	0.3872	0.4133	(x)	0.2260	0.2368	Real descriptor	Not acceptable; problematic distinct groups and distribution
		–C=O electr char	0.8210	0.8055	(+)	0.3832	0.3845	Real descriptor	Acceptable; moderate distribution problems
26	MLR	α	0.9430	0.9235	(+)	0.9747	0.9557	Real descriptor	Good
	QSAR	E_{HOMO}	0.0743	0.1572	(x)	0.1810	0.2064	Real noise	Not acceptable; serious problems with distinct groups and dispersion
		qH^+	-0.6168	-0.6583	(-)	-0.1309	-0.2099	Real descriptor	Acceptable after removing distinct groups; distribution problems
27	MLR	$lgEnr_M$	0.8444	0.8400	0.8646	0.8857	0.8935	Real descriptor	Acceptable; moderate distribution problems
	QSAR	$GAP_{h1\text{-}M}$	-0.2712	-0.2644	-0.2936	-0.3304	-0.3299	Hidden noise	Not acceptable; problematic distinct groups and distribution
		$GAPV_{mM}$	-0.1497	-0.1226	-0.1419	-0.2753	-0.2628	Real noise	Not acceptable; problematic distinct groups and distribution
		μ_M	0.0451	0.0261	0.1430	0.1747	0.1541	Real noise	Not acceptable; problematic distinct groups and distribution
28	MLR	ALFA	0.9371	0.9416	0.9267	0.9677	0.9713	Real descriptor	Good
	QSAR	MVC	**-0.1846**	**-0.2264**	**0.0428**	**0.2402**	**0.2281**	Unstable noise	Not acceptable; problematic distinct groups and distribution
		FPSA	**-0.5021**	**-0.4558**	**-0.7380**	**0.0763**	**0.0683**	Anti descriptor	Acceptable after linearization and removing the outliers
29	MLR	$\mu_1\mu_2^{Std}$	**0.4336**	**0.4367**	**0.4249**	**-0.2929**	**-0.2934**	Anti descriptor	Not acceptable; problematic distinct groups and distribution
	QSPR	μ_{10}^{Std}	0.3974	0.3953	0.4074	0.4264	0.4262	Real descriptor	Not acceptable; problematic distinct groups and distribution
		$\mu_5^{Ab\text{-}R2}$	**0.4132**	**0.4107**	**0.4249**	**-0.7535**	**-0.7558**	Anti descriptor	Not acceptable; problematic distinct groups and distribution
		μ_1^{Hyd}	0.4624	0.4827	0.3900	0.0417	0.0430	Real descriptor	Acceptable after linearization; pronounced dispersion
		μ_1^{Dip2}	**0.2578**	**0.2659**	**0.2250**	**-0.0716**	**-0.0700**	Unstable noise	Not acceptable; problematic distinct groups and distribution
		μ_3^{Van}	0.5188	0.5228	0.5039	0.1802	0.1738	Real descriptor	Acceptable after linearization; pronounced dispersion
		$\mu_1\mu_4^{Dip2}$	0.3181	0.3216	0.3250	0.0678	0.0706	Real descriptor	Not acceptable; problematic distinct groups and distribution
		$\mu_4^{Ab\text{-}logL16}$	0.4678	0.4688	0.4645	0.1943	0.2102	Real descriptor	Acceptable after linearization; pronounced dispersion
		$\mu_4^{Ab\text{-}\Sigma\beta2o}$	0.4547	0.4544	0.4568	0.2879	0.2735	Real descriptor	Acceptable after linearization; pronounced dispersion
		μ_4^{Pols}	0.0943	0.0779	0.1485	0.0081	0.0116	Real noise	Not acceptable; problematic distinct groups and distribution
30	MLR	RDF020u	**0.2787**	**0.2087**	**(+)**	**-0.3265**	**-0.3576**	Unstable noise	Not acceptable; serious problems with distinct groups
	QSPR	Mor28e	**-0.0944**	**-0.0120**	**(+)**	**0.1528**	**0.1761**	Unstable noise	Acceptable after linearization; pronounced dispersion
		Mor07p	0.9273	0.9168	(+)	0.9328	0.9171	Real descriptor	Acceptable after linearization; distribution problems
31	MLR	RB	0.6409	0.6778	0.6224	0.4458	0.3902	Real descriptor	Acceptable after linearization; pronounced dispersion
	QSAR	HBA	**0.3827**	**0.3263**	**0.4149**	**-0.4460**	**-0.4385**	Anti descriptor	Acceptable after linearization; pronounced dispersion
		CHI	0.6731	0.7514	0.6324	0.7761	0.8096	Real descriptor	Accept; pronounced dispersion
32	MLR	DTsDeP1/dGP2	**-0.3841**	**-0.3792**	**-0.3991**	**0.0166**	**-0.3328**	Anti descriptor	Not acceptable; problematic distinct groups, outlier and distribution
	QSAR	lnDGsDiE1/pGE	**0.3294**	**0.4180**	**0.1606**	**-0.0635**	**0.4318**	Anti descriptor	Not acceptable; problematic distinct groups, outlier and distribution
		DTjDeMp/d2GP	0.6772	0.6845	0.6647	0.7842	0.5253	Real descriptor	Accept after removing the outlier
		lnDTjDeEp2/d2AE	**0.7422**	**0.7258**	**0.8420**	**-0.5539**	**0.5472**	Anti descriptor	Not acceptable; problematic distinct groups, outlier and distribution
		DTsDeP1/dGP2	**-0.0837**	**-0.0728**	**-0.1048**	**-0.0886**	**0.1212**	Unstable noise	Not acceptable; problematic distinct groups, outlier and distribution
		LnRGsDeMp2/d2AE	-0.6239	-0.5898	-0.8428	-0.2571	-0.3358	Real descriptor	Not acceptable; problematic distinct groups and distribution

33	MLR	X1	-0.1495	-0.1681	-0.0038	-0.2158	-0.3705	Hidden noise	Not acceptable; problematic distinct groups and distribution
	QSAR	X2	-0.3116	-0.3506	-0.1314	-0.1107	-0.1396	Quasi descriptor	Not acceptable; serious distribution problems
		X3	**0.4283**	**0.5493**	**-0.5945**	**0.2405**	**0.3533**	Anti descriptor	Acceptable after linearization and removing the outliers
		X4	0.3872	0.4389	0.5639	0.7113	0.6729	Real descriptor	Not acceptable; serious distribution problems; non-linearity
		X5	**0.6253**	**0.6811**	**-0.0024**	**0.2350**	**0.1665**	Anti descriptor	Not acceptable; problematic outliers and large dispersion
		X6	**0.0888**	**0.0311**	**0.5906**	**-0.5676**	**-0.4879**	Unstable noise	Not acceptable; serious problems with distinct groups; non-linearity
34	MLR	Dip	0.6328	0.6088	(+)	0.9477	0.9505	Real descriptor	Accept after removing the outlier
	QSAR	IP	0.1335	0.0224	(x)	0.2000	0.1622	Real noise	Not acceptable; problematic outliers and distribution
		Polar	-0.1662	-0.1979	(x)	-0.2489	-0.2652	Real noise	Not acceptable; problematic outliers and distribution
35	MLR	FH20	0.4514	0.4318	0.5197	0.2439	0.2687	Real descriptor	Accept after removing the outliers
	MI-QSAR	Dipole	**0.1283**	**0.1294**	**0.1254**	**-0.1335**	**-0.1281**	Unstable noise	Not acceptable; problematic distinct groups and distribution
		$\Delta\Sigma h(r)$	0.8203	0.8004	0.8914	0.9606	0.9547	Real descriptor	Acceptable after removing distinct groups
36	MLR	SX_{1CH}	**-0.9738**	**-0.9699**	**-0.9787**	**0.3574**	**0.3248**	Anti descriptor	Acceptable after linearization
	QSPR	SX_{1CC}	**0.9207**	**0.9167**	**0.9263**	**-0.2403**	**-0.2415**	Anti descriptor	Acceptable after linearization
		SV_{ij}	0.5176	0.5016	0.5377	0.0961	0.0841	Real descriptor	Not acceptable; serious problems with distinct groups; non-linearity
		OEI	0.9607	0.9547	0.9685	0.1946	0.1602	Real descriptor	Acceptable after linearization
		$N^{2/3}$	0.9910	0.9910	0.9912	0.8760	0.8964	Real descriptor	Acceptable after linearization
37	MLR	ASMmVQt	0.5171	0.5351	0.5214	0.4832	0.5344	Real descriptor	Not acceptable; serious problems with distinct groups; non-linearity
	QSAR	lfDdOQg	-0.5599	-0.4503	-0.8974	-0.7862	-0.7553	Real descriptor	Accept; pronounced dispersion
		InMrLQg	**-0.1839**	**-0.1877**	**-0.3048**	**0.2683**	**0.2576**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		LsDMpQg	**0.1422**	**0.0048**	**0.5864**	**-0.2766**	**-0.2786**	Unstable noise	Not acceptable; problematic outliers and large dispersion
38	MLR	ΔV	0.2472	0.2736	(x)	0.1580	0.1780	Real noise	Acceptable; moderate distribution problems
	QSAR	MR2	-0.5139	-0.6454	(x)	-0.1136	-0.1653	Real descriptor	Acceptable; moderate distribution problems
		ΔE_1	-0.2482	-0.2744	(x)	-0.6421	-0.6530	Hidden noise	Acceptable; moderate distribution problems
		$(\Delta E_1)^2$	**0.2036**	**0.2278**	**(x)**	**-0.7005**	**-0.6765**	Unstable noise	Not acceptable; serious problems with distinct groups
		ΔE_2	-0.4247	-0.4907	(-)	-0.2431	-0.2388	Real descriptor	Accept; pronounced dispersion
39	MLR	MR	0.6417	0.7267	0.4586	0.3170	0.3192	Real descriptor	Acceptable; moderate distribution problems
	QSAR	DM	0.4237	0.4886	0.2422	0.0289	0.0201	Quasi descriptor	Acceptable after linearization
		SASA	0.7350	0.8234	0.5289	0.5832	0.5940	Real descriptor	Acceptable after linearization
		Polrz	**0.6390**	**0.7265**	**0.4444**	**-0.7296**	**-0.6693**	Anti descriptor	Acceptable after linearization and removing the outliers
		LogPo/w	-0.4484	-0.5103	-0.2744	-0.1003	-0.2362	Quasi descriptor	Accept; pronounced dispersion
		LogS	-0.1628	-0.2813	-0.0477	-0.1272	-0.2026	Real noise	Not acceptable; serious problems with distinct groups and dispersion
40	MLR	β_{xxx}	0.2353	0.1393	0.2420	0.2579	0.1168	Real noise	Not acceptable; pronounced dispersion
	QSAR	β_{xyy}	-0.2370	-0.0362	-0.3758	-0.1795	-0.0048	Real noise	Not acceptable; serious problems with distinct groups and dispersion
		Ω_{xyz}	**0.3254**	**0.5188**	**-0.0675**	**0.1837**	**0.2373**	Anti descriptor	Not acceptable; serious problems with distinct groups and dispersion
		Ω_{zzz}	**0.0518**	**0.0959**	**-0.0931**	**-0.1799**	**-0.1191**	Unstable noise	Not acceptable; serious problems with distinct groups and dispersion
		$\Delta\alpha$	0.8645	0.8970	0.8463	0.9139	0.9570	Real descriptor	Good
41	MLR	x_2	-0.3397	-0.3584	(-)	-0.1222	-0.1371	Real descriptor	Not acceptable; problematic outliers and distribution
	QSAR	x_9	-0.4627	-0.5122	(-)	-0.6901	-0.7406	Real descriptor	Not acceptable; problematic outliers and distribution
		x_{21}	**0.0146**	**-0.0422**	**(+)**	**0.1052**	**0.1623**	Unstable noise	Not acceptable; problematic outliers
		x_{62}	0.5638	0.4745	(+)	0.7055	0.6375	Real descriptor	Not acceptable; problematic outlier

42	MLR	HOMO	-0.8265	-0.7854	(-)	-0.9448	-0.9219	Real descriptor	Acceptable; moderate distribution problems
	QSAR	TOE	-0.1535	-0.2709	(-)	-0.3278	-0.3875	Hidden noise	Not acceptable; serious problems with distinct groups
43	MLR	qC_{13}	0.1768	0.1326	0.3621	0.3538	0.3169	Hidden noise	Not acceptable; problematic distinct groups and distribution
	QSAR	GAP	0.2194	0.1618	0.4688	0.5264	0.5042	Hidden noise	Not acceptable; serious distribution problems
		SA	-0.3407	-0.2687	-0.5814	-0.2695	-0.2635	Quasi descriptor	Acceptable; pronounced dispersion
		nHAcc	-0.4166	-0.4035	-0.5050	-0.7246	-0.7589	Real descriptor	Not acceptable; only two distinct values of indicator variable
44	MLR	CRI	0.4006	0.3977	0.3985	0.4753	0.4976	Real descriptor	Accept after removing the outliers
	QSAR	E_{LUMO}	-0.8088	-0.7719	-0.8818	-0.8798	-0.8674	Real descriptor	Acceptable after removing distinct groups
45	MLR	$^{1}\chi^{f}(x,y)$	0.9935	0.9947	0.9889	0.9823	0.9838	Real descriptor	Good
	QSPR	$^{m}EM_2$	0.9311	0.9290	0.9440	0.1871	0.1790	Real descriptor	Good
46	MLR	V^2	**-0.1967**	**-0.4514**	**0.2028**	**-0.5854**	**-0.6117**	Unstable noise	Not acceptable; pronounced dispersion; non-linearity
	ADME	V	**-0.2004**	**-0.4605**	**0.2557**	**0.6852**	**0.6928**	Unstable noise	Not acceptable; pronounced dispersion; non-linearity
		PSA	-0.7767	-0.8561	-0.6886	-0.4333	-0.3819	Real descriptor	Good
47	MLR	x_5	0.5505	0.4746	0.7224	0.2999	0.2588	Real descriptor	Not acceptable; serious distribution problems
	QSAR	x_{21}	0.6323	0.6100	0.7137	0.4994	0.4957	Real descriptor	Acceptable; moderate distribution problems
		x_{26}	0.5940	0.5891	0.6104	0.6869	0.6959	Real descriptor	Acceptable; moderate distribution problems
		x_{32}	0.2951	0.2641	0.3559	0.4029	0.4063	Hidden noise	Not acceptable; serious problems with distinct groups; non-linearity
		x_{51}	**0.2118**	**0.2983**	**-0.1562**	**0.1628**	**0.1949**	Unstable noise	Not acceptable; serious distribution problems
48	MLR	Human Liver	0.7308	0.6675	(+)	0.3154	0.2991	Real descriptor	Good
	QSAAR	LUMO	-0.8367	-0.7880	(-)	-0.9111	-0.8961	Real descriptor	Acceptable; moderate distribution problems
		N_O	-0.0429	-0.1603	(x)	-0.2654	-0.3280	Hidden noise	Not acceptable; serious problems with distinct groups; non-linearity
49	MLR	α	0.4061	0.5739	(+)	0.7527	0.7604	Real descriptor	Acceptable; moderate distribution problems
	QSAR	α^2	**0.3330**	**0.4869**	**(+)**	**-0.6443**	**-0.6433**	Anti descriptor	Acceptable; moderate distribution problems
		F^{N}_{C*}	-0.4939	-0.2166	(-)	-0.1331	-0.0857	Real noise	Not acceptable; serious problems with distinct groups
		Q_{N**}	0.3333	0.3059	(+)	0.0239	0.0269	Real descriptor	Not acceptable; serious problems with distinct groups
50	MLR	$^{2}\Omega_p^{C}(q)$	-0.8798	-0.7368	-0.8671	-0.8537	-0.7527	Real descriptor	Good
	QSPR	$^{6}\varepsilon_{Ch}(\rho)$	-0.6902	-0.6520	-0.7495	-0.5207	-0.6584	Real descriptor	Acceptable after linearization and removing distinct groups

[a]Descriptors with the sign change problem (criterion I) are marked by bold values of regression and correlation coefficients. Descriptors that satisfy criterion IV (*i.e.*, all the criteria I, II and III) have names marked in green.

[b]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the complete dataset.

[c]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the training set after data split.

[d]Pearson correlation coefficient between a descriptor and the dependent variable **y** for the external validation set after data split. The correlation coefficient was not calculated for external sets with less than seven samples, but a qualitative parameter for correlation was determined from scatterplots: (+) – positive correlation, (-) – negative correlation, and (x) – direction of correlation could not be determined.

[e]Normalized regression vector for the complete dataset.

[f]Normalized regression vector for the training set after data split.

[g]Three types of descriptors (real, quasi and anti) and three types of noise variables (real, hidden and unstable), based on criterion II. Variables that did not fail according to this criterion are marked in pink.

[h]Variable characterization according to criterion III: good, acceptable (with our without some modest changes), and not acceptable descriptors. Diagnostics details are given so it can be understood why, how and how much a descriptor is problematic, and whether some action may be made to remedy this descriptor.

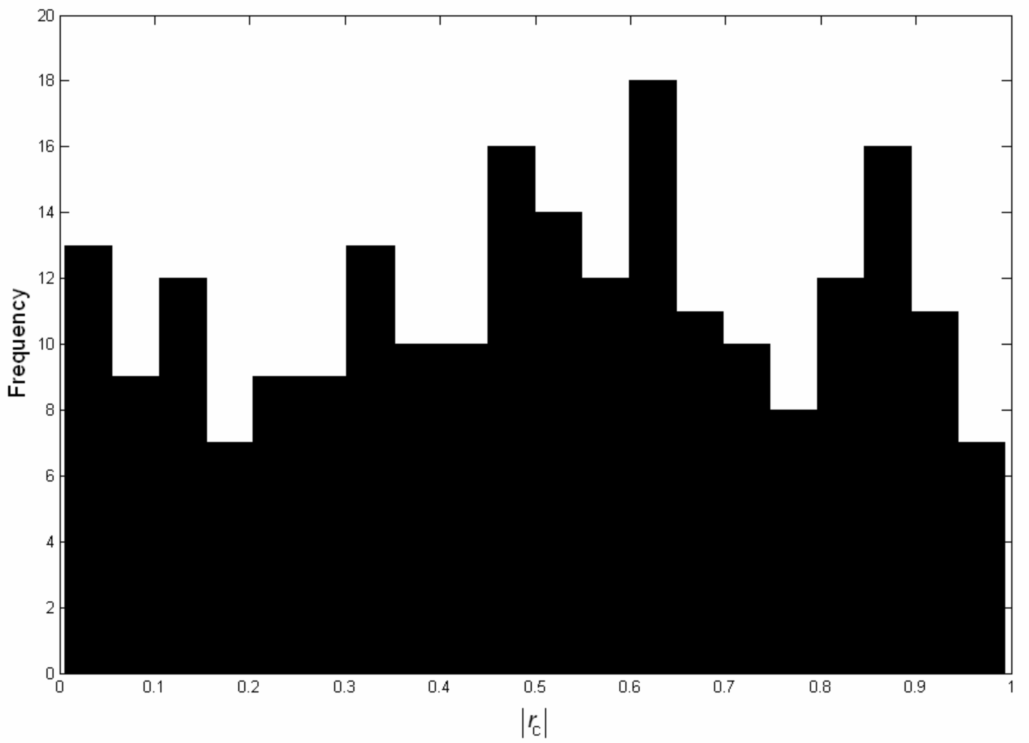

Figure S1. Frequency distribution of the absolute value of the Pearson correlation coefficient r_c for independent variables from complete datasets. Three main regions are visible: from 0 to 0.2, from 0.2 to 0.75, and from 0.75 to 1, roughly corresponding to very low to low, medium to high, and very high correlations.

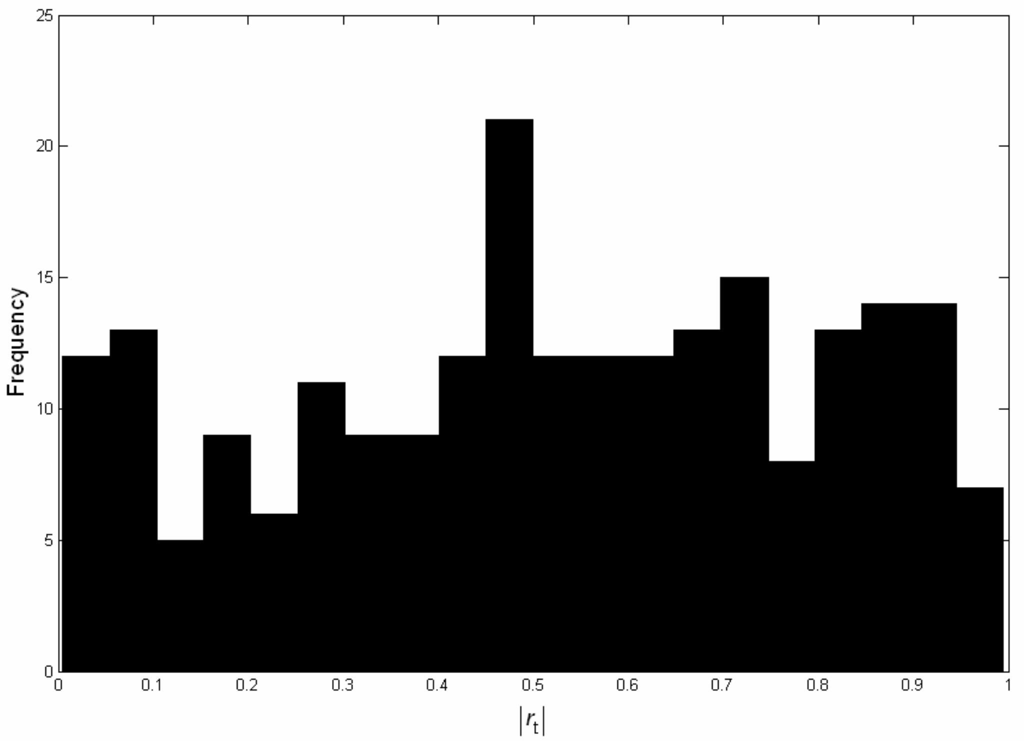

Figure S2. Frequency distribution of the absolute value of the Pearson correlation coefficient r_t for independent variables from training datasets. Three main regions are visible: from 0 to 0.15, from 0.15 to 0.75, and from 0.75 to 1, roughly corresponding to very low to low, medium to high, and very high correlations.

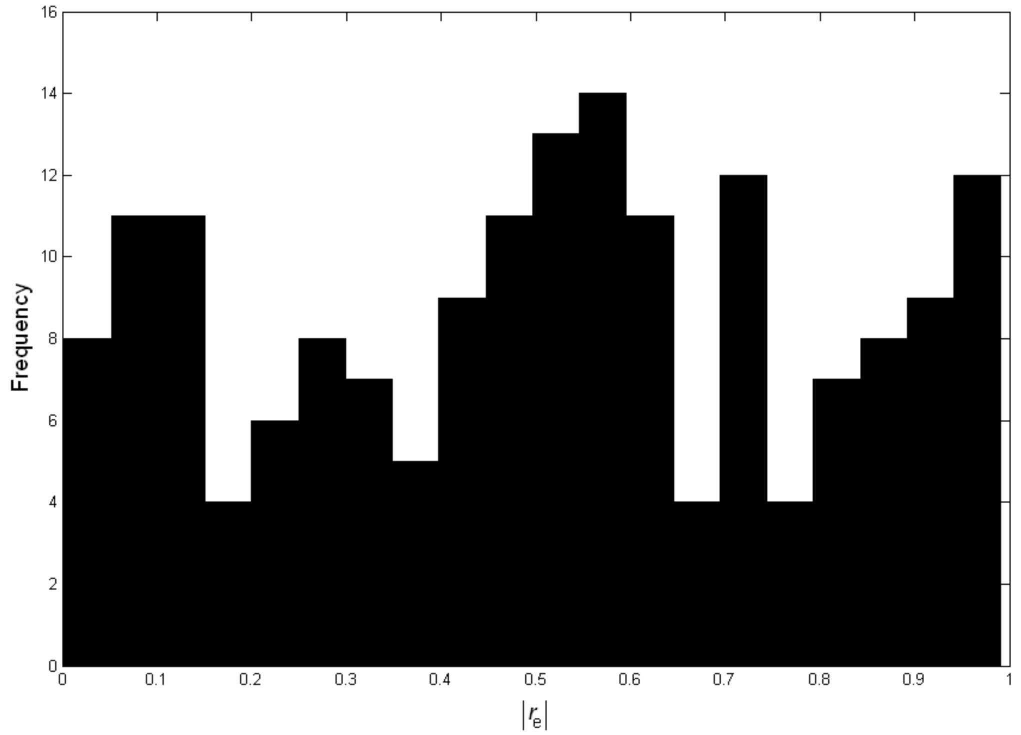

Figure S3. Frequency distribution of the absolute value of the Pearson correlation coefficient r_e for independent variables from external validation datasets. Three main regions are visible: from 0 to 0.15, from 0.15 to 0.75 (with three local peaks), and from 0.75 to 1, roughly corresponding to very low to low, medium to high, and very high correlations.

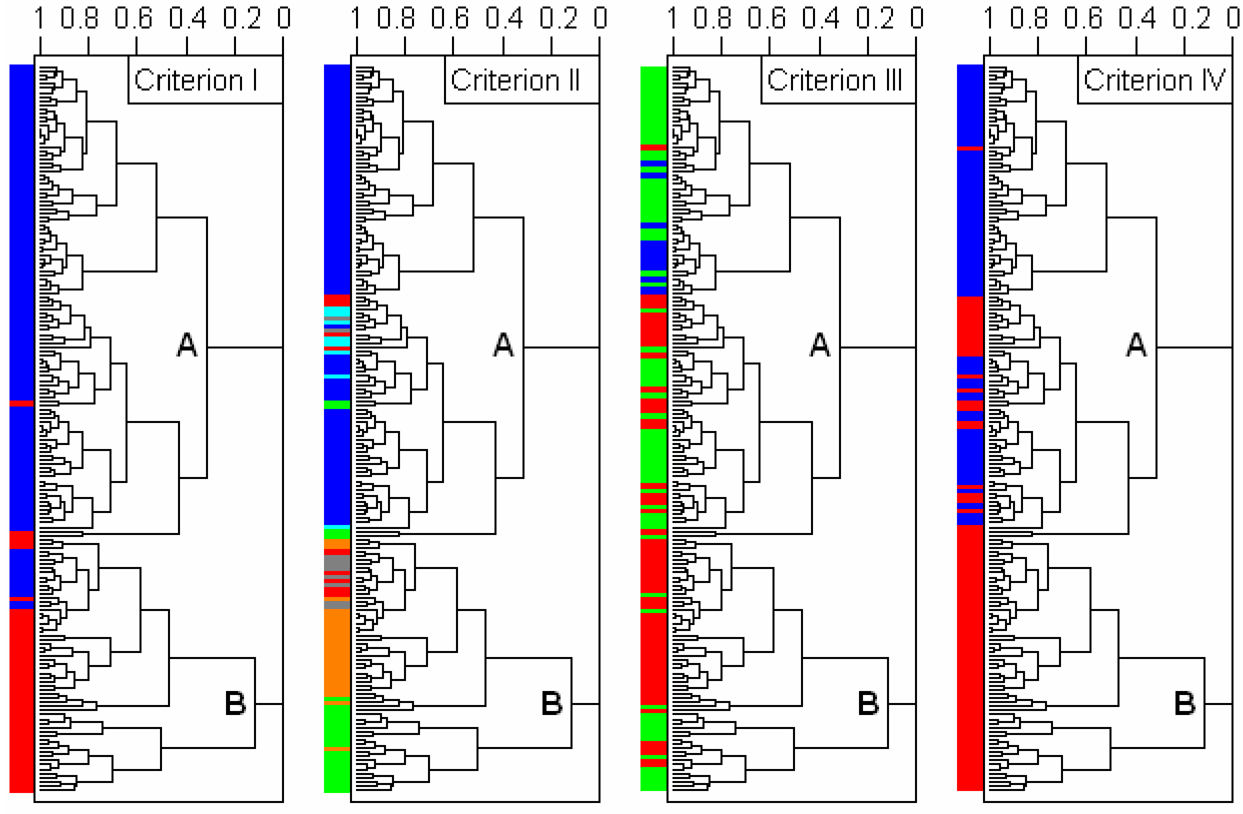

Figure S4. The samples dendogram from the HCA analysis of the four *F*-functions (F_1, F_2, F_3 and F_4), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

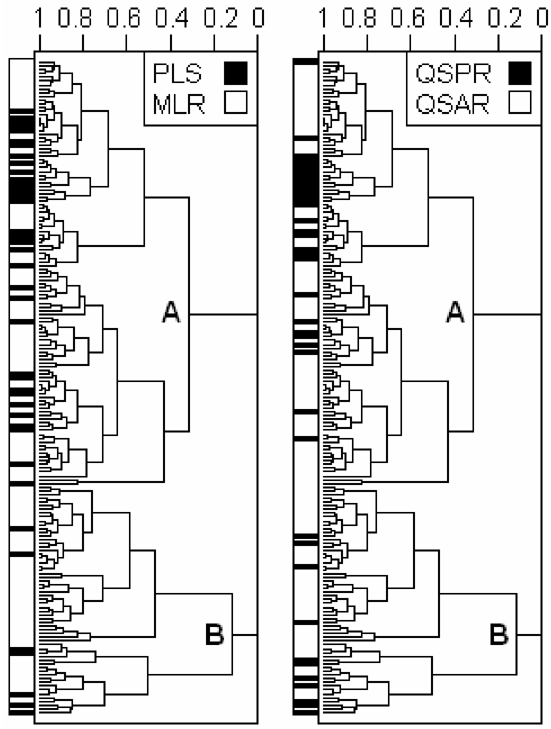

Figure S5. The samples dendogram from the HCA analysis of the four *F*-functions (F_1, F_2, F_3 and F_4), with distinction of PLS from MLR models, and QSPR from QSAR models.

<u>Comments for Figure S5.</u> This HCA dendrogram was inspected for two more classifications: 1) descriptors from PLS and from MLR models, and 2) descriptors from QSAR/QSAR-like models and from QSPR/QSPR-like models. The classes are not uniformly distributed over clusters A and B and their subclusters. PLS models in one classification and QSPR/QSPR-like models in the other one, although making minorities, tend to be concentrated more in cluster A than in B. Perhaps it reflects the basic distinction between MLR and PLS (use of significant fraction of original descriptors in PLS), and between QSAR and QSPR (the latter does not need frequent transformation of **y**, and is easier to interpret since corresponding chemical background is simpler).

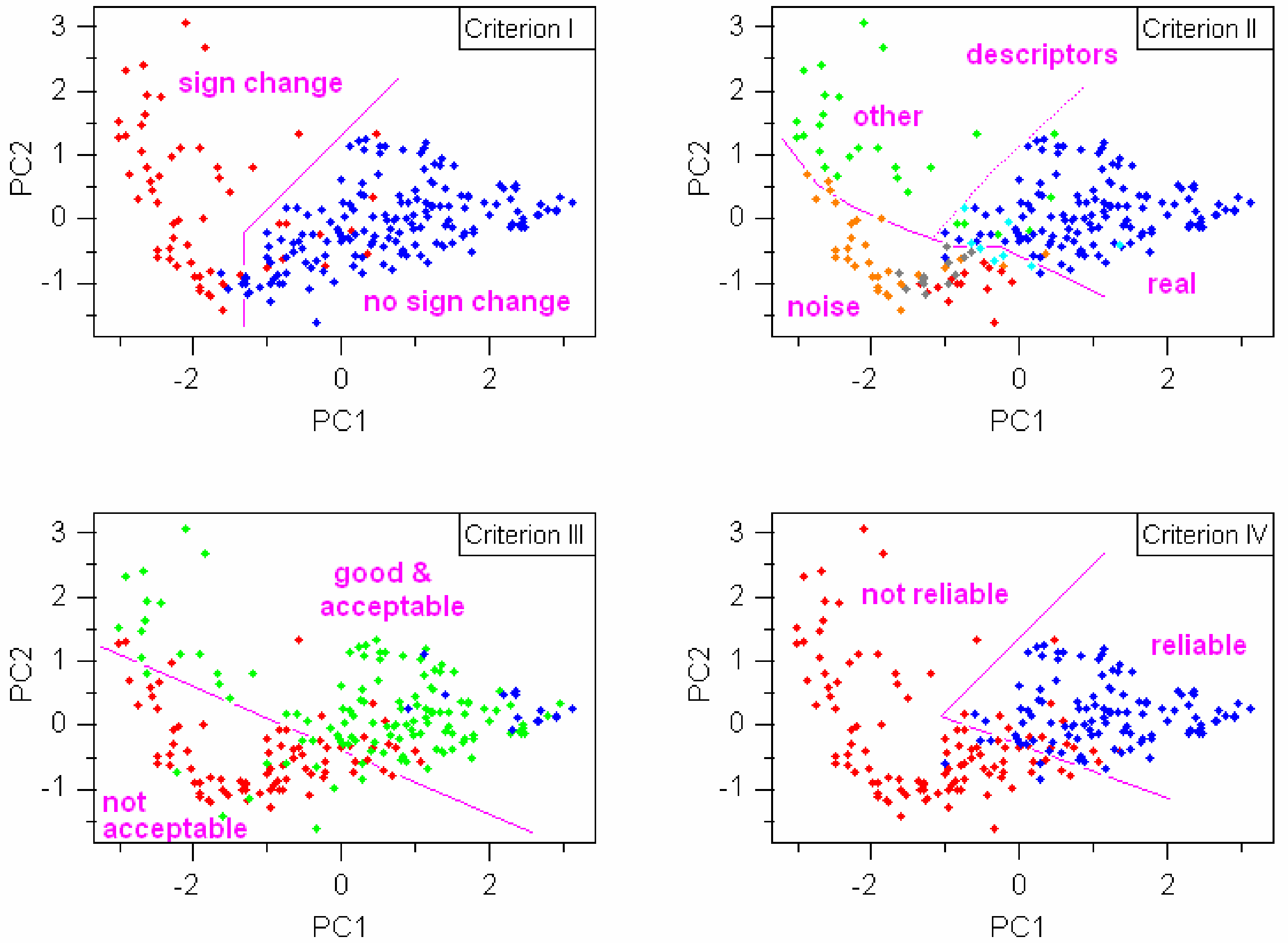

Figure S6. The PC1-PC2 scores plots from the PCA analysis of the three *F*-functions (F_1, F_3 and F_4), with classes of descriptors marked in different colors and separated by arbitrarily drawn magenta lines. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

<u>Comments for Figure S6.</u> The three *F*-functions for 227 descriptors, when explored with PCA, show similar trends as the four *F*-functions for 174 descriptors. The scores plots (PC1: 78% and PC2: 20% of the original variance) in Figure S6 show somewhat more mixing between classes of variables than in Figure 5.

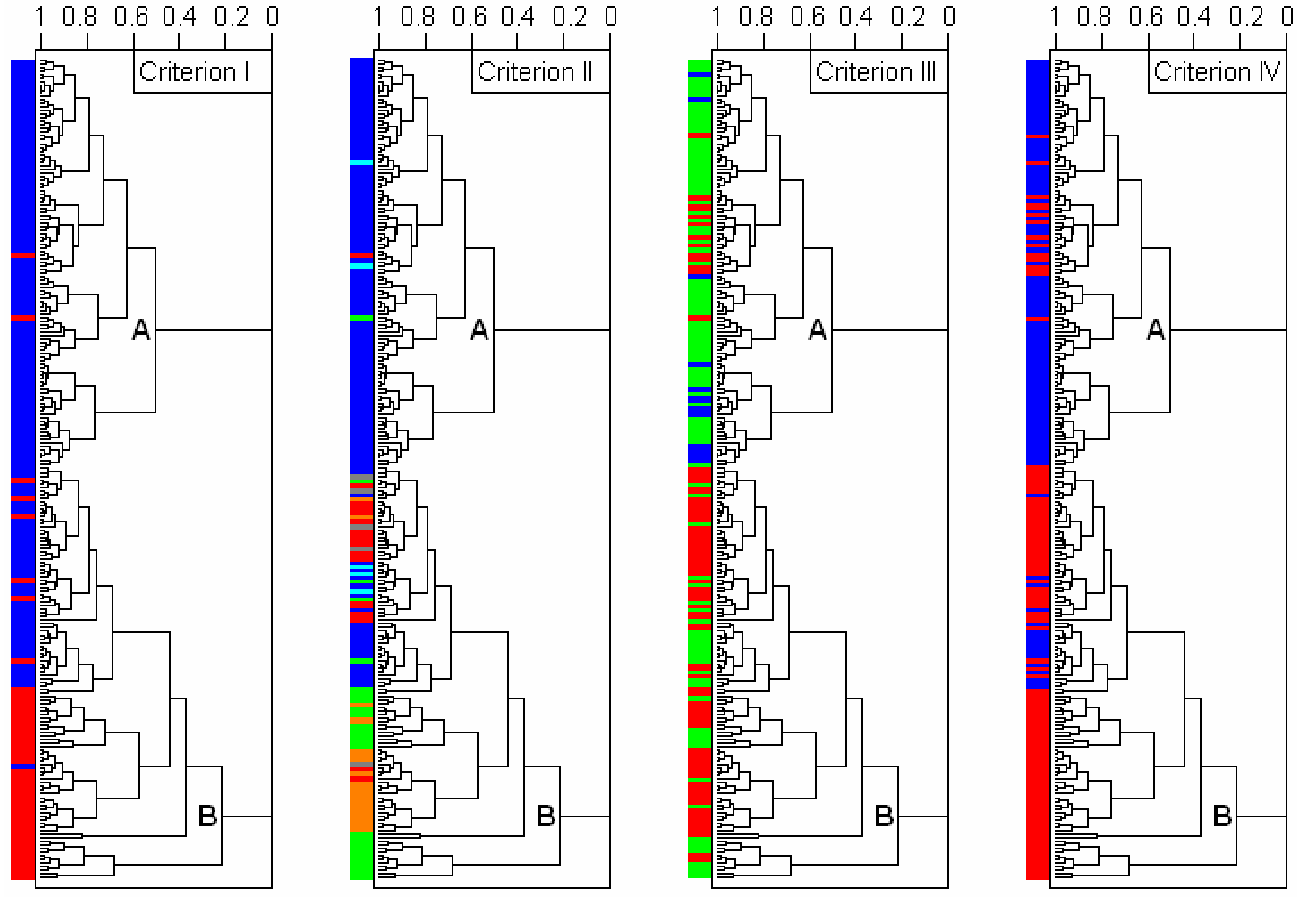

Figure S7. The samples dendogram from the HCA analysis of the three *F*-functions (F_1, F_3 and F_4), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

<u>Comments for Figure S7.</u> The three *F*-functions for 227 descriptors, when explored with HCA, show similar trends as the four *F*-functions for 174 descriptors. This notable similarity can be seen when comparing dendograms for descriptor accounting for the four classifications in Figures S4 and S8.

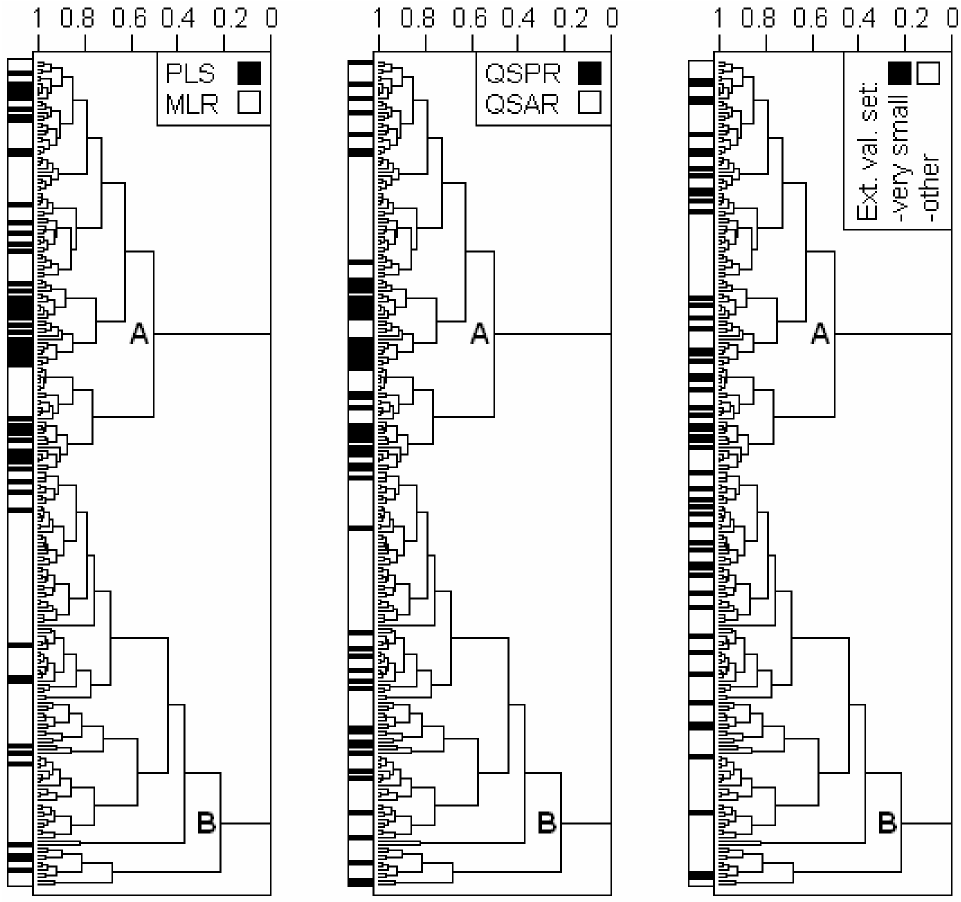

Figure S8. The samples dendogram from the HCA analysis of the three *F*-functions (F_1, F_3 and F_4), with distinction of PLS from MLR models, QSPR from QSAR models, and small from moderate and large external validation sets (small sets have less than seven samples).

<u>Comments for Figure S8.</u> The three *F*-functions for 227 descriptors, when explored with HCA, show similar trends as the four *F*-functions for 174 descriptors. This notable similarity can be seen in dendograms when considering PLS-MLR and QSAR-QSPR distinctions (Figures S5 and S8). This similarity may be due to rather uniform distribution of descriptors from datasets with very small external sets (less than seven samples), as visible in Figure S8 right.

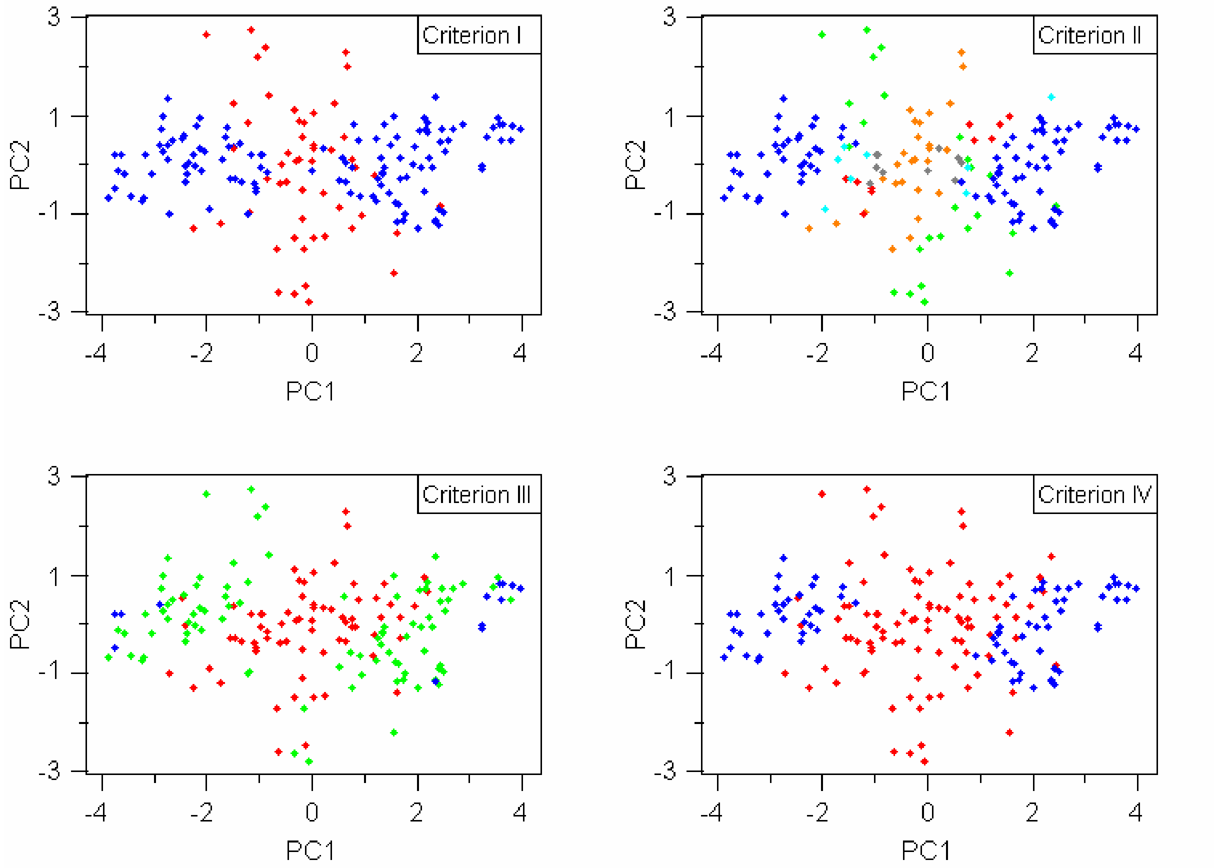

Figure S9. The PC1-PC2 scores plots from the PCA analysis of the five parameters (r_c, r_t, r_e, β_c and β_t), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

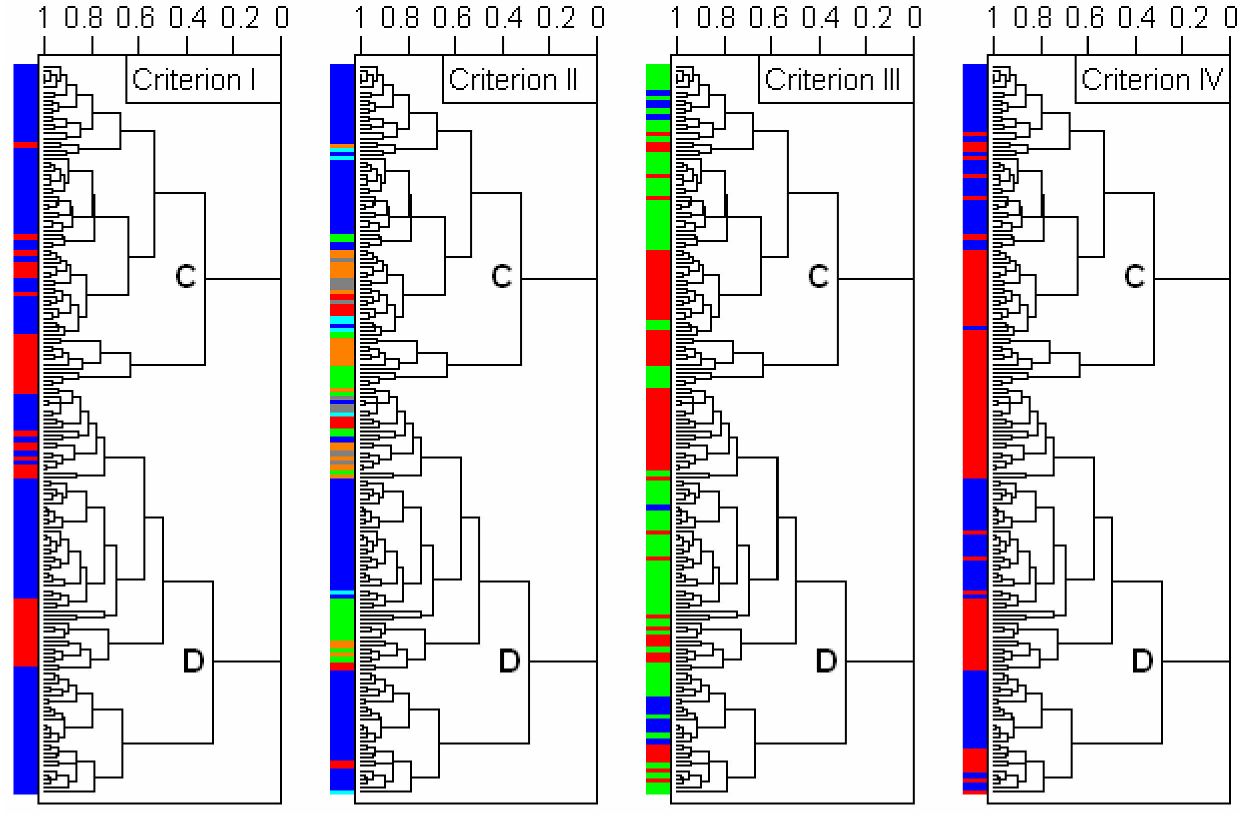

Figure S10. The samples dendogram from the HCA analysis of the five parameters (r_c, r_t, r_e, β_c and β_t), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

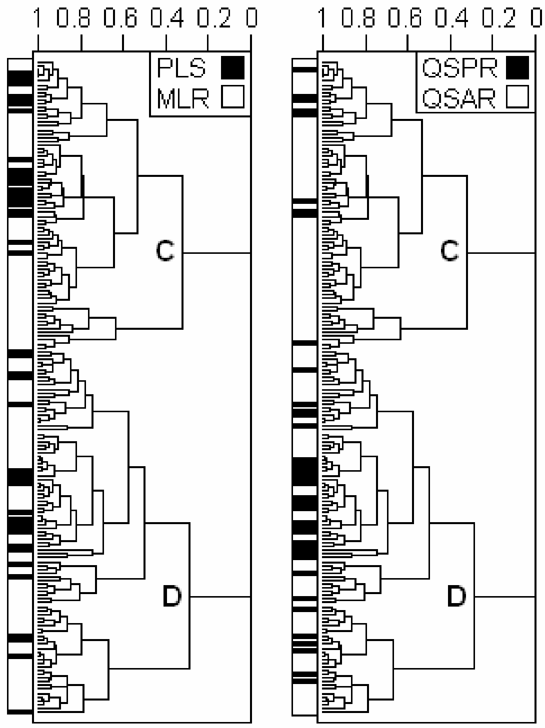

Figure S11. The samples dendogram from the HCA analysis of the five parameters (r_c, r_t, r_e, β_c and β_t), with distinction of PLS from MLR models, and QSPR from QSAR models.

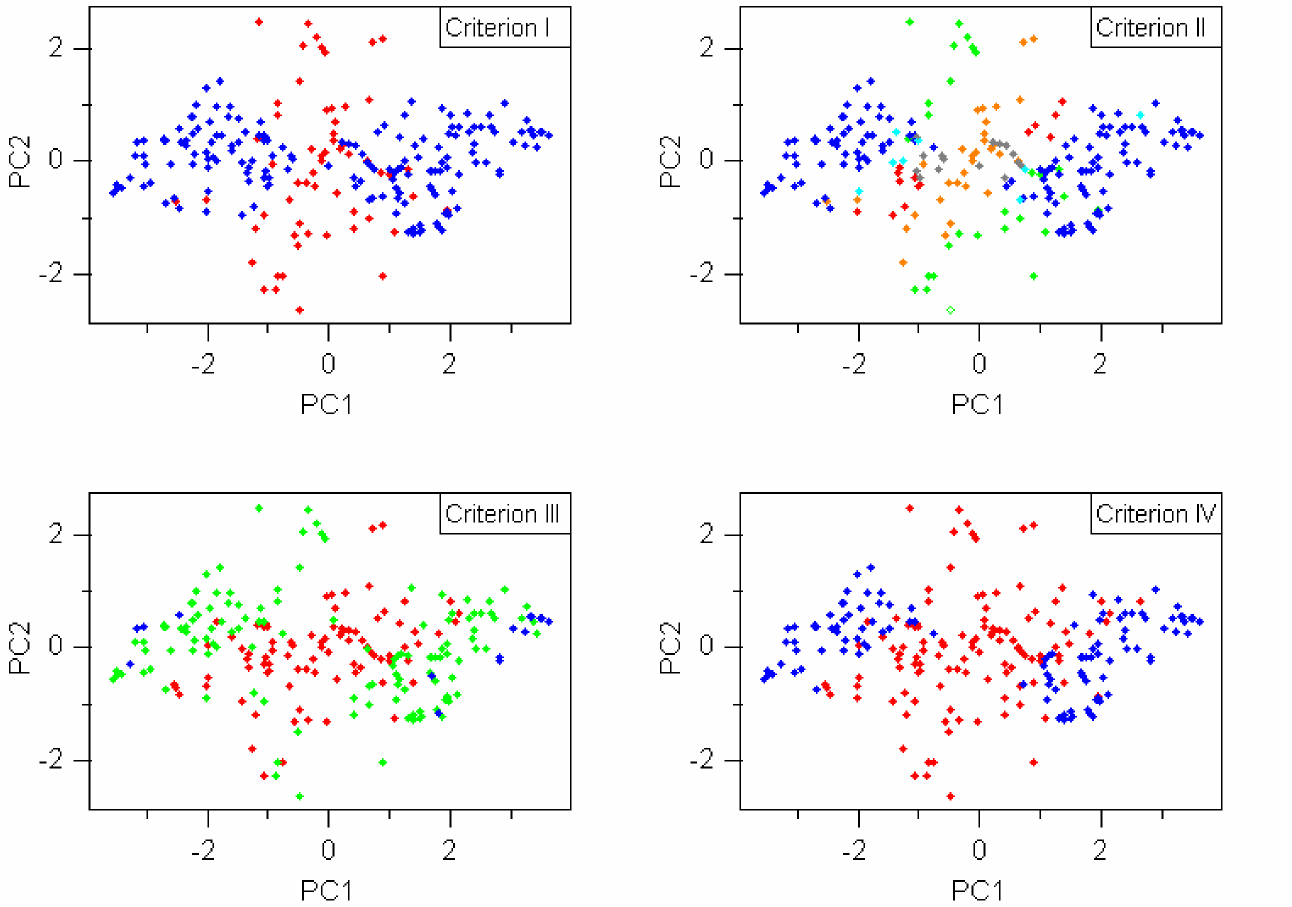

Figure S12. The PC1-PC2 scores plots from the PCA analysis of the four parameters (r_c, r_t, β_c and β_t), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

Comments for Figure S12. The four parameters (r_c, r_t, β_c and β_t) for 227 descriptors, when explored with PCA, show similar trends as the five parameters (r_c, r_t, r_e, β_c and β_t) for 174 descriptor. The scores plots (PC1: 82% and PC2: 17%) in Figure S12 show practically the same trend as the analogous plots in Figure S9.

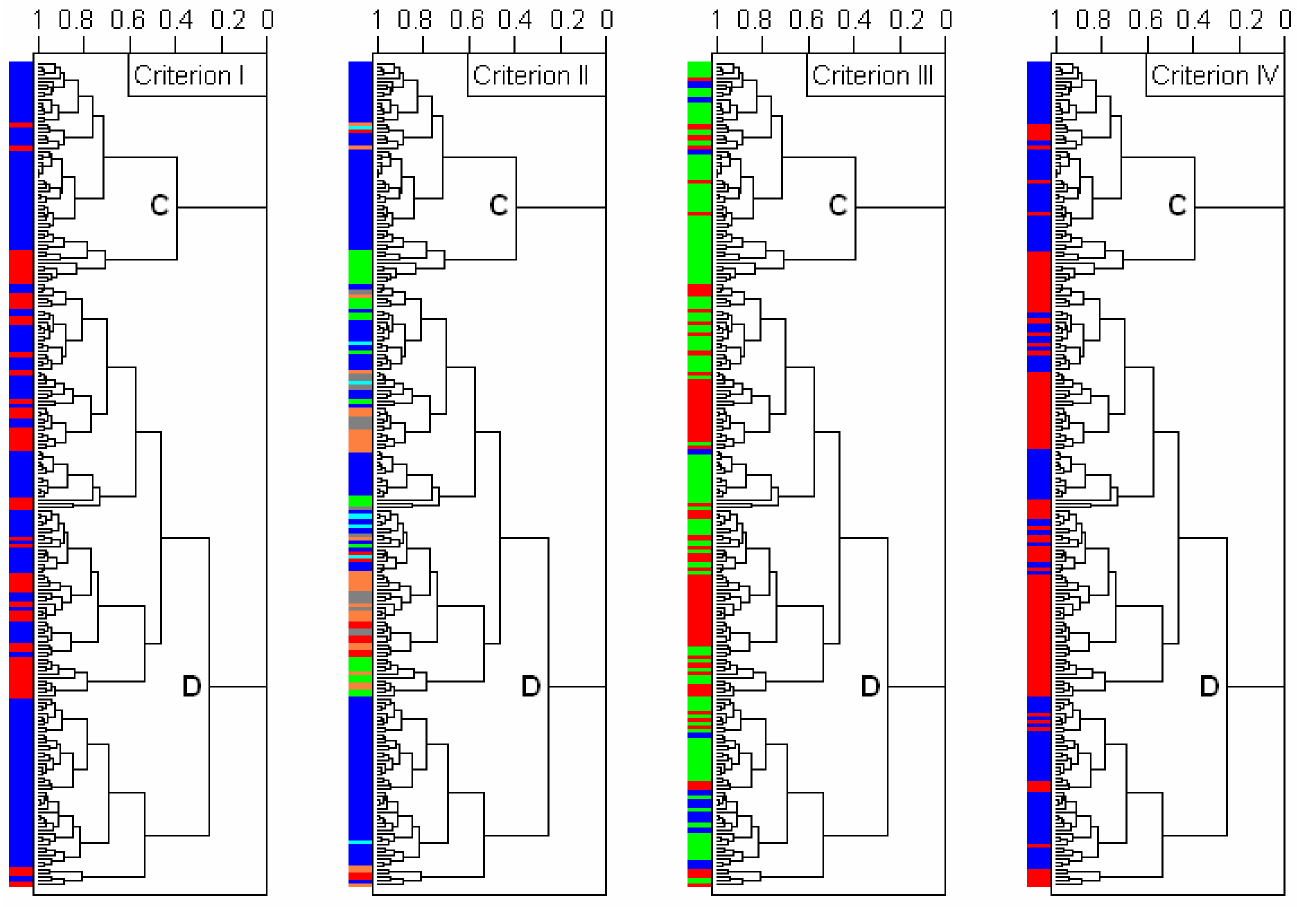

Figure S13. The samples dendogram from the HCA analysis of the four parameters (r_c, r_t, β_c and β_t), with classes of descriptors marked in different colors. Criterion I: blue – no sign change, red – sign change present; Criterion II: blue, cyan, green – real, quasi, anti descriptors, respectively, and gray, orange, red – real, unstable, hidden noise, respectively; Criterion III: blue – good, green – acceptable, red – not acceptable **x-y** scatterplots; Criterion IV: blue – reliable, red – not reliable descriptors.

Comments for Figure S13. The four parameters (r_c, r_t, β_c and β_t) for 227 descriptors, when explored with HCA, show similar trends as the five parameters (r_c, r_t, r_e, β_c and β_t) for 174 descriptor. This notable similarity can be seen when comparing dendograms for descriptor accounting for the four classifications in Figures S10 and S13.

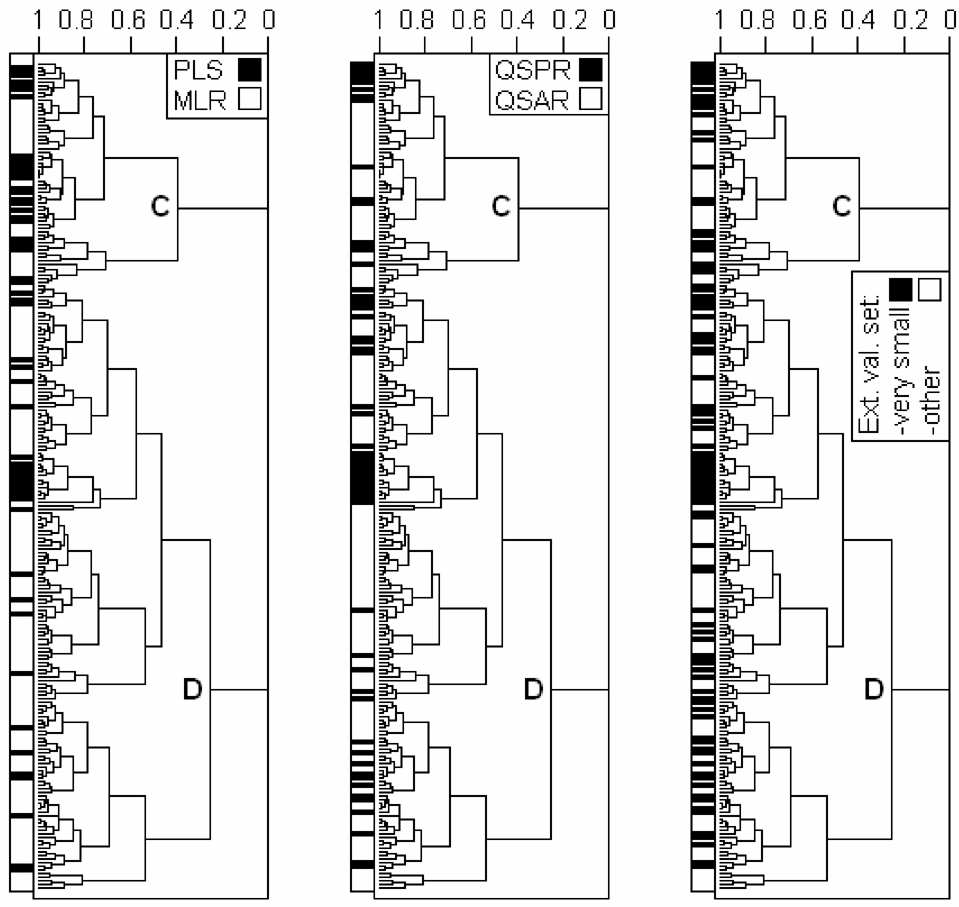

Figure S14. The samples dendogram from the HCA analysis of the four parameters (r_c, r_t, β_c and β_t), with distinction of PLS from MLR models, and QSPR from QSAR models, and small from moderate and large external validation sets (small sets have less than seven samples).

Comments for Figure S14. The four parameters (r_c, r_t, β_c and β_t) for 227 descriptors, when explored with HCA, show similar trends as the five parameters (r_c, r_t, r_e, β_c and β_t) for 174 descriptor. This notable similarity can be seen in dendograms when considering PLS-MLR and QSAR-QSPR distinctions (Figures S11 and S14). This similarity may be due to rather uniform distribution of descriptors from datasets with very small external sets (less than seven samples), as visible in Figure S14 right.

Table S3. Bivariate linear regression models of the form $\hat{\mathbf{y}} = \alpha_c + \beta_{1c}\,\mathbf{x}_1 + \beta_{2c}\,\mathbf{x}_2$ related to the dataset 48.*

Regression	Human Liver (r_1 = 0.7308)				LUMO (r_1 = -0.8367)				N_O (r_1 = -0.0429)			
No.	β_{1c}	β_{2c}	r_2	r_{12}	β_{1c}	β_{2c}	r_2	r_{12}	β_{1c}	β_{2c}	r_2	r_{12}
1	0.4350	-0.9004	-0.8367	-0.6752	-0.9004	0.4350	0.7308	-0.6752	-0.0657	0.9978	0.7308	0.0072
2	0.9978	-0.0657	-0.0429	0.0072	-0.9637	-0.2670	-0.0429	-0.2290	-0.2670	-0.9637	-0.8367	-0.2290
3	0.9149	0.4036	0.6165	0.6413	-0.9918	0.1275	0.6165	-0.6720	-0.3608	0.9327	0.6165	0.3260
4	0.8827	0.4699	0.6270	0.5994	-0.9863	0.1648	0.6270	-0.6656	-0.2357	0.9718	0.6270	0.1771
5	0.7221	0.6918	0.7274	0.8040	-0.9661	0.2583	0.7274	-0.7842	-0.3364	0.9417	0.7274	0.3047
6	1.0000	-0.0094	0.5187	0.7144	-0.9920	0.1265	0.5187	-0.5346	-0.0051	1.0000	0.5187	-0.0777
7	0.9986	0.0533	0.5533	0.7334	-0.9904	0.1384	0.5533	-0.5746	-0.0313	0.9995	0.5533	-0.0463
8	0.9735	-0.2288	-0.6181	-0.7624	-0.9999	-0.0104	-0.6181	0.7340	-0.2458	-0.9693	-0.6181	-0.1875
9	0.9680	-0.2510	-0.6270	-0.7700	-0.9998	0.0202	-0.6270	0.7581	-0.2788	-0.9603	-0.6270	-0.2264
10	0.8680	-0.4965	-0.6038	-0.4823	-0.9813	-0.1923	-0.6038	0.6123	-0.4479	-0.8941	-0.6038	-0.4458
11	0.9899	-0.1420	-0.2449	-0.2013	-0.9989	-0.0467	-0.2449	0.2493	-0.5860	-0.8103	-0.2449	-0.6276
13	0.8887	-0.4585	-0.5566	-0.4047	-0.9999	0.0153	-0.5566	0.6736	-0.0092	-1.0000	-0.5566	0.0679
13	0.9963	-0.0859	-0.0376	0.0350	-0.9962	0.0869	-0.0376	0.1316	-0.8175	-0.5759	-0.0376	0.4485
14	0.9914	-0.1309	0.3333	0.5548	-0.9995	0.0307	0.3333	-0.3722	-0.2916	0.9565	0.3333	0.1834
15	0.6345	0.7729	0.7502	0.7641	-0.9176	0.3974	0.7502	-0.7579	-0.2741	0.9617	0.7502	0.2316
16	0.9998	0.0197	0.5417	0.7322	-0.9827	0.1853	0.5417	-0.5227	-0.3315	0.9435	0.5417	0.2799
17	0.9033	0.4291	0.5659	0.4736	-0.9859	0.1671	0.5659	-0.5725	-0.5233	0.8521	0.5659	0.5646
18	0.6410	-0.7675	-0.7654	-0.5898	-0.9847	-0.1744	-0.7654	0.8803	-0.4104	-0.9119	-0.7654	-0.4043
19	0.9594	0.2819	0.2377	0.0347	-0.9983	-0.0578	0.2377	-0.3364	-0.5772	0.8166	0.2377	0.6033
20	0.9087	0.4174	0.6516	0.7323	-0.9937	0.1117	0.6516	-0.7303	-0.4469	0.8946	0.6516	0.4485
21	0.9423	0.3348	0.5145	0.4651	-0.9004	0.4350	0.7308	-0.5483	-0.5040	0.8637	0.5145	0.5257

*The dataset 48 from Table S1, with extension to the full descriptor pool (22 independent variables in total) [31]. The dependent variable is human toxicity HAP. Variable 1 is one of the three descriptors in the final trivariate model for prediction of HAP (see Table S1): Human Liver (human liver toxicity as –logIC_{50}), LUMO or N_o (number of oxygen atoms). Descriptor 2 is any other of the 21 variables. Index "c" means complete dataset, *i.e.*, no data split was applied. Regression coefficients β_{1c} and β_{2c} are for descriptors 1 and 2, respectively. Pearson correlation coefficients r_1 and r_2 account for correlation between descriptor 1 and HAP, and descriptor 2 and HAP, respectively. The correlation coefficient r_{12} is for correlation between descriptors 1 and 2.

Table S4. Multicollinearity effects on multiple linear regression in analytical form.

m[a]	Linear regressions[b]	Multiple linear regression[c]	No multicollinearity[d]	Multicollinearity present[e]
1[f]	$\hat{\mathbf{y}} = a + b\mathbf{x}$; r; R; $R = \|r\|$	–	–	–
2	$\hat{\mathbf{y}} = a_1 + b_1\mathbf{x}_1$; r_1; R_1 $\hat{\mathbf{y}} = a_2 + b_2\mathbf{x}_2$; r_2; R_2; r_{12}	$\hat{\mathbf{y}} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2$; R	$\beta_1 = b_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} r_1$ $\beta_2 = b_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} r_2$ $\alpha = a_1 + a_2 - \overline{y}$ $R^2 = R_1^2 + R_2^2$	$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} \frac{r_1 - r_2 r_{12}}{1 - r_{12}^2} \neq b_1$ $\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \frac{r_2 - r_1 r_{12}}{1 - r_{12}^2} \neq b_2$ $\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} \neq a_1 + a_2 - \overline{y}$ $R^2 < R_1^2 + R_2^2$
3	$\hat{\mathbf{y}} = a_1 + b_1\mathbf{x}_1$; r_1; R_1 $\hat{\mathbf{y}} = a_2 + b_2\mathbf{x}_2$; r_2; R_2; r_{12} $\hat{\mathbf{y}} = a_3 + b_3\mathbf{x}_3$; r_3; R_3; r_{13}; r_{23}	$\hat{\mathbf{y}} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3$; R	$\beta_1 = b_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} r_1$ $\beta_2 = b_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} r_2$ $\beta_3 = b_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} r_3$ $\alpha = a_1 + a_2 + a_3 - \overline{2y}$ $R^2 = R_1^2 + R_2^2 + R_3^2$	$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} \frac{f_1(r_1, r_2, r_3, r_{12}, r_{13}, r_{23})}{g_1(r_{12}, r_{13}, r_{23})} \neq b_1$ $\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \frac{f_2(r_1, r_2, r_3, r_{12}, r_{13}, r_{23})}{g_1(r_{12}, r_{13}, r_{23})} \neq b_2$ $\beta_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} \frac{f_3(r_1, r_2, r_3, r_{12}, r_{13}, r_{23})}{g_1(r_{12}, r_{13}, r_{23})} \neq b_3$ $\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - \beta_3 \overline{x_3} \neq a_1 + a_2 + a_3 - \overline{2y}$ $R^2 < R_1^2 + R_2^2 + R_3^2$
4	$\hat{\mathbf{y}} = a_1 + b_1\mathbf{x}_1$; r_1; R_1 $\hat{\mathbf{y}} = a_2 + b_2\mathbf{x}_2$; r_2; R_2; r_{12}; r_{13} $\hat{\mathbf{y}} = a_3 + b_3\mathbf{x}_3$; r_3; R_3; r_{23}; r_{24} $\hat{\mathbf{y}} = a_4 + b_4\mathbf{x}_4$; r_4; R_4; r_{14}; r_{34}	$\hat{\mathbf{y}} = \alpha + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4$; R	$\beta_1 = b_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} r_1$ $\beta_2 = b_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} r_2$ $\beta_3 = b_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} r_3$ $\beta_4 = b_4 = \sqrt{\frac{\sigma_y}{\sigma_{x4}}} r_4$	$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} \frac{f_1(r_1, r_2, r_3, r_4, r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})}{g(r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})} \neq b_1$ $\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \frac{f_2(r_1, r_2, r_3, r_4, r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})}{g(r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})} \neq b_2$ $\beta_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} \frac{f_3(r_1, r_2, r_3, r_4, r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})}{g(r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})} \neq b_3$ $\beta_4 = \sqrt{\frac{\sigma_y}{\sigma_{x4}}} \frac{f_4(r_1, r_2, r_3, r_4, r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})}{g(r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34})} \neq b_4$

			$\alpha = a_1 + a_2 + a_3 + a_4 - 3\overline{y}$ $R^2 = R_1^2 + R_2^2 + R_3^2 + R_4^2$	$\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - \beta_3 \overline{x_3} - \beta_4 \overline{x_4} \neq$ $\neq a_1 + a_2 + a_3 + a_4 - 3\overline{y}$ $R^2 < R_1^2 + R_2^2 + R_3^2 + R_4^2$
m[g]	$\hat{\mathbf{y}} = a_j + b_j \mathbf{x}_j$; r_j; R_j; $j = 1, 2, ..., m$; $k = 1, 2, ..., m\text{-}1$; r_{kl}, $m \geq l \geq k+1$	$\hat{\mathbf{y}} = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + ... + \beta_m \mathbf{x}_m$; R	$\beta_j = b_j = \sqrt{\frac{\sigma_y}{\sigma_{xj}}} r_j$ $\alpha = a_1 + a_2 + ... + a_m - (m-1)\overline{y}$ $R^2 = R_1^2 + R_2^2 + ... + R_m^2$	$\beta_j = \sqrt{\frac{\sigma_y}{\sigma_{xj}}} \frac{f_j(r_1, r_2, ..., r_m, r_{12}, r_{13}, ..., r_{m-1,m})}{g(r_{12}, r_{13}, ..., r_{m-1,m})} \neq b_j$ $\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - ... - \beta_m \overline{x_m} \neq$ $\neq a_1 + a_2 + ... + a_m - (m-1)\overline{y}$ $R^2 < R_1^2 + R_2^2 + ... + R_m^2$

[a]Number of independent variables (descriptors). For univariate or simple linear regression is $m = 1$.
[b]Univariate regressions for all descriptors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$. The predicted values of the dependent variable $\mathbf{y}$ are marked with $\hat{\mathbf{y}}$, a_1, a_2, …, a_m are the constant coefficients, and b_1, b_2, …, b_m are regression coefficients for descriptors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$, respectively. The Pearson correlation coefficients for correlations between descriptors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$ and $\mathbf{y}$ are r_1, r_2, …, r_m, respectively, and r_{12}, r_{13}, ..., $r_{m-1,m}$ are correlation coefficients for descriptors' intercorrelations. R_1, R_2, ..., R_m are the coefficients of univariate determinations for descriptors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$, respectively.
[c]Multiple linear regression (MLR) for descriptors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$, and β_1, β_2, …, β_m are regression coefficients, respectively, and α is the constant coefficient. R is the coefficients of multiple determinations.
[d]Expressions for estimated regression coefficients and statistical parameters for MLR, connecting the multivariate with corresponding univariate regressions in the absence of multicollinearity ($r_{12} = r_{13} = ... = r_{m-1,m} = 0$). The average value of $\mathbf{y}$ is marked with symbol $\overline{y}$. Standard deviations for $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_m$ are σ_y, σ_{x1}, σ_{x2}, …, σ_{xm}, respectively.
[e]Expressions for estimated regression coefficients and statistical parameters for MLR in the presence of multicollinearity, showing significant deviations from the simple relationships between MLR and respective univariate regressions in the absence of multicollinearity. Expressions f_1, f_2, …, f_m and g are complicated polynomial functions of Pearson correlation coefficients r_1, r_2, …, r_m, and r_{12}, r_{13}, ..., $r_{m-1,m}$.
[f]Indices to distinguish regression coefficients and statistical parameters for descriptors are not applicable for univariate regression model as the final model.
[g]Index j is a general index for m descriptors and m univariate regression equations. Index l accounts for intercorrelations together with the index k.

Comments on Table S4:

Multicollinearity in MLR causes several effects:

1) A regression coefficient is not more a measure of variation of **y** caused by variation of the corresponding independent variable whilst all other variables are held constant. Its respective regression coefficients in univariate and multivariate regression models are not more equal

2) The sum of coefficients of determination from the corresponding univariate regressions is not equal but is significantly greater than the coefficient of determination of the MLR model, since significant portion of the original variance is repeated in mutually correlated descriptors.

APPENDIX 1: EXPRESSIONS FOR REGRESSION COEFFICIENTS IN MLR WITH 2, 3 AND 4 INDEPENDENT VARIABLES

ESTIMATED PARAMETERS FOR SOME MULTIPLE LINEAR REGRESSIONS (BI-, TRI- AND QUADRIVARIATE)

Legend

$\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ – independent (predictor) variables;

$\mathbf{y}$ – observed values of the dependent (response) variable;

$\hat{\mathbf{y}}$ – predicted values of the dependent (response) variable;

$\overline{x_1}$, $\overline{x_2}$, $\overline{x_3}$, $\overline{x_4}$, $\overline{y}$ – average values of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{y}$, respectively;

α – intercept of the linear regression (hyper)plane from the multivariate regression equation;

β_1, β_2, β_3, β_4 – regression coefficients for $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ in multivariate models, respectively;

σ_y, σ_{x1}, σ_{x2}, σ_{x3}, σ_{x4} – standard deviations of $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$, respectively;

r_1, r_2, r_3, r_4 – Pearson correlation coefficients for pairs of variables ($\mathbf{x}_1$, $\mathbf{y}$), ($\mathbf{x}_2$, $\mathbf{y}$), ($\mathbf{x}_3$, $\mathbf{y}$) and ($\mathbf{x}_4$, $\mathbf{y}$), respectively;

r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34} – Pearson correlation coefficients for pairs of variables ($\mathbf{x}_1$, $\mathbf{x}_2$), ($\mathbf{x}_1$, $\mathbf{x}_3$), ($\mathbf{x}_1$, $\mathbf{x}_4$), ($\mathbf{x}_2$, $\mathbf{x}_3$), ($\mathbf{x}_2$, $\mathbf{x}_4$) and ($\mathbf{x}_3$, $\mathbf{x}_4$), respectively.

Note.

Expressions for regression coefficients were derived from expressions for normal equations. The following literature was used:

-examples of deriving β_1 and β_2 in bivariate regression [Janke, S. J.; Tinsley, F. C. *Introduction to Linear Models and Statistical Inference*. Wiley: Hoboken, NJ, 2005, p 354; Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. *Applied Linear Statistical Models*, 5th ed. McGraw-Hill: Boston, MA, 2005, p. 276.];

-expressions for inverse matrices of dimensions 2×2, 3×3 and 4×4 [Inverse matrix of 2-by-2 matrix, 3-by-3 matrix, 4-by-4 matrix. The site of D. Miyazaki at the University of Tokyo, accessed on October 9, 2009. http://www.cvl.iis.u-tokyo.ac.jp/~miyazaki/tech/teche23.html];

-general information about Laplace expansion in matrix inversion [Laplace expansion. Wikipedia - the free encyclopedia. Wikimedia Foundations, San Francisco, CA. Accessed on October 8, 2009. http://en.wikipedia.org/wiki/Laplace_expansion].

BIVARIATE LINEAR REGRESSION

$$\hat{\mathbf{y}} = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$$

$$\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2}$$

$$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} \frac{r_1 - r_2 r_{12}}{1 - r_{12}^2}$$

$$\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \frac{r_2 - r_1 r_{12}}{1 - r_{12}^2}$$

TRIVARIATE LINEAR REGRESSION

$$\hat{\mathbf{y}} = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3$$

$$\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - \beta_3 \overline{x_3}$$

$$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x1}}} \frac{r_1\left(1 - r_{23}^2\right) + r_2\left(r_{13} r_{23} - r_{12}\right) + r_3\left(r_{12} r_{23} - r_{13}\right)}{1 + 2 r_{12} r_{13} r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2}$$

$$\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \frac{r_1\left(r_{13} r_{23} - r_{12}\right) + r_1\left(1 - r_{13}^2\right) + r_3\left(r_{12} r_{13} - r_{23}\right)}{1 + 2 r_{12} r_{13} r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2}$$

$$\beta_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} \frac{r_1\left(r_{12} r_{23} - r_{13}\right) + r_2\left(r_{12} r_{13} - r_{23}\right) + r_3\left(1 - r_{12}^2\right)}{1 + 2 r_{12} r_{13} r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2}$$

QUADRIVARIATE LINEAR REGRESSION

$$\hat{\mathbf{y}} = \alpha + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \beta_4 \mathbf{x}_4$$

$$\alpha = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - \beta_3 \overline{x_3} - \beta_4 \overline{x_4}$$

$$\beta_1 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \Big\{ r_1 \left(1 + 2r_{23} r_{24} r_{34} - r_{23}^2 - r_{24}^2 - r_{34}^2\right) + r_2 \left[r_{12}\left(r_{34}^2 - 1\right) + r_{13}\left(r_{23} - r_{24} r_{34}\right) + r_{14}\left(r_{24} - r_{23} r_{34}\right)\right] + r_3 \left[r_{13}\left(r_{34}^2 - 1\right) + r_{12}\left(r_{23} - r_{24} r_{34}\right) + r_{14}\left(r_{34} - r_{23} r_{24}\right)\right] +$$
$$+ r_4 \left[r_{14}\left(r_{33}^2 - 1\right) + r_{12}\left(r_{24} - r_{23} r_{34}\right) + r_{13}\left(r_{34} - r_{23} r_{24}\right)\right]\Big\} \Big/ \Big\{1 + 2r_{12} r_{13} r_{23} + 2r_{12} r_{14} r_{24} + 2r_{13} r_{14} r_{34} + 2r_{23} r_{24} r_{34} - r_{12}^2 r_{34}^2 - r_{13}^2 r_{24}^2 - r_{14}^2 r_{23}^2 -$$
$$- 2r_{12} r_{13} r_{23} r_{34} - 2r_{12} r_{14} r_{24} r_{34} - 2r_{13} r_{14} r_{23} r_{34} - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2 \Big\}$$

$$\beta_2 = \sqrt{\frac{\sigma_y}{\sigma_{x2}}} \Big\{ r_1 \left[r_{12}\left(r_{34}^2 - 1\right) + r_{13}\left(r_{23} - r_{24} r_{34}\right) + r_{14}\left(r_{24} - r_{23} r_{34}\right)\right] + r_2 \left(1 + 2r_{13} r_{14} r_{34} - r_{13}^2 - r_{14}^2 - r_{34}^2\right) + r_3 \left[r_{23}\left(r_{14}^2 - 1\right) + r_{12}\left(r_{13} - r_{14} r_{34}\right) + r_{24}\left(r_{34} - r_{12} r_{14}\right)\right] +$$
$$+ r_4 \left[r_{24}\left(r_{13}^2 - 1\right) + r_{12}\left(r_{14} - r_{13} r_{34}\right) + r_{23}\left(r_{34} - r_{13} r_{14}\right)\right]\Big\} \Big/ \Big\{1 + 2r_{12} r_{13} r_{23} + 2r_{12} r_{14} r_{24} + 2r_{13} r_{14} r_{34} + 2r_{23} r_{24} r_{34} - r_{12}^2 r_{34}^2 - r_{13}^2 r_{24}^2 - r_{14}^2 r_{23}^2 -$$
$$- 2r_{12} r_{13} r_{23} r_{34} - 2r_{12} r_{14} r_{24} r_{34} - 2r_{13} r_{14} r_{23} r_{34} - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2 \Big\}$$

$$\beta_3 = \sqrt{\frac{\sigma_y}{\sigma_{x3}}} \Big\{ r_1 \left[r_{13}\left(r_{24}^2 - 1\right) + r_{14}\left(r_{34} - r_{23} r_{24}\right) + r_{12}\left(r_{23} - r_{24} r_{34}\right)\right] + r_2 \left[r_{23}\left(r_{14}^2 - 1\right) + r_{12}\left(r_{13} - r_{14} r_{34}\right) + r_{24}\left(r_{34} - r_{12} r_{24}\right)\right] + r_3 \left(1 + 2r_{12} r_{14} r_{24} - r_{12}^2 - r_{14}^2 - r_{24}^2\right) +$$
$$+ r_4 \left[r_{34}\left(r_{12}^2 - 1\right) + r_{23}\left(r_{24} - r_{12} r_{14}\right) + r_{13}\left(r_{14} - r_{12} r_{24}\right)\right]\Big\} \Big/ \Big\{1 + 2r_{12} r_{13} r_{23} + 2r_{12} r_{14} r_{24} + 2r_{13} r_{14} r_{34} + 2r_{23} r_{24} r_{34} - r_{12}^2 r_{34}^2 - r_{13}^2 r_{24}^2 - r_{14}^2 r_{23}^2 -$$
$$- 2r_{12} r_{13} r_{23} r_{34} - 2r_{12} r_{14} r_{24} r_{34} - 2r_{13} r_{14} r_{23} r_{34} - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2 \Big\}$$

$$\beta_4 = \sqrt{\frac{\sigma_y}{\sigma_{x4}}}\{r_1[r_{14}(r_{23}^2 - 1) + r_{12}(r_{24} - r_{23}r_{24}) + r_{13}(r_{34} - r_{23}r_{24})] + r_2[r_{24}(r_{13}^2 - 1) + r_{12}(r_{14} - r_{13}r_{34}) + r_{23}(r_{34} - r_{13}r_{14})] + r_3[r_{34}(r_{12}^2 - 1) + r_{24}(r_{23} - r_{12}r_{13}) + r_{14}(r_{13} - r_{12}r_{23})] +$$

$$+ r_4(1 + 2r_{12}r_{13}r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2)\} / \{1 + 2r_{12}r_{13}r_{23} + 2r_{12}r_{14}r_{24} + 2r_{13}r_{14}r_{34} + 2r_{23}r_{24}r_{34} - r_{12}^2 r_{34}^2 - r_{13}^2 r_{24}^2 - r_{14}^2 r_{23}^2 -$$

$$- 2r_{12}r_{13}r_{23}r_{34} - 2r_{12}r_{14}r_{24}r_{34} - 2r_{13}r_{14}r_{23}r_{34} - r_{13}^2 - r_{14}^2 - r_{23}^2 - r_{24}^2 - r_{34}^2\}$$

APPENDIX 2: EXPRESSIONS FOR REGRESSION COEFFICIENTS IN PLS WITH 2 INDEPENDENT VARIABLES

Legend

$\mathbf{x}_1$, $\mathbf{x}_2$ – vectors of independent (predictor) variables;

$\mathbf{X}$ – matrix of the independent variables, where the first column contains $\mathbf{x}_1$, and the second $\mathbf{x}_2$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \cdots & \cdots \\ x_{m1} & x_{m2} \end{bmatrix}$$

$\mathbf{y}$ – vector of the observed values of the dependent (response) variable:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix}$$

n – number of observations (samples);

y_i – i-th element of vector $\mathbf{y}$;

x_{i1} – i-th element of the first column of matrix $\mathbf{X}$, *i.e.*, i-th element of independent variable $\mathbf{x}_1$;

x_{i2} – i-th element of the second column of matrix $\mathbf{X}$, *i.e.*, i-th element of independent variable $\mathbf{x}_2$;

β_1, β_2 – regression coefficients for $\mathbf{x}_1$ and $\mathbf{x}_2$ in the bivariate PLS models, respectively;

r_1, r_2 – Pearson correlation coefficients for pairs of variables ($\mathbf{x}_1$, $\mathbf{y}$) and ($\mathbf{x}_2$, $\mathbf{y}$), respectively;

r_{12} – Pearson correlation coefficients for the pair of variables ($\mathbf{x}_1$, $\mathbf{x}_2$).

Note 1.

Data **X** and **y** are autoscaled, *i.e.*, scaled to unit variance, and therefore, the following equalities are used:

$$\sum_{i=1}^{n} y_i^2 = 1 \qquad \sum_{i=1}^{n} x_{i1}^2 = 1 \qquad \sum_{i=1}^{n} x_{i2}^2 = 1$$

$$\sum_{i=1}^{n} x_{i1} y_i = r_1 \qquad \sum_{i=1}^{n} x_{i2} y_i = r_2 \qquad \sum_{i=1}^{n} x_{i1} x_{i2} = r_{12}$$

Note.

Expressions for regression coefficients were derived for the PLS model with one latent variable. The PLS algorithm of Geladi and Kowalski was used to derive these expressions:

Geladi, P.; Kowalski, B. R. Partial least-squares: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1-17.

Algorithm step 1 - Calculation of **w**:

$$\mathbf{w} = \left(\mathbf{y}^{\mathrm{T}}\mathbf{y}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} = 1 \quad \text{and} \quad \mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \quad \text{give} \quad \mathbf{w} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

Algorithm step 2 - Calculation of $\mathbf{w}_1$, the first column of matrix **W**:

$$\mathbf{w}_1 = \left(\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)^{-1/2}\mathbf{w}$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{w} = r_1^2 + r_2^2 \quad \text{and} \quad \mathbf{w} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \quad \text{give} \quad \mathbf{w}_1 = \begin{bmatrix} w_{11} \\ w_{21} \end{bmatrix} = \frac{1}{\sqrt{r_1^2 + r_2^2}} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

Algorithm step 3 - Calcultion of $\mathbf{t}_1$, the first column of matrix $\mathbf{T}$:

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \dots & \dots \\ x_{m1} & x_{m2} \end{bmatrix} \quad \text{and} \quad \mathbf{w}_1 = \frac{1}{\sqrt{r_1^2 + r_2^2}} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \quad \text{give} \quad \mathbf{t}_1 = \begin{bmatrix} t_{11} \\ t_{21} \\ \dots \\ t_{m1} \end{bmatrix} = \frac{1}{\sqrt{r_1^2 + r_2^2}} \begin{bmatrix} x_{11}r_1 + x_{12}r_2 \\ x_{21}r_1 + x_{22}r_2 \\ \dots \\ x_{m1}r_1 + x_{m2}r_2 \end{bmatrix}$$

Algorithm step 4 - Calculation of q_1, the first element of line vector $\mathbf{q}$:

$$q_1 = \left(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\right)^{-1}\mathbf{t}_1^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1 = \frac{(x_{11}r_1 + x_{12}r_2)^2 + (x_{21}r_1 + x_{22}r_2)^2 + \dots + (x_{m1}r_1 + x_{m2}r_2)^2}{r_1^2 + r_2^2} = \frac{r_1^2 + r_2^2 + 2r_1r_2r_{12}}{r_1^2 + r_2^2} \quad \text{and}$$

$$\mathbf{t}_1^{\mathrm{T}}\mathbf{y} = \frac{y_1(x_{11}r_1 + x_{12}r_2) + y_2(x_{21}r_1 + x_{22}r_2) + \dots + y_m(x_{m1}r_1 + x_{m2}r_2)}{\sqrt{r_1^2 + r_2^2}} = \sqrt{r_1^2 + r_2^2} \quad \text{give}$$

$$q_1 = \sqrt{r_1^2 + r_2^2}\,\frac{r_1^2 + r_2^2}{r_1^2 + r_2^2 + 2r_1r_2r_{12}}$$

Algorithm step 5 - Calculation of $\mathbf{l}_1$, the first column of matrix $\mathbf{L}$:

$$\mathbf{l}_1 = \left(\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{t}_1$$

$$\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1 = \frac{r_1^2 + r_2^2 + 2r_1r_2r_{12}}{r_1^2 + r_2^2} \quad \text{and} \quad \mathbf{X}^{\mathrm{T}}\mathbf{t}_1 = \frac{1}{\sqrt{r_1^2 + r_2^2}} \begin{bmatrix} x_{11}(x_{11}r_1 + x_{12}r_2) + x_{21}(x_{21}r_1 + x_{22}r_2) + \dots + x_{m1}(x_{m1}r_1 + x_{m2}r_2) \\ x_{12}(x_{11}r_1 + x_{12}r_2) + x_{22}(x_{21}r_1 + x_{22}r_2) + \dots + x_{m2}(x_{m1}r_1 + x_{m2}r_2) \end{bmatrix} = \frac{1}{\sqrt{r_1^2 + r_2^2}} \begin{bmatrix} r_1 + r_2r_{12} \\ r_1r_{12} + r_2 \end{bmatrix} \quad \text{give}$$

$$\mathbf{l}_1 = \begin{bmatrix} l_{11} \\ l_{21} \end{bmatrix} = \sqrt{r_1^2 + r_2^2} \begin{bmatrix} \dfrac{r_1 + r_2 r_{12}}{r_1^2 + r_2^2 + 2 r_1 r_2 r_{12}} \\ \\ \dfrac{r_1 r_{12} + r_2}{r_1^2 + r_2^2 + 2 r_1 r_2 r_{12}} \end{bmatrix}$$

Algorithm step 6 - Calculation of the regression vector β:

$$\boldsymbol{\beta} = \mathbf{w}_1 \left(\mathbf{l}_1^{\mathrm{T}} \mathbf{w}_1 \right)^{-1} q_1$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \frac{q_1}{l_{11} w_{11} + l_{21} w_{21}} \begin{bmatrix} w_{11} \\ w_{21} \end{bmatrix} = \begin{bmatrix} \dfrac{r_1 \left(r_1^2 + r_2^2 \right)}{r_1^2 + r_2^2 + 2 r_1 r_2 r_{12}} \\ \\ \dfrac{r_2 \left(r_1^2 + r_2^2 \right)}{r_1^2 + r_2^2 + 2 r_1 r_2 r_{12}} \end{bmatrix}$$

$$\beta_1 = \frac{r_1 \left(r_1^2 + r_2^2 \right)}{r_1 + r_2 + r_1 r_2 r_{12}^2} \qquad \beta_2 = \frac{r_2 \left(r_1^2 + r_2^2 \right)}{r_1 + r_2 + r_1 r_2 r_{12}^2}$$

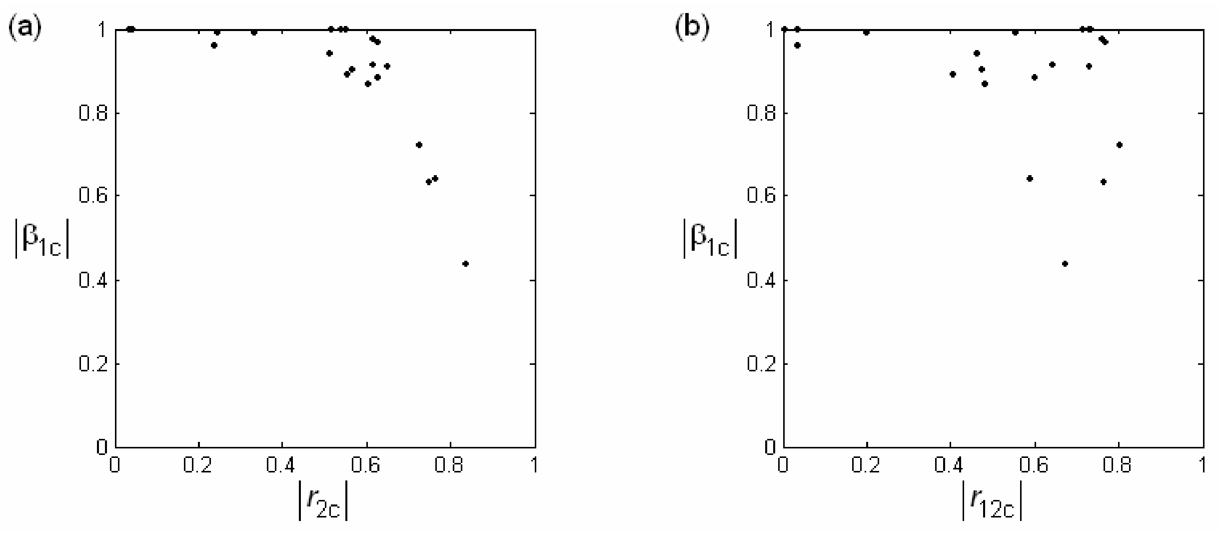

Figure S15. Descriptors' stability in terms of $|\beta_{1c}|$ plotted against a) $|r_{2c}|$ and b) $|r_{12c}|$ for all bivariate regressions generated from the complete descriptors pool for the dataset 48. Descriptor 1 is Human Liver from the final model for prediction of human toxicity HAP (Reference No. 44 associated to Table S1). Index 1 stand for the descriptor 1, index 2 for other descriptors, and 12 for intercorrelation between descriptors 1 and 2.

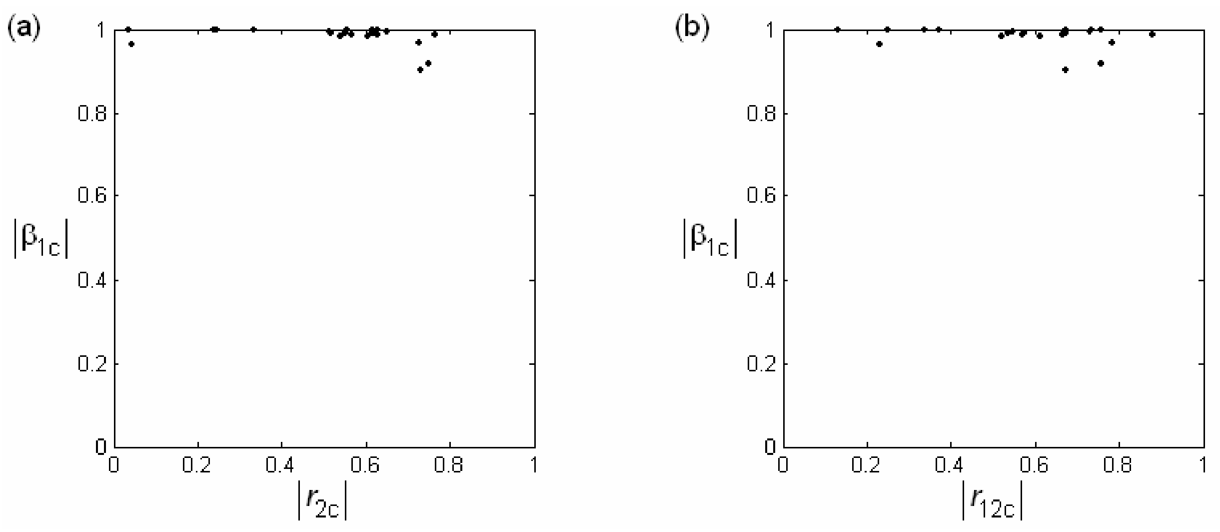

Figure S16. Descriptors' stability in terms of $|\beta_{1c}|$ plotted against a) $|r_{2c}|$ and b) $|r_{12c}|$ for all bivariate regressions generated from the complete descriptors pool for the dataset 48. Descriptors 1 is LUMO from the final model for prediction of human toxicity HAP (Reference No. 44 associated to Table S1). Index 1 stand for the descriptor 1, index 2 for other descriptors, and 12 for intercorrelation between descriptors 1 and 2.

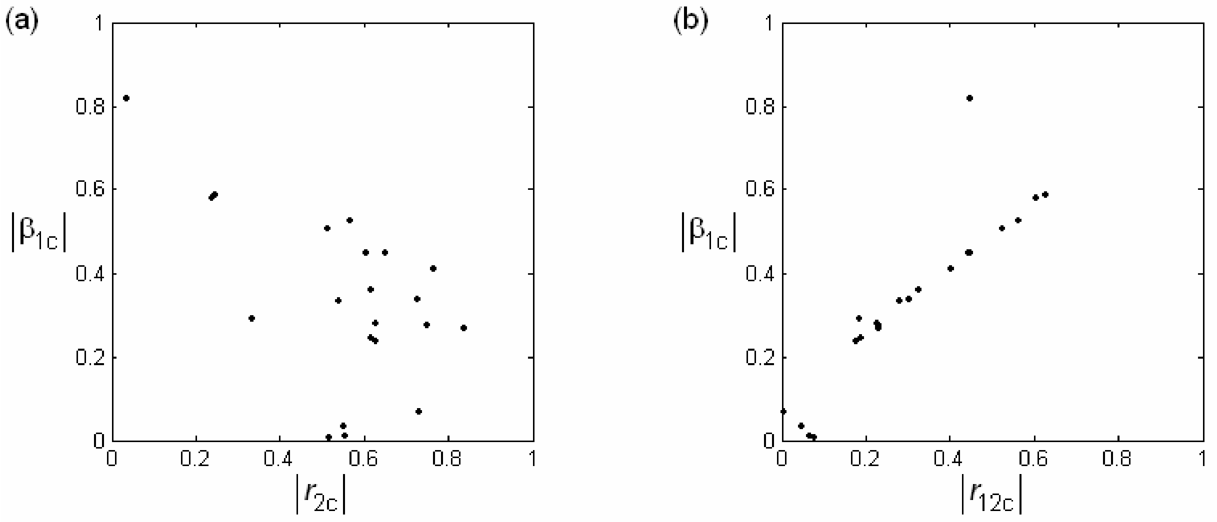

Figure S17. Descriptors' stability in terms of $|\beta_{1c}|$ plotted against a) $|r_{2c}|$ and b) $|r_{12c}|$ for all bivariate regressions generated from the complete descriptors pool for the dataset 48. Descriptors 1 is N_O from the final model for prediction of human toxicity HAP (Reference No. 44 associated to Table S1). Index 1 stand for the descriptor 1, index 2 for other descriptors, and 12 for intercorrelation between descriptors 1 and 2.

Table S5. Parameters for assessment of PLS overfitting for eight datasets[a].

Parameter[b]	Dataset	N_{LV} =1	N_{LV} =2	N_{LV} =3	N_{LV} =4	N_{LV} =5	N_{LV} =6	N_{LV} =7	N_{LV} =8	N_{LV} =9
$\overline{F_3}$	2	0.4686	0.3883	0.2874	0.2315	**0.1844**	0.1843	0.1002	0.0726	-
	3	0.5582	**0.4932**	0.4932	0.3102	0.2261	0.1960	0.1403	0.1392	-
	5	0.5516	**0.4496**	0.1772	0.0961	0.0676	0.0584	0.0729	0.0878	-
	8	0.5090	0.5104	**0.5026**	0.4561	0.4323	-	-	-	-
	11	**0.5495**	0.5409	0.5339	0.5337	0.5337	-	-	-	-
	14	**0.6985**	0.6365	0.6230	-	-	-	-	-	-
	15	0.4673	0.2991	**0.3093**	0.3013	0.2869	0.2653	0.2626	0.0186	0.0390
	18	0.6081	0.5063	**0.2833**	-	-	-	-	-	-
w_3	2	0.0000	0.0000	0.0019	0.1339	**0.4137**	0.4149	0.3421	0.2896	-
	3	0.0000	**0.0000**	0.0000	0.0108	0.1621	0.2866	0.6168	0.6033	-
	5	0.0000	**0.0000**	0.1454	0.2756	0.2743	0.3027	0.3557	0.5390	-
	8	0.0000	0.0000	**0.0000**	0.0000	0.0000	-	-	-	-
	11	**0.0000**	0.0000	0.0000	0.0000	0.0000	-	-	-	-
	14	**0.0000**	0.0000	0.0000	-	-	-	-	-	-
	15	0.0000	0.0021	**0.0001**	0.0060	0.0081	0.0201	0.1344	0.3885	0.5235
	18	0.0000	0.0004	**0.0084**	-	-	-	-	-	-
N_3	2	0	0	2	2	**2**	2	3	3	-
	3	0	**0**	0	3	3	2	2	2	-
	5	0	**0**	3	3	4	4	4	3	-
	8	0	0	**0**	0	0	-	-	-	-
	11	**0**	0	0	0	0	-	-	-	-
	14	**0**	0	0	-	-	-	-	-	-
	15	0	2	**1**	1	2	2	2	3	3
	18	0	1	**1**	-	-	-	-	-	-

[a]Complete datasets which were used in the literature to build PLS regression models, defined in Table S1 under numbers 2, 3, 5, 8, 11, 14, 15 and 18.

[b]Three parameters calculated as functions of the number of latent variables (N_{LV}): $\overline{F_3}$ - average F_3-function for a model, defined as the sum of F_3-function values for all descriptors in a model, divided by the number of these descriptors (m); w_3 - contribution of the sign-changed coefficients, defined as the sum of squares of regression coefficients β_c for descriptors which show the sign change (*i.e.*, their respective F_3-function values are negative); N_3 - number of these descriptors with the sign change problem. The sign change problem was considered as existing only with respect to the sign of $\overline{F_3}$. Values typed in bold are for the numbers of latent variables of the final PLS models in original publications (see references after Table S1).

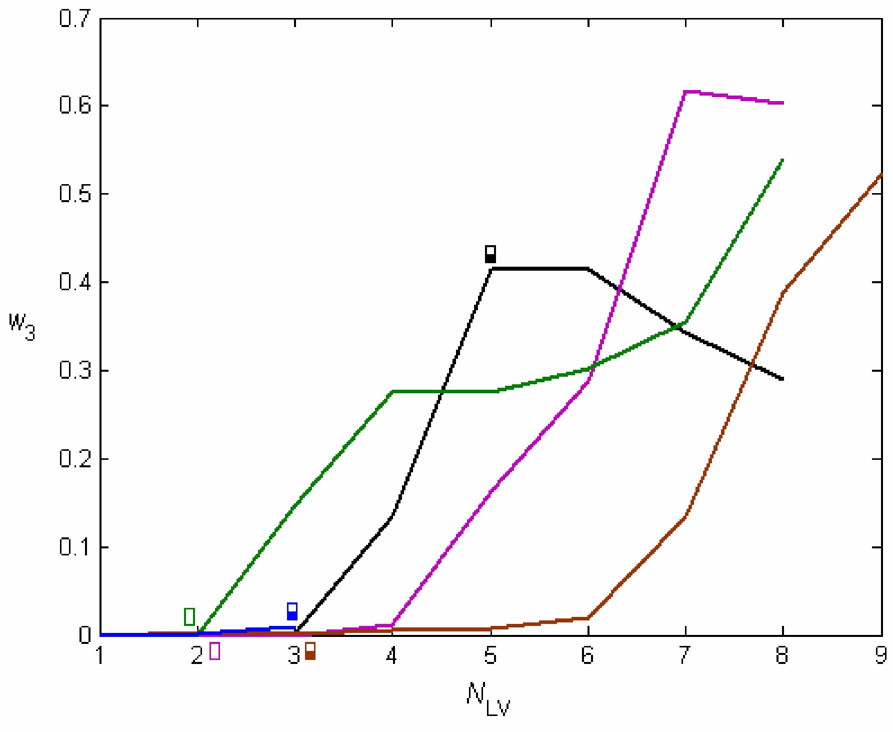

Figure S18. The contribution of the sign-changed coefficients w_3 as a function of the number of latent variables (N_{LV}) for the eight PLS datasets colored differently (2 – black, 3 – magenta, 5 – green, 8 – red, 11 – cyan, 14 – gray, 15 – brown and 18 – blue). N_{LV} for the published PLS models (see Table 1) is marked with small rectangular boxes, which are semi-solid if the sign change has been detected.

Comments on Figure S18:

Five out of eight datasets (2, 3, 5, 15 and 18) have non-zero contribution of the sign-changed coefficients w_3, *i.e.*, the sign change problem has been identified. When this parameter is plotted as a function of the number of latent varibles (N_{LV}), its increasing trend is visible, even a very small increase for dataset 18 can noticed. The extent of this increase can be even up to values around 0.6, meaning that the contribution of the sign-changed regression coefficients is around 60%, what is an obvious consequence of overfitting.

Table S6. Parameters for assessment of the sign change problem for all 50 datasets.

Dataset[a]	m[b]	$\overline{F_3}$ [c]	$\overline{F_4}$ [d]	w_3[e]	w_4[f]	N_3[g]	N_4[h]	w_{m3}[i]	w_{m4}[j]
1	5	0.2006	0.2094	0.0026	0.0177	2	2	0.4000	0.4000
2	8	0.1844	0.1868	0.4137	0.4097	2	2	0.2500	0.2500
3	8	0.4932	0.4823	0	0	0	0	0	0
4	3	0.4366	0.4389	0	0	0	0	0	0
5	8	0.4496	0.4652	0	0	0	0	0	0
6	3	0.5187	0.5225	0	0	0	0	0	0
7	8	0.1010	0.1033	0.5717	0.5654	2	2	0.2500	0.2500
8	6	0.5026	0.5069	0	0	0	0	0	0
9	4	0.1437	0.1367	0.2322	0.2344	2	2	0.5000	0.5000
10	5	0.2490	0.2488	0.0243	0.0227	1	1	0.2000	0.2000
11	5	0.5495	0.5499	0	0	0	0	0	0
12	5	0.1049	0.0999	0.3826	0.4070	2	2	0.4000	0.4000
13	4	0.3292	0.3333	0.0196	0.0673	0	0	0	0
14	3	0.6985	0.6999	0	0	0	0	0	0
15	9	0.3093	0.2803	0.0001	0.0445	1	2	0.1111	0.2222
16	3	0.1719	0.1620	0.1981	0.2380	1	1	0.3333	0.3333
17	3	0.6557	0.6525	0	0	0	0	0	0
18	3	0.2833	0.6119	0.0084	0	1	0	0.3333	0
19	7	0.1316	0.1111	0.4918	0.5138	1	2	0.1429	0.2857
20	6	0.1587	0.1610	0.3985	0.3842	1	1	0.1667	0.1667
21	5	0.4781	0.4818	0	0	0	0	0	0
22	2	0.4603	0.4566	0	0	0	0	0	0
23	6	0.1160	0.1197	0.1915	0.2783	2	2	0.3333	0.3333
24	4	0.3703	0.3703	0.1058	0.0982	1	1	0.2500	0.2500
25	4	0.2576	0.2593	0.2333	0.2327	1	1	0.2500	0.2500
26	3	0.4529	0.4776	0	0	0	0	0	0
27	4	0.3640	0.3624	0	0	0	0	0	0
28	3	0.1820	0.1879	0.0635	0.0567	2	2	0.6667	0.6667
29	10	0.0644	0.0652	0.6587	0.6622	3	3	0.3000	0.3000
30	3	0.1694	0.1592	0.1300	0.1589	2	2	0.6667	0.6667
31	3	0.2814	0.2762	0.1989	0.1923	1	1	0.3333	0.3333
32	6	0.0583	0.3876	0.3111	0.0147	3	1	0.5000	0.1667
33	6	0.2283	0.2430	0.3222	0.2380	1	1	0.1667	0.1667
34	3	0.3804	0.3775	0	0	0	0	0	0
35	3	0.3629	0.3683	0.0178	0.0164	1	1	0.3333	0.3333
36	5	0.1054	0.1019	0.1855	0.1638	2	2	0.4000	0.4000
37	4	0.1857	0.1898	0.1485	0.1440	2	2	0.5000	0.5000
38	6	0.1564	0.1702	0.4907	0.4577	1	1	0.1667	0.1667
39	6	0.1483	0.1764	0.5323	0.4480	1	1	0.1667	0.1667
40	5	0.2979	0.2617	0.0324	0.0142	1	1	0.2000	0.2000
41	4	0.3597	0.3624	0	0	0	0	0	0
42	2	0.5540	0.5584	0	0	0	0	0	0
43	4	0.3606	0.3578	0	0	0	0	0	0
44	2	0.6400	0.6421	0	0	0	0	0	0
45	2	0.7027	0.6985	0	0	0	0	0	0
46	3	0.1830	0.1730	0.4695	0.4800	1	1	0.3333	0.3333
47	5	0.4275	0.4260	0	0	0	0	0	0
48	3	0.4866	0.4840	0	0	0	0	0	0
49	4	0.1088	0.0983	0.4151	0.4138	1	1	0.2500	0.2500
50	2	0.7331	0.7440	0	0	0	0	0	0

[a]Datasets numbering as in Table S1.

[b]The total number of descriptors in dataset (or model) is marked with m.

[c] $\overline{F_3}$ - average F_3-function for a model, defined as the sum of F_3-function values for all descriptors in a model, divided by the number of these descriptors (m).

[d] $\overline{F_4}$ - average F_4-function for a model, defined as the sum of F_4-function values for all descriptors in a model, divided by the number of these descriptors (m).
[e] w_3 - contribution of the sign-changed coefficients, defined as the sum of squares of regression coefficients β_c for descriptors which show the sign change (*i.e.*, their respective F_3-function values are negative).
[f] w_4 - contribution of the sign-changed coefficients, defined as the sum of squares of regression coefficients β_t for descriptors which show the sign change (*i.e.*, their respective F_4-function values are negative).
[g] N_3 - number of descriptors with the sign change problem. The sign change problem was considered as existing only with respect to the sign of $\overline{F_3}$.
[h] N_4 - number of descriptors with the sign change problem. The sign change problem was considered as existing only with respect to the sign of $\overline{F_4}$.
[i] w_{m3} - percentage of the sign-changed descriptors defined as $w_{m3} = N_3/m$.
[j] w_{m4} - percentage of the sign-changed descriptors defined as $w_{m4} = N_4/m$.

Comments on Table S6:

When parameters are plotted against each other, various trends can be observed as illustrated in the following Figures S19 - S21.

1) $\overline{F_3}$ and $\overline{F_4}$ always decrease with m, meaning that the risk of the sign change problem increases for models with many descriptors.

2) The main risk region of the sign change problem is 0.2 - 0.4 in terms of values of $\overline{F_3}$ and $\overline{F_4}$.

3) The higher the values of the w parameters, the more pronounced is the sign change problem, in some cases reaching values of almost 70% of the contribution of the sign-changed regression coefficients (w_3 and w_4) and the same percentage of the number of descriptors with the sign change problem (w_{m3} and w_{m4}).

4) All these trends do not distinguish MLR from PLS models, and also not QSAR/QSAR-like models from QSPR/QSPR-like models.

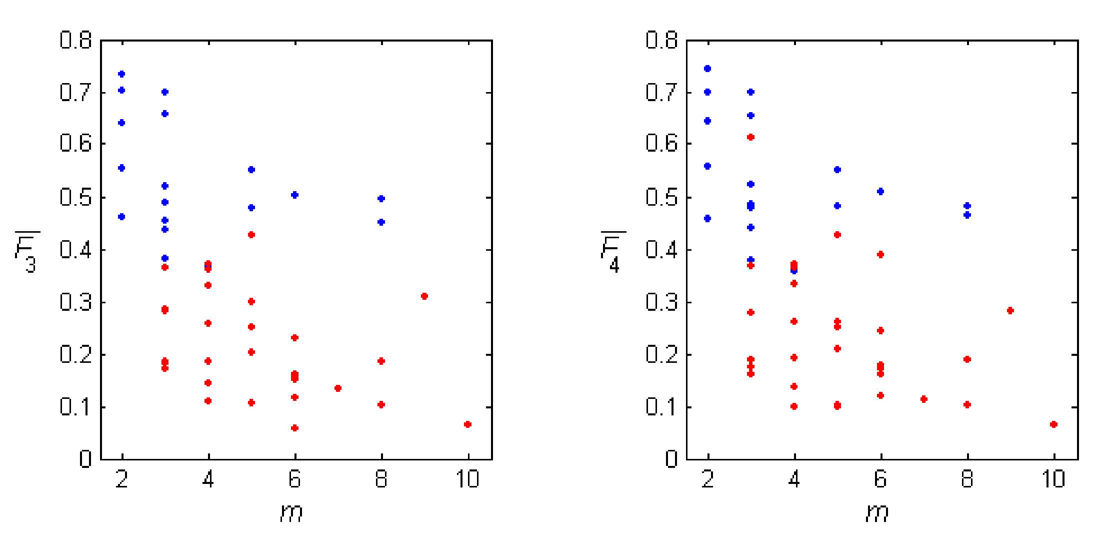

Figure S19. Scatterplot of the number of descriptors in a model (m) versus the average F_3-function ($\overline{F_3}$) for the complete dataset (left) or the average F_4-function ($\overline{F_4}$) for the training set after data split (right). The 50 models are split into those with (red) and without (blue) the sign change problem, according to classifications in Table S2.

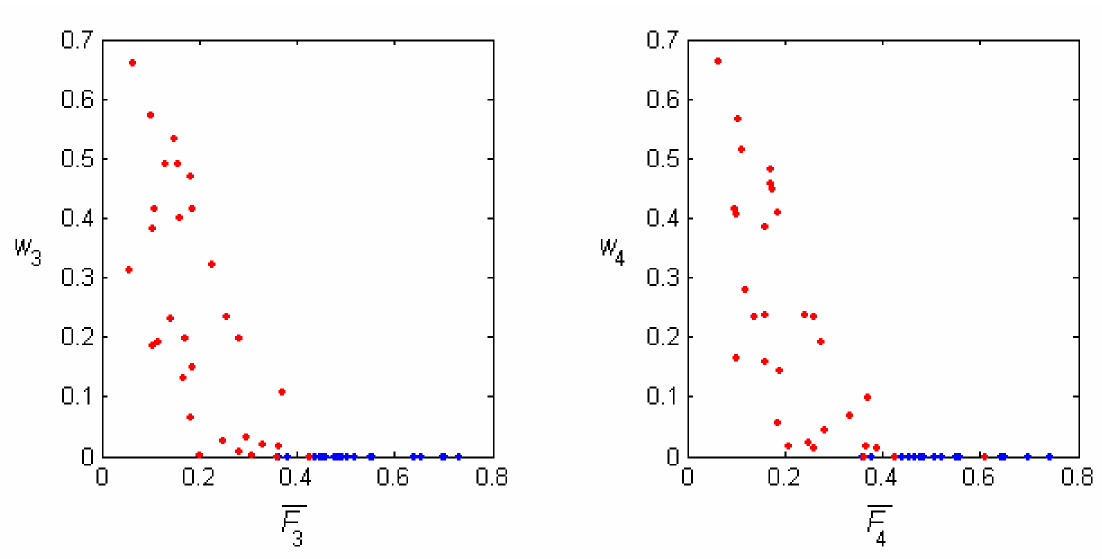

Figure S20. The average F_3-function ($\overline{F_3}$) for the complete dataset (left) or the average F_4-function ($\overline{F_4}$) for the training set after data split (right) plotted against the respective contributions of the sign-changed coefficients w_3 and w_4. The 50 models are split into those with (red) and without (blue) the sign change problem, according to classifications in Table S2.

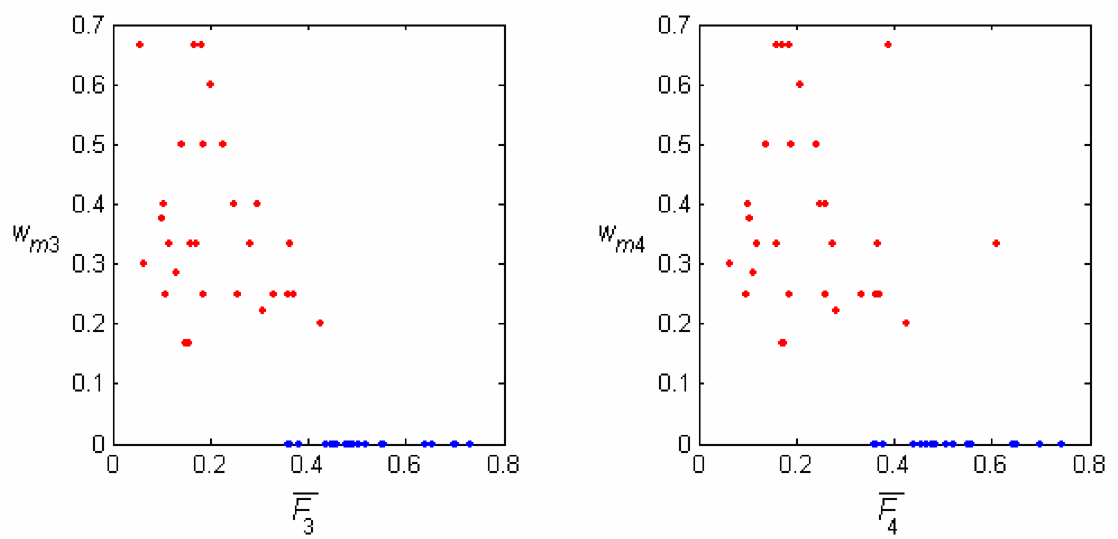

Figure S21. The average F_3-function ($\overline{F_3}$) for the complete dataset (left) or the average F_4-function ($\overline{F_4}$) for the training set after data split (right) plotted against the respective percentage of the sign-changed descriptors w_{m3} and w_{m4}. The 50 models are split into those with (red) and without (blue) the sign change problem, according to classifications in Table S2.