J. Braz. Chem. Soc., Vol. 20, No. 4, S1-S37, 2009.
Printed in Brazil - ©2009 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Basic Validation Procedures for Regression Models in QSAR and QSPR Studies: Theory and Applications

Rudolf Kiralj, Márcia M. C. Ferreira*

Laboratory of Theoretical and Applied Chemometrics, Institute of Chemistry, State University of Campinas, P.O. Box 6154, 13084-971 Campinas-SP, Brazil

DATA SETS 1-4

Table T1. The QSAR data set 1 for phenols toxicity to ciliate *T. pyriformis*.[a,b,c]

No.	Log*Kow*	p*Ka*	E_{LUMO}/eV	E_{HOMO}/eV	N_{hdon}	–log[IGC_{50}/molL^{-1}]
			Training set			
1	0.16	8.45	0.25	–9.16	3	–1.26
3	3.80	11.39	0.46	–8.76	1	1.30
6	3.58	6.75	–0.56	–9.49	1	2.37
8	2.92	10.63	0.36	–8.84	1	0.28
10	4.37	5.72	–0.66	–9.31	2	1.06
11	3.37	6.35	–0.50	–9.39	1	1.41
13	2.97	10.88	0.42	–8.70	1	0.28
15	0.92	8.96	0.53	–8.65	1	–0.52
17	2.96	7.58	–0.29	–9.31	1	1.13
19	2.47	10.22	0.38	–8.85	1	0.08
20	5.63	12.55	0.50	–8.62	1	1.80
21	2.63	6.78	–0.26	–9.37	1	0.74
25	2.55	10.28	0.36	–9.04	1	0.33
29	2.65	8.54	0.06	–9.07	1	0.39
31	1.60	6.98	–0.43	–9.58	1	0.03
32	1.72	8.73	0.02	–9.28	1	0.19
35	2.86	10.37	0.35	–8.76	1	0.71
38	1.98	9.67	–0.28	–9.33	1	0.55
41	3.58	9.67	–0.51	–9.03	1	1.42
43	0.44	9.92	0.20	–9.24	2	–0.95
44	1.56	9.92	–0.54	–9.42	2	–0.08
45	2.90	10.40	0.41	–8.99	1	0.80
47	2.58	9.88	–0.05	–8.50	1	0.75
48	3.30	11.10	0.42	–8.97	1	1.30
49	3.17	8.63	–0.20	–9.26	1	1.75
50	2.42	10.32	0.43	–8.80	1	0.12
51	3.42	6.20	–0.93	–9.66	1	1.64
56	3.87	6.37	–0.90	–9.72	1	2.34
58	2.47	10.15	0.39	–8.98	1	0.11
60	0.49	9.92	0.22	–8.73	2	-0.16
61	2.72	8.96	–0.26	–9.25	1	1.13
63	2.49	9.10	0.04	–9.30	1	0.87
65	1.80	7.80	–0.44	–9.09	1	0.02
69	1.69	9.94	0.32	–8.36	1	–0.30

*e-mail: marcia@iqm.unicamp.br

Table T1. continuation

No.	Log*Kow*	p*Ka*	E_{LUMO}/eV	E_{HOMO}/eV	N_{hdon}	–log[IGC_{50}/molL^{-1}]
Training set						
71	1.46	9.19	–0.46	–9.40	1	–0.38
77	2.90	8.88	–0.04	–9.34	1	1.12
78	2.90	10.10	0.42	–9.02	1	0.61
79	1.57	9.65	0.36	–8.88	1	–0.33
80	2.40	10.23	0.25	–8.89	1	0.42
84	3.34	10.60	0.24	–8.61	1	1.04
85	3.63	10.00	0.10	–8.99	1	1.17
88	3.63	9.70	0.11	–9.06	1	1.27
89	3.61	8.20	–0.22	–9.23	1	1.28
94	4.41	10.03	0.14	–8.93	1	1.85
95	2.98	9.55	0.13	–9.04	1	0.80
100	1.92	9.89	0.07	–9.09	1	0.02
101	4.22	10.70	0.35	–8.60	1	1.64
102	3.45	9.63	0.30	–8.80	2	1.80
106	1.95	10.05	–0.35	–9.33	1	0.19
107	0.29	9.90	0.17	–9.03	2	–0.70
109	0.28	9.27	0.19	–8.87	2	–0.97
110	1.46	8.05	–0.38	–9.43	1	–0.30
111	0.90	9.52	0.06	–9.30	1	–0.38
112	0.75	4.49	–0.21	–9.48	2	–1.50
114	1.98	8.85	–0.36	–9.41	1	0.05
115	2.90	9.20	0.02	–9.24	1	0.85
118	1.57	10.20	0.31	–8.65	1	–0.14
119	3.20	9.55	–0.10	–8.68	1	1.39
120	3.03	10.30	0.43	–8.91	1	0.64
121	3.43	10.30	0.45	–8.91	1	0.98
125	1.92	6.06	–0.70	–9.33	1	0.62
126	2.17	9.79	–0.64	–9.39	1	0.04
129	2.88	8.68	–0.34	–9.80	1	0.62
130	2.51	8.92	–0.38	–9.50	1	0.57
133	1.28	8.89	–0.49	–9.11	1	–0.14
135	1.99	9.21	–0.48	–9.45	1	–0.05
139	6.20	10.40	0.43	–8.92	1	2.47
141	1.97	10.26	0.40	–8.96	1	–0.30
142	1.97	10.26	0.43	–8.88	1	–0.18
143	0.80	9.44	0.28	–8.98	2	–0.65
144	1.81	8.34	–0.43	–9.50	1	0.42
147	1.28	8.36	–0.19	–9.51	2	–0.24
148	0.88	8.78	–0.46	–9.66	3	0.38
150	2.19	2.98	–0.46	–9.51	2	–0.51
151	1.28	7.40	–0.48	–9.14	1	–0.03
External validation set						
2	2.92	10.48	0.38	–8.81	1	0.36
4	2.84	7.44	–0.25	–9.39	1	1.28
5	2.42	10.50	0.38	–8.93	1	0.12
7	3.58	7.37	–0.51	–9.32	1	2.10
9	3.92	6.80	–0.62	–9.50	1	2.03
12	3.31	7.87	–0.30	–9.33	1	1.40
14	2.96	7.97	–0.19	–9.23	1	1.04
16	1.95	8.58	–0.32	–9.29	1	0.60
18	2.47	10.58	0.44	–8.77	1	0.07
22	2.80	6.75	–0.57	–9.38	1	0.80
23	1.75	7.51	–0.32	–9.46	1	0.47

Table T1. continuation

No.	Log*Kow*	p*Ka*	E_{LUMO}/eV	E_{HOMO}/eV	N_{hdon}	$-\log[IGC_{50}/molL^{-1}]$
External validation set						
24	1.10	9.92	0.39	–8.61	1	–0.60
26	2.85	8.67	0.02	–9.05	1	0.60
27	2.36	9.34	–0.01	–9.24	1	0.33
28	3.10	8.85	0.09	–8.89	1	0.69
30	2.16	8.55	0.07	–9.19	1	0.18
33	1.85	10.11	0.42	–8.73	1	–0.36
34	2.50	10.20	0.40	–8.99	1	0.16
36	2.41	10.23	–0.32	–9.19	1	0.31
37	1.92	9.90	–0.33	–9.32	1	0.08
39	1.32	9.99	0.40	–8.79	1	–0.51
40	3.09	9.55	–0.04	–8.74	1	1.09
42	2.87	10.25	0.41	–8.76	1	0.93
46	3.29	8.18	–0.28	–9.54	1	1.57
52	3.07	6.20	–0.90	–9.58	1	1.55
53	1.60	9.35	0.42	–8.94	1	–0.09
54	5.13	10.32	0.47	–8.93	1	1.64
55	2.64	9.03	–0.05	–9.34	1	1.15
57	2.50	8.85	0.03	–9.22	1	0.76
59	1.60	8.61	–0.50	–9.59	1	–0.06
62	2.50	10.07	0.40	–9.04	1	0.23
64	1.92	9.29	0.03	–9.37	1	0.38
66	0.29	9.67	0.33	–8.68	2	–0.99
67	1.44	9.00	–0.53	–9.43	1	0.09
68	1.56	4.08	–0.57	–9.52	2	–0.81
70	0.44	9.83	0.12	–9.26	2	–1.04
72	3.23	9.63	–0.15	–8.95	1	1.35
73	3.30	10.10	0.43	–9.01	1	0.73
74	5.16	9.92	0.47	–8.85	1	2.10
75	3.30	10.30	0.47	–8.91	1	0.91
76	2.08	7.28	–0.25	–9.03	2	0.97
81	3.52	6.40	–0.52	–9.44	1	1.78
82	2.64	9.34	0.02	–9.20	1	0.68
83	3.16	10.60	0.34	–8.61	1	0.70
86	2.98	9.67	0.12	–9.01	1	0.70
87	3.48	9.65	0.14	–8.99	1	1.20
90	3.51	9.54	0.14	–9.04	1	1.08
91	2.49	9.43	0.10	–9.12	1	0.55
92	1.58	8.11	0.01	–8.98	2	0.13
93	1.60	7.95	–0.40	–9.57	1	0.52
96	2.10	10.50	0.33	–8.61	1	0.01
97	2.50	10.00	0.43	–8.92	1	0.21
98	4.75	10.70	0.35	–8.59	1	2.03
99	1.27	9.85	–0.39	–9.00	1	–0.12
103	1.42	7.96	–0.44	–9.10	1	–0.03
104	0.47	9.92	0.28	–8.78	2	–0.18
105	1.44	7.62	–0.44	–9.49	1	0.27
108	0.33	9.23	–0.25	–9.44	2	–0.78
113	1.56	4.58	–0.49	–9.60	2	–1.02
116	3.07	8.89	–0.49	–9.40	1	1.02
117	0.52	9.92	0.29	–9.06	2	–0.83
122	2.90	10.30	0.44	–8.92	1	0.47
123	3.83	10.30	0.46	–8.90	1	1.23

Table T1. continuation

No.	Log*Kow*	p*Ka*	E_{LUMO}/eV	E_{HOMO}/eV	N_{hdon}	–log[IGC_{50}/molL^{-1}]
External validation set						
124	1.60	9.34	–0.13	–9.29	2	0.34
127	1.31	9.46	0.26	–8.95	2	–0.39
128	3.42	9.49	0.29	–8.93	2	1.31
131	4.30	12.50	0.44	–8.69	1	1.16
132	2.51	9.09	–0.45	–9.44	1	0.48
134	1.53	9.92	0.04	–8.86	1	–0.23
136	1.97	10.10	0.39	–9.02	1	–0.06
137	1.99	9.05	–0.40	–9.54	1	0.08
138	2.49	9.21	–0.31	–9.35	1	0.62
140	1.65	7.91	–0.45	–9.12	1	0.38
145	3.54	9.92	0.43	–8.90	1	1.29
146	1.48	9.99	0.40	–9.11	1	–0.21
149	1.10	9.92	–0.19	–8.99	2	0.25
152	0.85	9.92	–0.36	–9.57	3	0.18
153	0.99	7.62	–0.50	–8.94	1	0.17

[a]This data set was generated by Aptula *et al.*, *Quant. Struct.-Act. Relat.* **2002**, *21*, 12. It was used to build MLR models by Yao *et al.*, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257. [b]Names of samples and variables are from the original publications. [c]The current data split is made in this work.

Table T2. The QGSAR data set 2 for fungal resistance (*P. digitatum* strains) to demethylation inhibitors.[a,b,c]

No.	Toxicant/ Straind	CYP51-g	CYP51-e	PMR1-t	CYP51-e*Npi	PCR*Npi	PMR1-e*Lpi	CYP51-e*Lpi	PCR*Lpi	-log[EC_{50}/ molL^{-1}]
Training set										
1	T/PD5	1	1	100	1	0.25	1	1	0.25	6.636
2	T/DF1	1	1	100	1	0.25	1	1	0.25	6.937
4	T1/LC2	5	100	100	100	0.75	7	100	0.75	5.225
6	T/I1	5	100	100	100	0.75	7	100	0.75	5.093
7	T/DIS03	5	100	1	100	0.75	0	100	0.75	5.895
9	T/DIS96	5	100	1	100	0.75	0	100	0.75	5.858
10	T/PD5	1	1	100	1	0.25	1	1	0.25	6.840
12	T/PD5-7	5	100	100	100	0.75	1	100	0.75	5.335
13	T/PD5-15	5	100	100	100	0.75	1	100	0.75	5.335
15	T/PD5	1	1	100	1	0.25	1	1	0.25	6.948
17	T/DISp21	1	1	1	1	0.25	0	1	0.25	7.694
19	T/LC2	5	100	100	100	0.75	7	100	0.75	5.146
20	T/DIS33	5	100	1	100	0.75	0	100	0.75	6.177
22	T/DIS33-Y8	5	100	100	100	0.75	7	100	0.75	5.179
23	T/DIS33-B0	5	100	10	100	0.75	7	100	0.75	6.007
27	T/DIS33	5	100	1	100	0.75	0	100	0.75	6.283
29	T/PD5	1	1	100	1	0.25	1	1	0.25	6.937
30	T/LC2M	2	50	100	50	0.37	7	50	0.37	5.840
32	T/DIS33	5	100	1	100	0.75	0	100	0.75	6.177
33	T/LC2	5	100	1000	100	0.75	7	100	0.75	5.189
34	T/DISp21	1	1	1	1	0.25	0	1	0.25	7.539
36	T/DIS5-P26	1	1	100	1	0.25	1	1	0.25	6.937
37	F/PD5	1	1	100	3	0.75	2	2	0.50	6.073
39	F/U1	1	1	100	3	0.75	2	2	0.50	6.058
40	F/LC2	5	100	100	300	2.25	14	200	1.50	5.160
42	F/I1	5	100	100	300	2.25	14	200	1.50	4.799
43	F/PD5	1	1	100	3	0.75	2	2	0.50	6.073
45	F/DIS33	5	100	1	300	2.25	0	200	1.50	6.073
46	F/DIS96	5	100	1	300	2.25	0	200	1.50	5.952

Table T2. continuation

No.	Toxicant/ Straind	CYP51-g	CYP51-e	PMR1-t	CYP51-e*Npi	PCR*Npi	PMR1-e*Lpi	CYP51-e*Lpi	PCR*Lpi	-log[EC_{50}/ $molL^{-1}$]
					Training set					
48	F/PD5-21	1	3	100	9	0.75	2	6	0.50	6.122
49	F/PD5-7	5	100	100	300	2.25	2	200	1.50	5.714
51	F/LC2	5	100	100	300	2.25	14	200	1.50	5.241
52	F/PD5	1	1	100	3	0.75	2	2	0.50	6.475
54	F/DISp21	1	1	1	3	0.75	0	2	0.50	6.520
56	F/LC2	5	100	100	300	2.25	14	200	1.50	5.230
57	F/DIS33	5	100	1	300	2.25	0	200	1.50	6.241
59	B/DF1	1	1	100	2	0.50	3	3	0.75	6.081
61	B/LC2	5	100	100	200	1.50	21	300	2.25	5.024
63	B/I1	5	100	100	200	1.50	21	300	2.25	4.546
65	B/DIS33	5	100	1	200	1.50	0	300	2.25	6.298
67	B/LC2	5	100	100	200	1.50	21	300	2.25	5.024
69	B/PD5-21	1	3	100	6	0.50	3	9	0.75	5.984
70	B/PD5-7	5	100	100	200	1.50	3	300	2.25	5.449
72	B/LC2	5	100	100	200	1.50	21	300	2.25	5.113
74	B/DISp12	1	1	1	2	0.50	0	3	0.75	7.324
75	B/DISp21	1	1	1	2	0.50	0	3	0.75	7.324
77	B/LC2	5	100	100	200	1.50	21	300	2.25	5.112
78	B/DIS33	5	100	1	200	1.50	0	300	2.25	6.382
79	P/PD5	1	1	100	3	0.75	3	3	0.75	5.902
80	P/DF1	1	1	100	3	0.75	3	3	0.75	5.664
82	P/PD5	1	1	100	3	0.75	3	3	0.75	6.993
83	P/PD5-21	1	3	100	9	0.75	3	9	0.75	6.023
85	P/PD5-15	5	100	100	300	2.25	3	300	2.25	4.979
87	P/PD5	1	1	100	3	0.75	3	3	0.75	6.993
88	P/DISp12	1	1	1	3	0.75	0	3	0.75	7.023
90	P/ECTp36	1	1	1	3	0.75	0	3	0.75	6.685
					External validation set					
5	T/M1	5	100	100	100	0.75	7	100	0.75	5.432
8	T/DIS33	5	100	1	100	0.75	0	100	0.75	5.971
11	T/PD5-21	1	3	100	3	0.25	1	3	0.25	6.363
14	T/LC2	5	100	100	100	0.75	7	100	0.75	5.141
16	T/DISp12	1	1	1	1	0.25	0	1	0.25	7.694
18	T/ECTp36	1	1	1	1	0.25	0	1	0.25	7.047
21	T/DIS33-Y4	5	100	100	100	0.75	7	100	0.75	5.641
24	T/DIS33-B13	5	100	10	100	0.75	7	100	0.75	6.124
28	T/LC2	5	100	100	100	0.75	7	100	0.75	5.107
31	T/PD5	1	1	100	1	0.25	1	1	0.25	6.937
35	T/DIS5-L22	5	100	100	100	0.75	7	100	0.75	5.202
38	F/DF1	1	1	100	3	0.75	2	2	0.50	6.015
41	F/M1	5	100	100	300	2.25	14	200	1.50	5.093
44	F/DIS07	5	100	1	300	2.25	0	200	1.50	6.002
47	F/PD5	1	1	100	3	0.75	2	2	0.50	6.479
50	F/PD5-15	5	100	100	300	2.25	2	200	1.50	5.542
53	F/DISp12	1	1	1	3	0.75	0	2	0.50	7.219
55	F/ECTp36	1	1	1	3	0.75	0	2	0.50	6.441
58	B/PD5	1	1	100	2	0.50	3	3	0.75	6.186
60	B/U1	1	1	100	2	0.50	3	3	0.75	6.037
62	B/MI	5	100	100	200	1.50	21	300	2.25	5.093
64	B/DIS07	5	100	1	200	1.50	0	300	2.25	6.227
66	B/DIS96	5	100	1	200	1.50	0	300	2.25	6.148
68	B/PD5	1	1	100	2	0.50	3	3	0.75	6.382

Table T2. continuation

No.	Toxicant/ Straind	CYP51-g	CYP51-e	PMR1-t	CYP51-e*Npi	PCR*Npi	PMR1-e*Lpi	CYP51-e*Lpi	PCR*Lpi	-log[EC_{50}/ molL^{-1}]
					External validation set					
71	B/PD5-15	5	100	100	200	1.50	3	300	2.25	5.324
73	B/PD5	1	1	100	2	0.50	3	3	0.75	6.424
76	B/ECTp36	1	1	1	2	0.50	0	3	0.75	6.503
81	P/U1	1	1	100	3	0.75	3	3	0.75	5.595
84	P/PD5-7	5	100	100	300	2.25	3	300	2.25	5.294
89	P/DISp21	1	1	1	3	0.75	0	3	0.75	7.039

[a]This data set was generated and used to build PLS models by Kiralj and Ferreira, *QSAR Comb. Sci.* **2008**, *27*, 289. It was denominated C1 and had six outliers which are already excluded in this table. [b]Names of samples and variables are from the original publication. Abbreviations for toxicants in the sample names: T - triflumizole, F - fenarimol, B - bitertanol and P - pyrifenox. [c]The current data split is from the original publication. [d]A sample is defined as combination toxicant-strain-experiment.

Table T3. The QSPR data set 3 for carbonyl oxygen chemical shift in substituted benzaldehydes.[a,b,c]

No.	E_e/eV	E_{CC}/eV	Δ_{HL}/eV	σ_b/Å	σ_r/Å	D_{CC}/Å	Q_{C2mul}	Q_{Omul}	δ/ppm
1	70.679	122.756	–9.567	0.071	0.003	1.484	–0.201	–0.317	563.2
2	71.061	122.671	–9.284	0.071	0.004	1.483	–0.209	–0.319	561.4
3	72.252	122.749	–9.047	0.071	0.007	1.481	–0.244	–0.322	526.9
4	72.092	122.819	–8.336	0.071	0.008	1.480	–0.240	–0.326	532.8
5	72.150	122.790	–9.012	0.071	0.006	1.481	–0.241	–0.322	545.7
6	65.144	122.617	–9.335	0.072	0.005	1.485	–0.215	–0.313	568.9
7	70.884	122.591	–8.930	0.072	0.003	1.485	–0.204	–0.313	570.1
8	70.633	122.540	–9.315	0.072	0.007	1.486	–0.196	–0.311	570.3
9	70.181	122.431	–9.140	0.072	0.003	1.487	–0.182	–0.305	593.6
10	68.864	122.302	–9.138	0.073	0.004	1.490	–0.143	–0.294	590.1
11	70.690	122.526	–9.298	0.072	0.005	1.486	–0.198	–0.319	575.0
12	74.926	123.365	–8.695	0.067	0.011	1.471	–0.324	–0.360	505.8
13	71.304	122.324	–9.336	0.073	0.006	1.489	–0.217	–0.307	555.0
14	67.203	121.716	–9.367	0.076	0.005	1.499	–0.093	–0.282	576.0
15	69.963	122.470	–9.366	0.072	0.006	1.487	–0.176	–0.309	573.0
16	70.688	122.499	–9.013	0.072	0.003	1.486	–0.198	–0.310	573.0
17	70.367	122.491	–8.962	0.072	0.003	1.486	–0.188	–0.312	569.3
18	69.915	122.457	–9.310	0.072	0.004	1.487	–0.175	–0.309	570.8
19	71.291	122.541	–9.483	0.072	0.003	1.485	–0.216	–0.311	568.4
20	69.495	122.448	–8.905	0.072	0.005	1.486	–0.162	–0.313	555.2
21	71.519	122.396	–9.437	0.073	0.004	1.488	–0.223	–0.303	574.5
22	70.580	122.597	–9.324	0.071	0.007	1.485	–0.195	–0.311	566.0
23	69.529	122.508	–8.894	0.072	0.006	1.486	–0.163	–0.313	562.3
24	74.885	123.200	–8.548	0.068	0.011	1.473	–0.323	–0.353	507.0
25	74.580	123.258	–8.341	0.068	0.011	1.473	–0.314	–0.355	516.2
26	75.864	123.104	–8.915	0.069	0.013	1.476	–0.351	–0.349	522.8
27	73.652	123.179	–8.188	0.068	0.009	1.474	–0.286	–0.355	512.1
28	74.466	123.250	–8.539	0.068	0.011	1.473	–0.310	–0.356	514.7
29	73.572	123.241	–8.215	0.068	0.012	1.473	–0.284	–0.355	511.8
30	74.560	123.263	–8.347	0.068	0.010	1.473	–0.313	–0.355	509.0
31	73.360	123.193	–8.353	0.068	0.011	1.474	–0.279	–0.357	510.0
32	74.192	123.283	–8.292	0.069	0.015	1.473	–0.302	–0.357	513.9
33	74.230	123.166	–8.570	0.069	0.011	1.475	–0.303	–0.351	518.2
34	75.069	123.302	–8.603	0.068	0.011	1.472	–0.327	–0.357	507.0

Table T3. continuation

No.	E_e/eV	E_{CC}/eV	Δ_{HL}/eV	σ_b/Å	σ_t/Å	D_{CC}/Å	Q_{C2mul}	Q_{Omul}	δ/ppm
35	74.719	123.219	–8.217	0.068	0.011	1.474	–0.317	–0.353	515.0
36	74.722	123.230	–8.629	0.068	0.010	1.473	–0.318	–0.354	509.0
37	74.194	123.138	–8.139	0.068	0.010	1.475	–0.303	–0.351	520.0
38	74.517	123.116	–8.155	0.069	0.010	1.475	–0.313	–0.349	512.0
39	75.052	123.188	–8.509	0.068	0.011	1.474	–0.328	–0.351	507.0
40	74.628	123.142	–8.202	0.069	0.012	1.475	–0.315	–0.350	513.0
41	70.289	122.517	–8.677	0.073	0.007	1.485	–0.186	–0.313	550.0
42	70.690	122.507	–9.224	0.072	0.005	1.486	–0.199	–0.315	585.0
43	73.273	122.264	–9.205	0.074	0.005	1.490	–0.275	–0.313	565.0
44	71.796	122.477	–9.004	0.071	0.008	1.486	–0.231	–0.322	545.0
45	70.738	122.356	–8.761	0.074	0.011	1.489	–0.200	–0.307	538.0
46	70.328	122.471	–8.461	0.073	0.008	1.487	–0.188	–0.305	560.0
47	74.358	123.155	–8.050	0.069	0.010	1.475	–0.307	–0.350	518.0
48	74.739	123.148	–8.062	0.069	0.010	1.475	–0.318	–0.348	505.0
49	74.237	123.024	–7.968	0.069	0.010	1.476	–0.304	–0.346	517.0
50	74.405	123.092	–7.888	0.069	0.010	1.476	–0.309	–0.346	513.0

[a]This data set was generated and used to build PLS models by Kiralj and Ferreira, *J. Phys. Chem. A* **2008**, *112*, 6134. [b]Names of samples and variables are from the original publication. [c]The ten samples in the external validation set are **2**, **5**, **7**, **10**, **22**, **26**, **27**, **34**, **41** and **49**. This selection is from the original publication. Table T4. The QSAR data set 4 for mouse cyclooxigenase-2 inhibition by imidazoles.[a,b,c]

Table T4. The QSAR data set 4 for mouse cyclooxigenase-2 inhibition by imidazoles.[a,b,c]

No.	Substituent	ClogP	MgVol/Lmol^{-1}	$B1_{X.2}$/Å	–log[IC_{50}/molL^{-1}]
1	H	3.09	2.37	1.00	6.16
3	3-F	3.24	2.39	1.00	6.62
4	4-F	3.24	2.39	1.00	6.72
5	2-Cl	3.56	2.49	1.80	5.89
6	3-Cl	3.81	2.49	1.00	7.10
8	2-CH_3	3.29	2.51	1.52	5.75
9	3-CH_3	3.59	2.51	1.00	6.22
10	4-CH_3	3.59	2.51	1.00	6.19
11	3-OCH_3	3.10	2.57	1.00	5.62
12	4-OCH_3	3.10	2.57	1.00	5.54
13	3,4-Cl_2	4.40	2.61	1.00	7.40
14	2,4-F_2	3.38	2.40	1.35	6.26
15	3,4-F_2	3.31	2.40	1.00	6.80
16	3-Cl-4-CH_3	4.24	2.63	1.00	6.64
17	2-CH_3-3-F	3.44	2.53	1.52	5.77

[a]This data set was generated and used to build a MLR model by Garg *et al.*, *Chem. Rev.* **2003**, *103*, 703 (Table 17). Two outliers are already excluded. [b]Names of samples and variables are from the original publication. [c]The two samples in the external validation set are **4** and **10**. This selection was made in this work.

ANALYSIS OF DATA SET 1

Table T5. Results for leave-*N*-out for the MLR model on data set 1.

N	Single LNO	Multiple LNO[a]	
	Q^2_{LNO}	$\langle Q^2_{LNO} \rangle$	$\sigma(Q^2_{LNO})$
1	0.7725	0.7725	0
2	0.7815	0.7717	0.0041
3	0.7741	0.7710	0.0123
4	0.7854	0.7732	0.0054
5	0.7841	0.7680	0.0101
6	0.7735	0.7649	0.0184
7	0.7710	0.7694	0.0159
8	0.7321	0.7692	0.0220
9	0.7786	0.7630	0.0255
10	0.7690	0.7721	0.0134
11	0.7718	0.7708	0.0116
12	0.7913	0.7578	0.0233
13	0.7576	0.7593	0.0381
14	0.7840	0.7611	0.0279
15	0.7815	0.7573	0.0217
16	0.7812	0.7676	0.0243
17	0.7818	0.7642	0.0505
18	0.7854	0.7527	0.0577
19	0.7948	0.7566	0.0368
20	0.8057	0.7515	0.0387
21	0.8070	0.7548	0.0262
22	0.7924	0.7599	0.0281
23	0.7699	0.7542	0.0324
24	0.7813	0.7486	0.0459
25	0.7669	0.7594	0.0218
26	0.7523	0.7594	0.0274
27	0.7757	0.7527	0.0268
28	0.7809	0.7486	0.0359
29	0.7837	0.7541	0.0274
30	0.7809	0.7302	0.0750
31	0.7913	0.7386	0.0587
32	0.7963	0.7433	0.0613
33	0.7801	0.7429	0.0589
34	0.7747	0.7503	0.0531
35	0.7791	0.7487	0.0552
36	0.7592	0.7446	0.0295
37	0.7842	0.7469	0.0310
Average	0.7787		
Standard Deviation	0.0140		

[a]Average values of Q^2_{LNO} and respective standard deviations are reported for multiple LNO (ten times reordered data set).

Table T6. Ten bootstrappings with HCA clustering (BH), PCA clustering (BP), and only with random selection (BR) for the MLR model on data set 1.

Bootstrapping	Q^2_{BH}	R^2_{BH}	Q^2_{BP}	R^2_{BP}	Q^2_{BR}	R^2_{BR}
1	0.7642	0.8229	0.8946	0.9141	0.7428	0.8045
2	0.8078	0.8597	0.7753	0.8371	0.7727	0.8163
3	0.7642	0.8200	0.7835	0.8375	0.7817	0.8439
4	0.7960	0.8485	0.7776	0.8442	0.7945	0.8483
5	0.8115	0.8628	0.8095	0.8520	0.7847	0.8315
6	0.8271	0.8675	0.7453	0.8259	0.7747	0.8330
7	0.7905	0.8377	0.7327	0.7994	0.8021	0.8446
8	0.7712	0.8309	0.7631	0.8218	0.7751	0.8307
9	0.7312	0.7981	0.7645	0.8221	0.7727	0.8247
10	0.7884	0.8431	0.8070	0.8523	0.8202	0.8595
Average	0.7852	0.8391	0.7853	0.8406	0.7821	0.8337
Stand. dev.	0.0281	0.0218	0.0453	0.0304	0.0206	0.0162
The model	0.7725	0.8301	0.7725	0.8301	0.7725	0.8301

Table T7. Twenty-five bootstrappings with HCA clustering (BH), PCA clustering (BP), and only with random selection (BR) for the MLR model on data set 1.

Bootstrapping	Q^2_{BH}	R^2_{BH}	Q^2_{BP}	R^2_{BP}	Q^2_{BR}	R^2_{BR}
1	0.7704	0.8253	0.7967	0.8417	0.8031	0.8565
2	0.8108	0.8603	0.7577	0.8466	0.7823	0.8434
3	0.7863	0.8406	0.7584	0.8111	0.7817	0.8295
4	0.7868	0.8482	0.7449	0.8269	0.7842	0.8299
5	0.7703	0.8243	0.7252	0.7924	0.7553	0.8237
6	0.7636	0.8285	0.8197	0.8763	0.7677	0.8178
7	0.7745	0.8342	0.8143	0.8540	0.8000	0.8582
8	0.8052	0.8538	0.7409	0.8009	0.7543	0.8171
9	0.7628	0.8245	0.7223	0.7878	0.7124	0.7801
10	0.7669	0.8303	0.6915	0.7286	0.7390	0.8175
11	0.8067	0.8587	0.8024	0.8501	0.7829	0.8462
12	0.8049	0.8509	0.6866	0.7879	0.7924	0.8401
13	0.7842	0.8447	0.8155	0.8598	0.8109	0.8562
14	0.7490	0.8210	0.8210	0.8640	0.7507	0.8109
15	0.8269	0.8761	0.8054	0.8533	0.7845	0.8439
16	0.8041	0.8457	0.8031	0.8442	0.7676	0.8282
17	0.7512	0.8149	0.8180	0.8661	0.7829	0.8318
18	0.8056	0.8542	0.7945	0.8512	0.8074	0.8585
19	0.7639	0.8226	0.7884	0.8455	0.8323	0.8751
20	0.7334	0.8035	0.7973	0.8602	0.8068	0.8483
21	0.7708	0.8365	0.7076	0.7917	0.8230	0.8726
22	0.8071	0.8515	0.7723	0.8268	0.9205	0.9331
23	0.7920	0.8457	0.7940	0.8448	0.7698	0.8263
24	0.7399	0.8036	0.7563	0.8151	0.7872	0.8426
25	0.8400	0.8808	0.8316	0.8689	0.7877	0.8293
Average	0.7831	0.8392	0.7746	0.8318	0.7875	0.8407
Stand. dev.	0.0271	0.0198	0.0430	0.0348	0.0384	0.0283
The model	0.7725	0.8301	0.7725	0.8301	0.7725	0.8301

Table T8. Ten **y**-randomizations for the MLR model on data set 1.

Randomization	Q^2	R^2
1	–0.0439	0.0912
2	–0.0380	0.1004
3	–0.1212	0.0442
4	–0.1801	0.0255
5	–0.0260	0.1561
6	–0.0847	0.0769
7	–0.1343	0.0458
8	–0.0167	0.1340
9	–0.1084	0.0571
10	–0.1866	0.0087
Maximum	–0.0167	0.1561
Average	–0.0940	0.0740
Standard deviation	0.0623	0.0471
The model	0.7725	0.8301

Table T9. Twenty-five **y**-randomizations for the MLR model on data set 1.

Randomization	Q^2	R^2
1	–0.0545	0.0881
2	–0.1419	0.0415
3	–0.0570	0.1166
4	–0.0697	0.1047
5	–0.0260	0.1561
6	–0.0847	0.0769
7	–0.1343	0.0458
8	–0.0167	0.1340
9	–0.1084	0.0571
11	–0.0326	0.1615
12	–0.0821	0.0825
13	–0.0895	0.0672
14	–0.1808	0.0212
15	–0.1146	0.0823
16	–0.1030	0.0619
17	–0.1217	0.0482
18	–0.1192	0.0584
19	–0.1153	0.0450
20	–0.0876	0.0547
21	–0.1808	0.0330
22	–0.0936	0.0709
23	–0.0998	0.0542
24	–0.1802	0.0111
25	–0.0378	0.0126
Maximum	–0.0167	0.1816
Average	–0.0943	0.0747
Standard deviation	0.0479	0.0455
The model	0.7725	0.8301

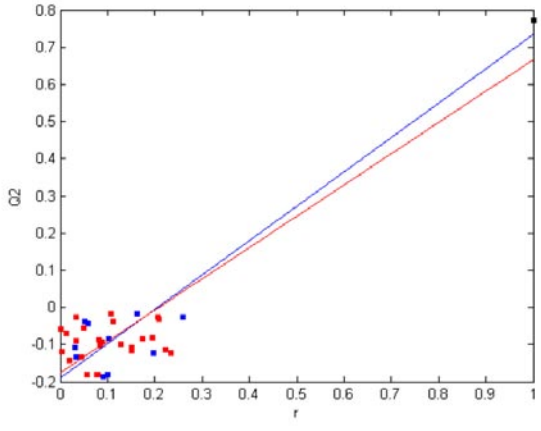

Figure F1. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the MLR model on data set 1: $Q^2 = –0.191 + 0.925\ r$ and $Q^2 = –0.176 + 0.842\ r$, respectively.

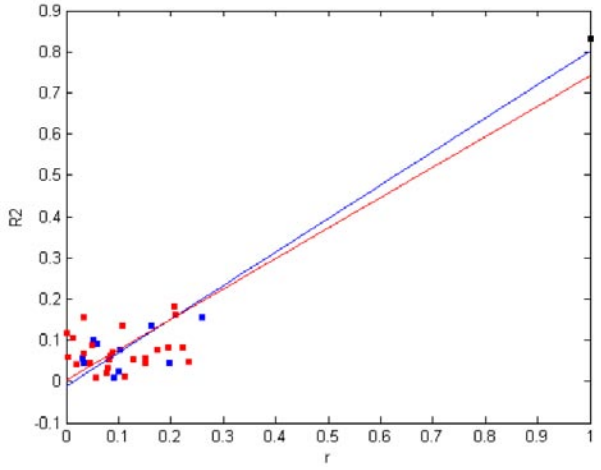

Figure F2. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the MLR model on data set 1: $R^2 = –0.012 + 0.813\ r$ and $R^2 = 0.003 + 0.738\ r$, respectively.

ANALYSIS OF DATA SET 2

Table T10. Results for leave-*N*-out for the PLS model on data set 2.

N	Single LNO	Multiple LNO[a]	
	Q^2_{LNO}	$<Q^2_{LNO}>$	$\sigma(Q^2_{LNO})$
1	0.8409	0.8409	0
2	0.8466	0.8415	0.0031
3	0.8394	0.8387	0.0059
4	0.8449	0.8387	0.0056
5	0.8448	0.8418	0.0059
6	0.8499	0.8346	0.0077
7	0.8277	0.8355	0.0048
8	0.8533	0.8394	0.0084
9	0.8342	0.8369	0.0105
10	0.8193	0.8418	0.0078
11	0.7739	0.8377	0.0128
12	0.8386	0.8313	0.0170
13	0.8493	0.8420	0.0098
14	0.7673	0.8405	0.0133
15	0.8291	0.8424	0.0060
16	0.8322	0.8418	0.0078
17	0.8293	0.8377	0.0085
18	0.7567	0.8125	0.0375
19	0.8316	0.8141	0.0345
20	0.8517	0.8292	0.0149
21	0.8554	0.8216	0.0206
22	0.8567	0.8308	0.0214
23	0.8553	0.8272	0.0192
24	0.8389	0.8210	0.0333
25	0.8481	0.8121	0.0348
26	0.8330	0.8194	0.0395
27	0.7881	0.8245	0.0417
28	0.7602	0.8367	0.0240

[a]Average values of Q^2_{LNO} and respective standard deviations are reported for multiple LNO (ten times reordered data set).

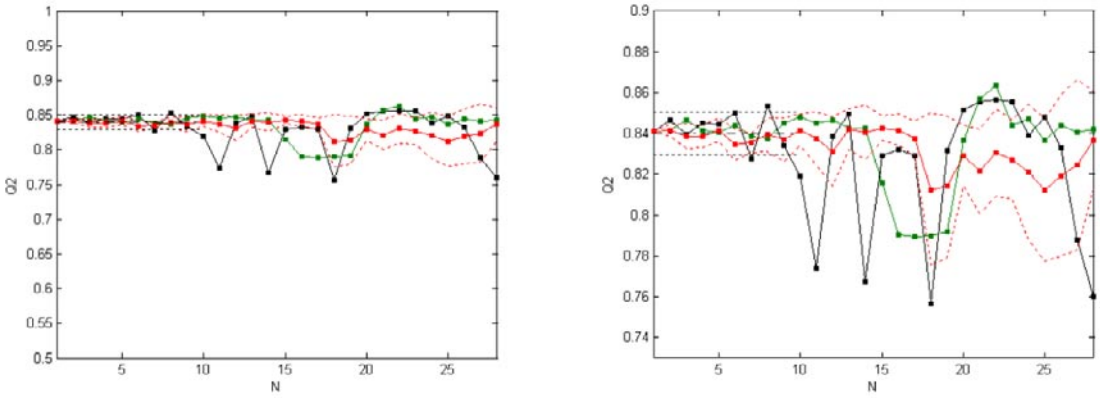

Figure F3. Leave-*N*-out crossvalidation plot for the PLS model on data set 2, showing the general trend (left) and details (right). Black - single LNO, red - multiple LNO (10 times). Single LNO: average Q^2 - dot-dash line, one standard deviation below and above the average - dotted lines. Multiple LNO: one standard deviation below and above the average - red dotted curved lines. Green - single LNO using the original (not randomized) data set.

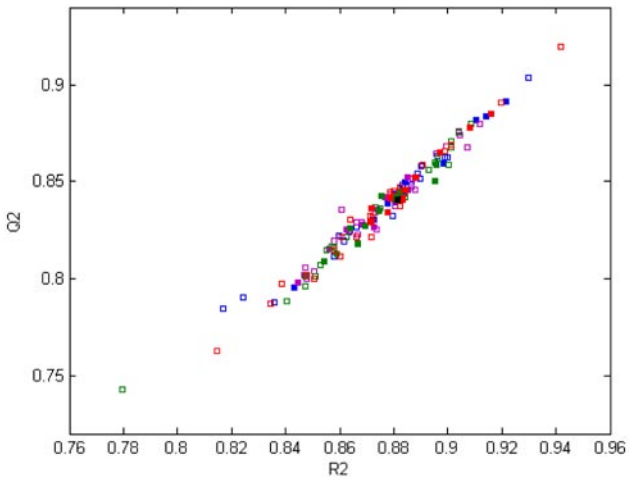

Figure F4. A comparative plot for bootstrappings for the PLS model on data set 2: the real model (black square), models from HCA-based bootstrappings (blue squares: 10 iterations - solid, 25 iterations - open), models from bootstrapping based on classes of Y (pink squares: 10 iterations - solid, 25 iterations - open), models from bootstrapping based MDR resistance classes (green squares: 10 iterations - solid, 25 iterations - open), and models from simple bootstrappings (red squares: 10 iterations - solid, 25 iterations - open).

Table T11. Ten bootstrappings with HCA clustering (BC), with classes of **y** (BY), with MDR resistance classes (BM) and only with random selection (BR) for the PLS model on data set 2.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BM}	R^2_{BM}	Q^2_{BR}	R^2_{BR}
1	0.8595	0.8983	0.8525	0.8853	0.8587	0.8958	0.8345	0.8778
2	0.8433	0.8820	0.8446	0.8822	0.8505	0.8954	0.8361	0.8716
3	0.8435	0.8810	0.7980	0.8447	0.8425	0.8755	0.8406	0.8833
4	0.8496	0.8841	0.8158	0.8559	0.8257	0.8640	0.8411	0.8787
5	0.8386	0.8777	0.8267	0.8729	0.8131	0.8590	0.8782	0.9083
6	0.8302	0.8729	0.8280	0.8687	0.8349	0.8744	0.8455	0.8848
7	0.8819	0.9104	0.8352	0.8739	0.8088	0.8542	0.8524	0.8878
8	0.8915	0.9215	0.8249	0.8624	0.8271	0.8695	0.8849	0.9159
9	0.8834	0.9140	0.8023	0.8468	0.8180	0.8665	0.8651	0.8971
10	0.7956	0.8434	0.8422	0.8766	0.8453	0.8825	0.8288	0.8717
Average	0.8517	0.8885	0.8270	0.8669	0.8325	0.8737	0.8507	0.8877
Stand. dev.	0.0288	0.0232	0.0178	0.0141	0.0167	0.0141	0.0192	0.0150
The model	0.8409	0.8814	0.8409	0.8814	0.8409	0.8814	0.8409	0.8814

Table T12. Twenty-five bootstrappings with HCA clustering (BC), with classes of **y** (BY), with MDR resistance classes (BM) and only with random selection (BR) for the PLS model on data set 2.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BM}	R^2_{BM}	Q^2_{BR}	R^2_{BR}
1	0.8630	0.8964	0.8272	0.8696	0.8416	0.8842	0.8118	0.8603
2	0.8623	0.8988	0.8459	0.8878	0.8599	0.8954	0.8654	0.8988
3	0.8119	0.8577	0.8246	0.8641	0.8751	0.9039	0.8018	0.8477
4	0.7878	0.8359	0.8001	0.8476	0.8159	0.8579	0.8445	0.8787
5	0.8222	0.8599	0.8227	0.8669	0.8801	0.9087	0.8322	0.8715
6	0.8484	0.8867	0.8800	0.9120	0.7887	0.8402	0.9198	0.9417
7	0.8484	0.8834	0.8058	0.8473	0.8147	0.8550	0.8210	0.8608
8	0.8195	0.8615	0.8676	0.9073	0.8563	0.8929	0.8909	0.9197
9	0.8539	0.8887	0.8763	0.9039	0.8072	0.8530	0.8213	0.8719
10	0.8467	0.8823	0.8375	0.8807	0.7426	0.7797	0.8003	0.8504
11	0.8241	0.8636	0.8642	0.8959	0.8679	0.9014	0.8689	0.9012
12	0.8450	0.8841	0.8431	0.8787	0.8442	0.8822	0.8454	0.8802
13	0.8320	0.8794	0.8356	0.8606	0.8368	0.8732	0.8376	0.8823
14	0.9035	0.9297	0.8738	0.9043	0.8358	0.8752	0.8585	0.8906
15	0.8455	0.8853	0.8439	0.8841	0.8410	0.8803	0.8315	0.8721
16	0.8440	0.8811	0.8251	0.8735	0.8620	0.8961	0.8438	0.8838
17	0.7844	0.8168	0.8468	0.8867	0.8584	0.9001	0.7630	0.8146
18	0.8416	0.8809	0.8468	0.8867	0.8710	0.9014	0.8214	0.8662
19	0.8517	0.8895	0.8681	0.8996	0.8213	0.8625	0.8146	0.8580
20	0.8624	0.8999	0.8195	0.8579	0.7962	0.8474	0.8300	0.8711
21	0.8578	0.8901	0.8342	0.8733	0.8013	0.8511	0.7871	0.8344
22	0.8395	0.8804	0.8290	0.8681	0.8403	0.8808	0.8520	0.8882
23	0.8262	0.8661	0.8292	0.8663	0.8016	0.8471	0.8302	0.8640
24	0.8458	0.8835	0.8037	0.8507	0.8419	0.8811	0.7972	0.8385
25	0.7902	0.8241	0.8514	0.8868	0.8166	0.8570	0.8470	0.8822
Average	0.8383	0.8762	0.8401	0.8784	0.8327	0.8723	0.8335	0.8732
Stand. dev.	0.0264	0.0244	0.0225	0.0188	0.0323	0.0284	0.0332	0.0268
The model	0.8409	0.8814	0.8409	0.8814	0.8409	0.8814	0.8409	0.8814

Table T13. Ten **y**-randomizations for the PLS model on data set 2.

Randomization	Q^2	R^2
1	0.0304	0.2332
2	–0.2744	0.0257
3	–0.1390	0.1047
4	–0.1468	0.1293
5	–0.2112	0.1165
6	–0.1667	0.1203
7	–0.1892	0.0728
8	–0.3055	0.0920
9	–0.1855	0.0938
10	–0.1616	0.1207
Maximum	0.0304	0.2332
Average	–0.1749	0.1109
Standard deviation	0.0901	0.0526
The model	0.8409	0.8814

Table T14. Twenty-five **y**-randomizations for the PLS model on data set 2.

Randomization	Q^2	R^2
1	–0.1859	0.0827
2	–0.1017	0.1257
3	–0.2452	0.0731
4	–0.1850	0.0602
5	–0.1861	0.1117
6	–0.0807	0.1708
7	–0.0462	0.1939
8	–0.1062	0.1289
9	–0.2632	0.0640
11	–0.0306	0.1461
12	–0.1365	0.1260
13	–0.1653	0.1317
14	–0.2246	0.0880
15	–0.1668	0.0995
16	–0.1735	0.1269
17	–0.1802	0.0521
18	–0.2296	0.0730
19	–0.2282	0.1212
20	–0.1912	0.0807
21	–0.2293	0.0516
22	–0.1348	0.1441
23	–0.2692	0.0333
24	–0.1812	0.1043
25	–0.0817	0.1492
Maximum	–0.0306	0.1939
Average	0.0662	0.0412
Standard deviation	0.0776	0.0508
The model	0.8409	0.8814

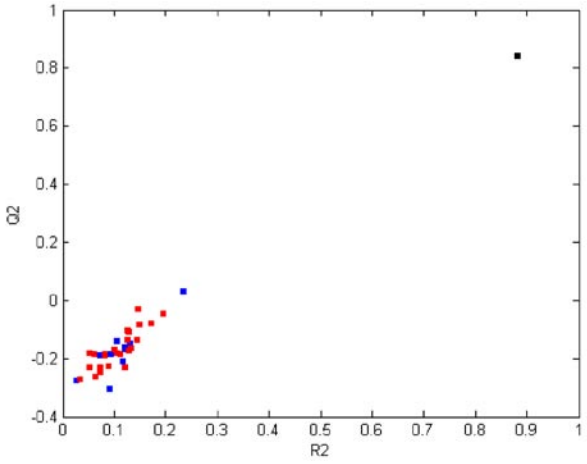

Figure F5. The **y**-randomization plot for the PLS model on data set 2: black square - the real model, blue squares - 10 randomized models, red squares - 25 randomized models.

Table T15. Comparative statistics of 10 and 25 **y**-randomizations for the PLS model on data set 2.

Parameter[a]	10 iterations	25 iterations
Maximum (Q^2_{yrand})	–0.030	–0.031
Maximum (R^2_{yrand})	0.233	0.194
Standard deviation (Q^2_{yrand})	0.090	0.078
Standard deviation (R^2_{yrand})	0.053	0.051
Minimum model-random. diff. (Q^2_{yrand})[b]	9.67	11.23
Minimum model-random. diff. (R^2_{yrand})[b]	12.32	13.53
Confidence level for min. diff. (Q^2_{yrand})[c]	<0.0001	<0.0001
Confidence level for min. diff. (R^2_{yrand})[c]	<0.0001	<0.0001
Randomizations %, conf. level > 0.0001 (Q^2_{yrand})[d]	0	0
Randomizations %, conf. level > 0.0001 (R^2_{yrand})[d]	0	0
y-Randomization intercept (r_{yrand} *vs.* Q^2_{yrand})[e]	–0.219	–0.262
y-Randomization intercept (r_{yrand} *vs.* R^2_{yrand})[e]	0.077	0.035

[a]Statistical parameters are calculated for Q^2 from from **y**-randomization (Q^2_{yrand}) and R^2 from **y**-randomization (R^2_{yrand}). [b]Minimum model-randomizations difference: the difference between the real model (Table 1) and the best **y**-randomization in terms of correlation coefficients Q^2_{yrand} or R^2_{yrand}, expressed in units of the standard deviations of Q^2_{yrand} or R^2_{yrand}, respectively. The best **y**-randomization is defined by the highest Q^2_{rand} or R^2_{rand}. [c]Confidence level for normal distribution of the minimum difference between the real and randomized models. [d]Percentage of randomizations characterized by the difference between the real and randomized models (in terms of Q^2_{yrand} or R^2_{yrand}) at confidence levels > 0.0001. [e]Intercepts obtained from two **y**-randomization plots for each regression model proposed. Q^2_{yrand} or R^2_{yrand} is the vertical axis, whilst the horizontal axis is the absolute value of the correlation coefficient r_{yrand} between the original and randomized vectors **y**. The randomization plots are completed with the data for the real model (r_{yrand} = 1.000, Q^2 or R^2).

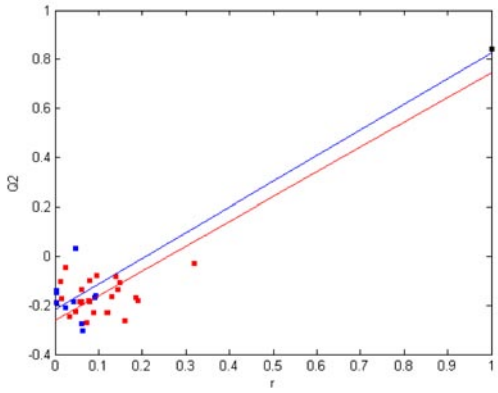

Figure F6. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the PLS model on data set 2: $Q^2 = -0.219 + 1.044\ r$ and $Q^2 = -0.262 + 1.007\ r$, respectively.

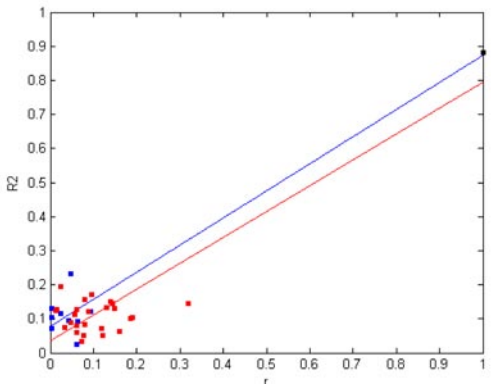

Figure F7. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the PLS model on data set 2: $R^2 = 0.077 + 0.7956\ r$ and $R^2 = 0.035 + 0.759\ r$, respectively.

ANALYSIS OF DATA SET 3

Table T16. Results for leave-*N*-out for the PLS model on data set 3.

N	Single LNO	Multiple LNO[a]	
	Q^2_{LNO}	$\langle Q^2_{LNO} \rangle$	$\sigma(Q^2_{LNO})$
1	0.8951	0.8951	0
2	0.8939	0.8949	0.0015
3	0.8774	0.8960	0.0028
4	0.8875	0.8917	0.0043
5	0.8965	0.8941	0.0058
6	0.8691	0.8896	0.0130
7	0.8982	0.8958	0.0039
8	0.8795	0.8935	0.0072
9	0.8868	0.8935	0.0068
10	0.8968	0.8926	0.0086
11	0.8937	0.8889	0.0127
12	0.8898	0.8849	0.0140
13	0.8926	0.8883	0.0105
14	0.9009	0.8902	0.0114
15	0.8717	0.8976	0.0067
16	0.8559	0.8963	0.0072
17	0.8892	0.8900	0.0187
18	0.8701	0.8863	0.0222
19	0.8915	0.8919	0.0125
20	0.8956	0.8750	0.0255
21	0.9120	0.8816	0.0242
22	0.9108	0.8781	0.0189
23	0.9091	0.8867	0.0110
24	0.9069	0.8854	0.0155
25	0.9066	0.8875	0.0156

[a]Average values of Q^2_{LNO} and respective standard deviations are reported for multiple LNO (ten times reordered data set).

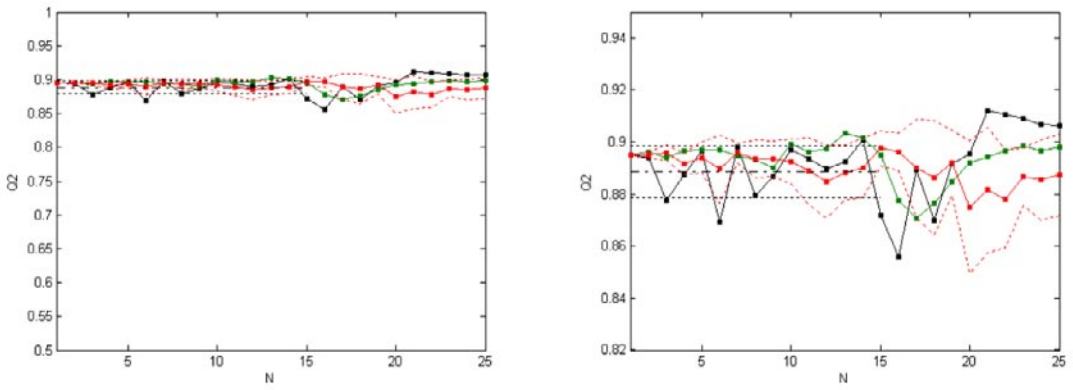

Figure F8. Leave-*N*-out crossvalidation plot for the PLS model on data set 3, showing the general trend (left) and details (right). Black - single LNO, red - multiple LNO (10 times). Single LNO: average Q^2 - dot-dash line, one standard deviation below and above the average - dotted lines. Multiple LNO: one standard deviation below and above the average - red dotted curved lines. Green - single LNO using the original (not randomized) data set.

Table T16. Ten bootstrappings with HCA clustering (BC), with classes of **y** (BY) and only with random selection (BR) for the PLS model on data set 3.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BR}	R^2_{BR}
1	0.8821	0.9108	0.8864	0.9130	0.8848	0.9102
2	0.8922	0.9185	0.8799	0.9066	0.8952	0.9209
3	0.8874	0.9178	0.8710	0.9003	0.8885	0.9209
4	0.8786	0.9050	0.8933	0.9179	0.9064	0.9324
5	0.8911	0.9161	0.8938	0.9175	0.8903	0.9172
6	0.9015	0.9246	0.8759	0.9038	0.8791	0.9076
7	0.8941	0.9222	0.8917	0.9148	0.9154	0.9394
8	0.8820	0.9086	0.8924	0.9202	0.8820	0.9092
9	0.8773	0.9095	0.9050	0.9281	0.8922	0.9190
10	0.8861	0.9111	0.9031	0.9262	0.8929	0.9176
Average	0.8872	0.9144	0.8892	0.9148	0.8927	0.9194
Standard deviation	0.0076	0.0064	0.0111	0.0092	0.0110	0.0101
The model	0.8951	0.9154	0.8951	0.9154	0.8951	0.9154

Table T17. Twenty-five bootstrappings with HCA clustering (BC), with classes of **y** (BY) and only with random selection (BR) for the PLS model on data set 3.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BR}	R^2_{BR}
1	0.8861	0.9119	0.9018	0.9215	0.8821	0.9106
2	0.9060	0.9309	0.8791	0.9061	0.9003	0.9233
3	0.8761	0.9034	0.8955	0.9225	0.8957	0.9203
4	0.8806	0.9076	0.8976	0.9178	0.8871	0.9128
5	0.8939	0.9182	0.8948	0.9225	0.8774	0.9039
6	0.8824	0.9092	0.9176	0.9390	0.9107	0.9316
7	0.8776	0.9069	0.8908	0.9163	0.8772	0.9036
8	0.8776	0.9069	0.8857	0.9106	0.8940	0.9182
9	0.8870	0.9124	0.9073	0.9260	0.8809	0.9055
10	0.8804	0.9090	0.8917	0.9151	0.8739	0.9013
11	0.8855	0.9126	0.8927	0.9190	0.9158	0.9347
12	0.8956	0.9172	0.9051	0.9288	0.8899	0.9171
13	0.8959	0.9216	0.8939	0.9172	0.8756	0.9028
14	0.8902	0.9193	0.8941	0.9165	0.8770	0.9042
15	0.8920	0.9086	0.8711	0.8998	0.8933	0.9177
16	0.8935	0.9168	0.8836	0.9091	0.8853	0.8982
17	0.8962	0.9216	0.8963	0.9199	0.8734	0.9013
18	0.8989	0.9146	0.8774	0.9037	0.8941	0.9177
19	0.9024	0.9292	0.9085	0.9308	0.8895	0.9148
20	0.8950	0.9231	0.8835	0.9089	0.8736	0.9017
21	0.8767	0.9048	0.8756	0.9035	0.9003	0.9293
22	0.9004	0.9310	0.8900	0.9165	0.9041	0.9139
23	0.8914	0.9154	0.8834	0.9102	0.9101	0.9347
24	0.8795	0.9119	0.8897	0.9154	0.8797	0.9057
25	0.8870	0.9124	0.9223	0.9294	0.8872	0.9108
Average	0.8891	0.9151	0.8932	0.9170	0.8891	0.9134
Standard deviation	0.0087	0.0078	0.0125	0.0094	0.0125	0.0110
The model	0.8951	0.9154	0.8951	0.9154	0.8951	0.9154

Table T18. Ten **y**-randomizations for the PLS model on data set 3.

Randomization	Q^2	R^2
1	–0.2014	0.0539
2	–0.1909	0.0582
3	–0.0200	0.1402
4	–0.0356	0.1401
5	–0.2101	0.0519
6	–0.1969	0.0178
7	–0.1950	0.0847
8	–0.1780	0.0358
9	–0.1458	0.0715
10	–0.5926	0.0692
Maximum	–0.0200	0.1402
Average	–0.1966	0.0723
Standard deviation	0.1553	0.0403
The model	0.8951	0.9154

Table T19. Twenty-five **y**-randomizations for the PLS model on data set 3.

Randomization	Q^2	R^2
1	–0.0928	0.1172
2	0.0671	0.2237
3	–0.2499	0.0504
4	–0.2531	0.0446
5	–0.1525	0.0732
6	–0.1101	0.0784
7	–0.0417	0.1719
8	–0.1675	0.0851
9	–0.0631	0.1259
11	–0.1041	0.0833
12	–0.1200	0.0750
13	–0.0770	0.1466
14	–0.1658	0.0514
15	–0.0236	0.1409
16	–0.0847	0.2243
17	–0.0874	0.1239
18	–0.2212	0.0261
19	–0.1092	0.0935
20	–0.1526	0.0813
21	–0.0401	0.1307
22	–0.1790	0.0864
23	–0.0561	0.1082
24	–0.1154	0.1088
25	–0.2619	0.0403
Maximum	0.0671	0.2243
Average	–0.1188	0.1037
Standard deviation	0.0776	0.0508
The model	0.8951	0.9154

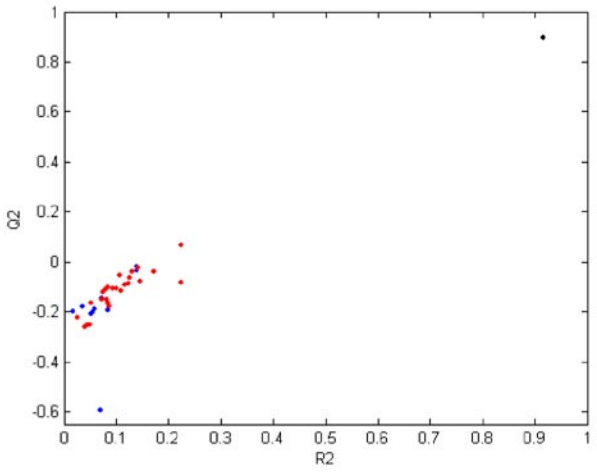

Figure F9. The **y**-randomization plot for the PLS model on data set 3: black ball - the real model, blue balls - 10 randomized models, red balls - 25 randomized models.

Table T20. Comparative statistics of 10 and 25 **y**-randomizations for the PLS model on data set 3.

Parameter[a]	10 iterations	25 iterations
Maximum (Q^2_{yrand})	–0.020	0.067
Maximum (R^2_{yrand})	0.140	0.224
Standard deviation (Q^2_{yrand})	0.155	0.078
Standard deviation (R^2_{yrand})	0.040	0.051
Minimum model-random. diff. (Q^2_{yrand})[b]	11.79	10.67
Minimum model-random. diff. (R^2_{yrand})[b]	19.24	13.60
Confidence level for min. diff. (Q^2_{yrand})[c]	<0.0001	<0.0001
Confidence level for min. diff. (R^2_{yrand})[c]	<0.0001	<0.0001
Randomizations %, conf. level > 0.0001 (Q^2_{yrand})[d]	0	0
Randomizations %, conf. level > 0.0001 (R^2_{yrand})[d]	0	0
y-Randomization intercept (r_{yrand} *vs.* Q^2_{yrand})[e]	–0.304	–0.231
y-Randomization intercept (r_{yrand} *vs.* R^2_{yrand})[e]	–0.004	0.016

[a]Statistical parameters are calculated for Q^2 from from **y**-randomization (Q^2_{yrand}) and R^2 from **y**-randomization (R^2_{yrand}). [b]Minimum model-randomizations difference: the difference between the real model (Table 1) and the best **y**-randomization in terms of correlation coefficients Q^2_{yrand} or R^2_{yrand}, expressed in units of the standard deviations of Q^2_{yrand} or R^2_{yrand}, respectively. The best **y**-randomization is defined by the highest Q^2_{rand} or R^2_{rand}. [c]Confidence level for normal distribution of the minimum difference between the real and randomized models. [d]Percentage of randomizations characterized by the difference between the real and randomized models (in terms of Q^2_{yrand} or R^2_{yrand}) at confidence levels > 0.0001. [e]Intercepts obtained from two **y**-randomization plots for each regression model proposed. Q^2_{yrand} or R^2_{yrand} is the vertical axis, whilst the horizontal axis is the absolute value of the correlation coefficient r_{yrand} between the original and randomized vectors **y**. The randomization plots are completed with the data for the real model (r_{yrand} = 1.000, Q^2 or R^2).

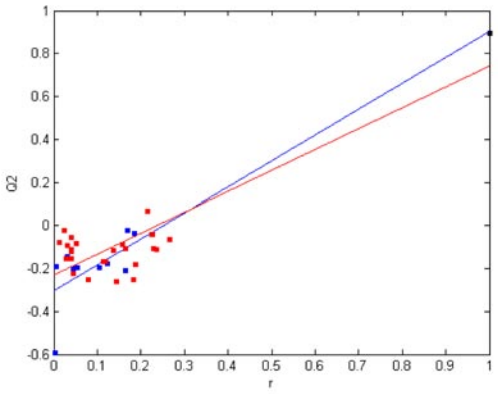

Figure F10. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the PLS model on data set 3: $Q^2 = -0.304 + 1.207\ r$ and $Q^2 = -0.231 + 0.972\ r$, respectively.

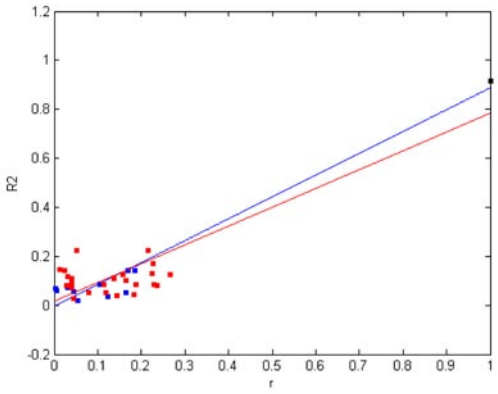

Figure F11. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from y-randomizations of the PLS model on data set 3: $R^2 = -0.004 + 0.891\ r$ and $R^2 = 0.016 + 0.768\ r$, respectively.

ANALYSIS OF DATA SET 4

Table T21. Results for leave-*N*-out for the MLR model on data set 4.

N	Single LNO	Multiple LNO[a]	
	Q^2_{LNO}	$\langle Q^2_{LNO} \rangle$	$\sigma(Q^2_{LNO})$
1	0.7977	0.7977	0
2	0.7093	0.8012	0.0284
3	0.7230	0.7460	0.0536
4	0.6846	0.7778	0.0664
5	0.8494	0.7713	0.0764
6	0.5663	0.6719	0.2721
7	0.7706	0.7178	0.1872

[a]Average values of Q^2_{LNO} and respective standard deviations are reported for multiple LNO (ten times reordered data set).

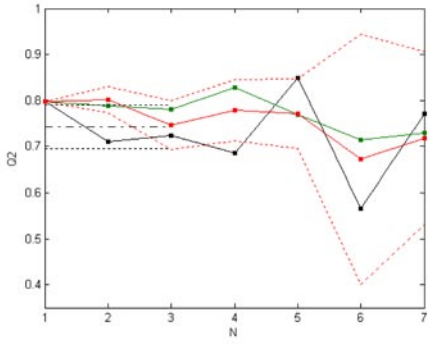

Figure F12. Leave-*N*-out crossvalidation plot for the MLR model on data set 4. Black - single LNO, red - multiple LNO (10 times). Single LNO: average Q^2 - dot-dash line, one standard deviation below and above the average - dotted lines. Multiple LNO: one standard deviation below and above the average - red dotted curved lines. Green - single LNO using the original (not randomized) data set.

Table T22. Ten bootstrappings with HCA clustering (BC) and only with random selection (BR) for the MLR model on data set 4.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BR}	R^2_{BR}
1	0.7868	0.8937	0.8043	0.8979
2	0.7059	0.8739	0.8160	0.9076
3	0.7948	0.8963	0.8209	0.9108
4	0.6379	0.8682	0.8372	0.9171
5	0.7620	0.8647	0.7741	0.9000
6	0.7705	0.8801	0.6337	0.8880
7	0.7705	0.8801	0.7490	0.8872
8	0.5760	0.8551	0.8216	0.9117
9	0.7560	0.8755	0.8100	0.9023
10	0.7620	0.8647	0.7873	0.8888
Average	0.7322	0.8752	0.7865	0.8997
Stand. dev.	0.0716	0.0130	0.0564	0.0112
The model	0.7977	0.8857	0.7977	0.8857

Table T23. Twenty-five bootstrappings with HCA clustering (BC) and only with random selection (BR) for the MLR model on data set 4.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BR}	R^2_{BR}
1	0.7868	0.8937	0.8876	0.9285
2	0.7796	0.8929	0.8543	0.9075
3	0.7298	0.8760	0.7764	0.8839
4	0.7721	0.8796	0.6932	0.8662
5	0.7664	0.8616	0.8042	0.9005
6	0.6773	0.8681	0.7314	0.8708
7	0.7462	0.8886	0.8334	0.9067
8	0.8372	0.9171	0.7796	0.8929
9	0.7620	0.8647	0.7033	0.8731
10	0.8106	0.9031	0.8145	0.8950
11	0.7835	0.8861	0.8065	0.9017
12	0.6306	0.8570	0.6998	0.8431
13	0.7721	0.8796	0.8074	0.9005
14	0.6953	0.8690	0.7939	0.8955
15	0.8160	0.9076	0.7879	0.8972
16	0.6379	0.8682	0.8158	0.9147
17	0.7059	0.8739	0.8100	0.9023
18	0.6993	0.8639	0.6827	0.8859
19	0.7873	0.8907	0.8069	0.9034
20	0.6379	0.8682	0.7879	0.8972
21	0.7705	0.8801	0.7764	0.8839
22	0.7930	0.8798	0.7796	0.8828
23	0.6345	0.8748	0.8224	0.8994
24	0.8074	0.9005	0.7344	0.8748
25	0.7764	0.8839	0.7873	0.8888
Average	0.7446	0.8811	0.7831	0.8919
Stand. dev.	0.0623	0.0153	0.0507	0.0175
The model	0.7977	0.8857	0.7977	0.8857

Table T24. Ten **y**-randomizations for the MLR model on data set 4.

Randomization	Q^2	R^2
1	–1.1201	0.0511
2	–1.0796	0.0494
3	–0.5260	0.1987
4	–0.6397	0.1189
5	–0.2905	0.2057
6	–0.5428	0.3060
7	–0.8073	0.0584
8	–0.3572	0.1894
9	–0.4147	0.1759
10	–0.2020	0.4037
Maximum	–0.2020	0.4037
Average	–0.5980	0.1757
Standard deviation	0.3164	0.1150
The model	0.7977	0.8857

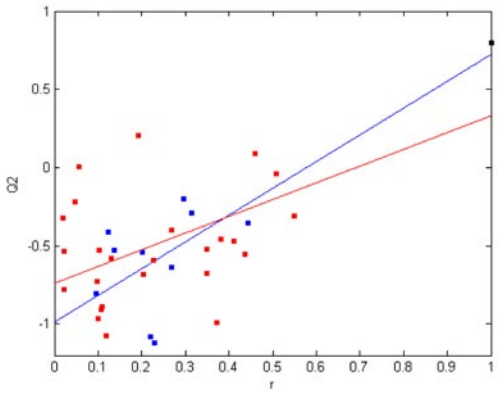

Figure F13. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the MLR model on data set 4: $Q^2 = –0.989 + 1.710\ r$ and $Q^2 = –0.739 + 1.069\ r$, respectively.

Table T25. Twenty-five **y**-randomizations for the MLR model on data set 4.

Randomization	Q^2	R^2
1	–0.6768	0.2281
2	–0.5245	0.2219
3	–0.6821	0.0685
4	–0.9890	0.1179
5	–0.3997	0.2255
6	–0.5940	0.1220
7	–0.5796	0.2874
8	–0.5530	0.2320
9	–0.3110	0.3728
10	–0.9104	0.0214
11	0.0926	0.5633
12	–0.7789	0.0541
13	–0.4595	0.1575
14	–0.2211	0.3617
15	–0.0381	0.4240
16	–0.8862	0.0185
17	–0.9625	0.0099
18	–1.0739	0.0114
19	0.0086	0.3491
20	–0.4680	0.2169
21	–0.5269	0.1176
22	–0.5318	0.1148
23	–0.7300	0.1343
24	0.2062	0.4832
25	–0.3210	0.2831
Maximum	0.2062	0.5633
Average	–0.5164	0.2079
Standard deviation	0.3409	0.1534
The model	0.7977	0.8857

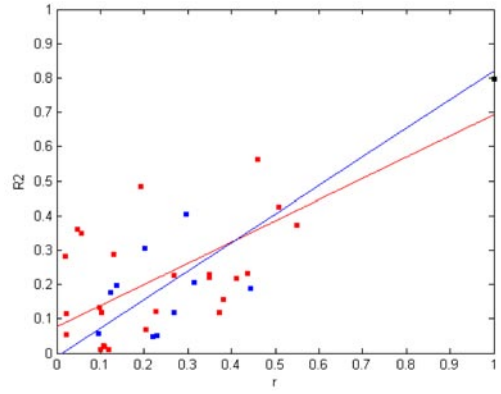

Figure F14. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the MLR model on data set 4: $R^2 = –0.011 + 0.830\ r$ and $R^2 = 0.077 + 0.616\ r$, respectively.

ANALYSIS OF DATA SUBSET 3

Table T26. The QSPR data subset 3 for carbonyl oxygen chemical shift in substituted benzaldehydes.[a,b,c]

No.	E_e/eV	E_{CC}/eV	Δ_{HL}/eV	σ_b/Å	σ_ρ/Å	D_{CC}/Å	Q_{C2mul}	Q_{Omul}	δ_{exp}/ppm
4	72.092	122.819	–8.336	0.071	0.008	1.480	–0.240	–0.326	532.8
6	65.144	122.617	–9.335	0.072	0.005	1.485	–0.215	–0.313	568.9
7	70.884	122.591	–8.930	0.072	0.003	1.485	–0.204	–0.313	570.1
8	70.633	122.540	–9.315	0.072	0.007	1.486	–0.196	–0.311	570.3
9	70.181	122.431	–9.140	0.072	0.003	1.487	–0.182	–0.305	593.6
13	71.304	122.324	–9.336	0.073	0.006	1.489	–0.217	–0.307	555.0
14	67.203	121.716	–9.367	0.076	0.005	1.499	–0.093	–0.282	576.0
19	71.291	122.541	–9.483	0.072	0.003	1.485	–0.216	–0.311	568.4
27	73.652	123.179	–8.188	0.068	0.009	1.474	–0.286	–0.355	512.1
32	74.192	123.283	–8.292	0.069	0.015	1.473	–0.302	–0.357	513.9
33	74.230	123.166	–8.570	0.069	0.011	1.475	–0.303	–0.351	518.2
36	74.722	123.230	–8.629	0.068	0.010	1.473	–0.318	–0.354	509.0
37	74.194	123.138	–8.139	0.068	0.010	1.475	–0.303	–0.351	520.0
46	70.328	122.471	–8.461	0.073	0.008	1.487	–0.188	–0.305	560.0
48	74.739	123.148	–8.062	0.069	0.010	1.475	–0.318	–0.348	505.0

[a]This data subset was generated from data set 3 (Table T3). [b]Names of samples and variables are from the original publication for data set 3. [c]The two samples in the external validation set are **7** and **37**. This selection is based on HCA analysis of this data subset.

Table T27. Results for leave-*N*-out for the PLS model on data subset 3.

N	Single LNO	Multiple LNO[a]	
	Q^2_{LNO}	$\langle Q^2_{LNO} \rangle$	$\sigma(Q^2_{LNO})$
1	0.7785	0.7785	0
2	0.7626	0.7811	0.0164
3	0.7787	0.7752	0.0281
4	0.8366	0.7738	0.0313
5	0.6233	0.7393	0.0672
6	0.6735	0.7593	0.0788
7	0.8214	0.7168	0.1231
Average[b]	0.7891		
Standard deviation[b]	0.0326		

[a]Average values of Q^2_{LNO} and respective standard deviations are reported for multiple LNO (ten times reordered data set). [b]Values calculated from Q^2_{LNO} for N =1, 2, 3, 4.

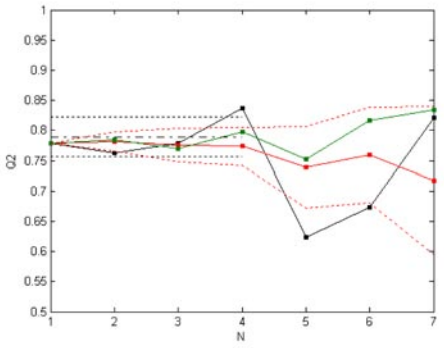

Figure F15. Leave-*N*-out crossvalidation plot for the PLS model on data subset 3. Black - single LNO, red - multiple LNO (10 times). Single LNO: average Q^2 - dot-dash line, one standard deviation below and above the average - dotted lines. Multiple LNO: one standard deviation below and above the average - red dotted curved lines. Green - single LNO using the original (not randomized) data set.

Table T28. Ten bootstrappings with HCA clustering (BC), with classes of Y (BY) and only with random selection (BR) for the PLS model on data subset 3.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BR}	R^2_{BR}
1	0.7879	0.9134	0.8176	0.8995	0.7479	0.8932
2	0.7773	0.9042	0.8176	0.8995	0.8541	0.9077
3	0.8567	0.9075	0.8495	0.9022	0.7501	0.8942
4	0.8176	0.8995	0.7970	0.9184	0.7556	0.8919
5	0.7505	0.8882	0.7695	0.8970	0.7890	0.9058
6	0.7251	0.8599	0.7186	0.8907	0.7611	0.8958
7	0.7293	0.8623	0.7498	0.8957	0.9065	0.9385
8	0.7959	0.9105	0.7479	0.8932	0.7342	0.8512
9	0.7511	0.9144	0.7627	0.8966	0.7501	0.8594
10	0.7695	0.8970	0.7505	0.8882	0.7405	0.8843
Average	0.7761	0.8957	0.7781	0.8981	0.7789	0.8922
Standard deviation	0.0406	0.0199	0.0407	0.0083	0.0568	0.0245
The model	0.7785	0.8968	0.7785	0.8968	0.7785	0.8968

Table T29. Twenty-five bootstrappings with HCA clustering (BC), with classes of Y (BY) and only with random selection (BR) for the PLS model on data subset 3.

Bootstrapping	Q^2_{BC}	R^2_{BC}	Q^2_{BY}	R^2_{BY}	Q^2_{BR}	R^2_{BR}
1	0.7475	0.8539	0.8063	0.9203	0.7792	0.9016
2	0.7705	0.8958	0.7099	0.8876	0.8567	0.9075
3	0.8176	0.8995	0.7374	0.8918	0.7214	0.8804
4	0.8495	0.9022	0.7302	0.8859	0.7293	0.8623
5	0.8495	0.9022	0.7879	0.9134	0.7444	0.8858
6	0.8567	0.9075	0.7296	0.8876	0.8176	0.8995
7	0.7959	0.9105	0.7348	0.8938	0.7501	0.8942
8	0.7479	0.8932	0.8063	0.9203	0.8541	0.9077
9	0.7479	0.8932	0.8098	0.9201	0.7296	0.8876
10	0.7342	0.8512	0.8063	0.9203	0.7706	0.8979
11	0.7394	0.8532	0.7879	0.9134	0.7382	0.8855
12	0.7505	0.8882	0.7695	0.8970	0.7475	0.8539
13	0.7292	0.8652	0.7879	0.9134	0.7231	0.9038
14	0.7511	0.9144	0.7348	0.8938	0.7501	0.8942
15	0.7342	0.8907	0.8017	0.9232	0.7465	0.8735
16	0.8093	0.8954	0.8567	0.9075	0.7367	0.8556
17	0.7292	0.8652	0.8531	0.9037	0.7183	0.8517
18	0.7498	0.8957	0.8314	0.9280	0.7542	0.8606
19	0.7496	0.8583	0.7970	0.9184	0.7342	0.8512
20	0.8148	0.9242	0.7380	0.8892	0.7628	0.9014
21	0.7501	0.8942	0.7099	0.8876	0.8202	0.9015
22	0.7511	0.9144	0.7879	0.9134	0.7496	0.8854
23	0.7556	0.8919	0.7498	0.8957	0.7475	0.8539
24	0.8531	0.9037	0.8196	0.9199	0.7705	0.8958
25	0.7296	0.8876	0.7348	0.8938	0.8559	0.9086
Average	0.7726	0.8901	0.7767	0.9056	0.7643	0.0201
Standard deviation	0.0437	0.0206	0.0432	0.0140	0.0427	0.0201
The model	0.7785	0.8968	0.7785	0.8968	0.7785	0.8968

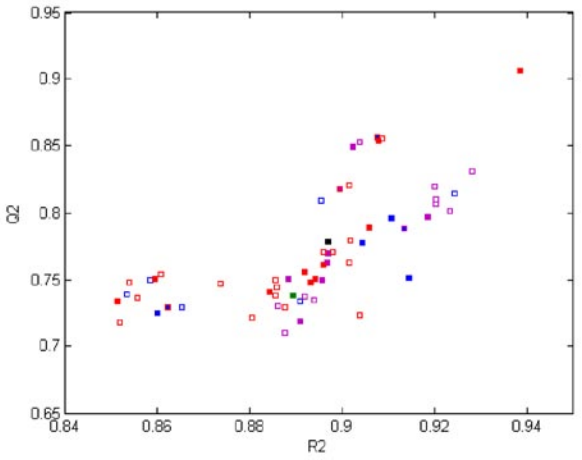

Figure F16. A comparative plot for bootstrappings for the PLS model on data subset 3: the real model (black square) its external validation (green square), models from HCA-based bootstrappings (blue squares: 10 iterations - solid, 25 iterations - open), models from bootstrapping based on classes of Y (pink squares: 10 iterations - solid, 25 iterations - open), , and models from simple bootstrappings (red squares: 10 iterations - solid, 25 iterations - open).

Table T30. Ten **y**-randomizations for the PLS model on data subset 3.

Randomization	Q^2	R^2
1	–1.0528	0.0931
2	–0.4425	0.3529
3	–0.1817	0.6157
4	–0.1583	0.4801
5	–0.1583	0.4801
6	0.0198	0.4445
7	–0.4868	0.4298
8	–0.8862	0.2171
9	–0.5164	0.3825
10	–1.2110	0.1653
Maximum	0.0198	0.6157
Average	–0.5074	0.3661
Standard deviation	0.4175	0.1619
The model	0.7785	0.8968

Table T31. Twenty-five **y**-randomizations for the PLS model on data subset 3.

Randomization	Q^2	R^2
1	–0.4923	0.2017
2	–0.4502	0.2120
3	–0.5006	0.3392
4	–0.6806	0.1838
5	–0.9556	0.2263
6	0.0029	0.4618
7	–0.1742	0.4167
8	0.2178	0.5924
9	–0.6725	0.1585
11	–0.5339	0.2831
12	–1.5136	0.1719
13	–0.4559	0.2486
14	–0.6581	0.3188
15	–1.1838	0.1467
16	–1.2932	0.2028
17	–0.4852	0.4049
18	–1.0034	0.0888
19	–1.2538	0.1834
20	–0.3780	0.3315
21	–0.6874	0.2471
22	–0.5331	0.1824
23	–1.3436	0.1321
24	0.1273	0.4754
25	–0.5868	0.3387
Maximum	0.2178	0.5924
Average	–0.6178	0.2830
Standard deviation	0.4669	0.1334
The model	0.7785	0.8968

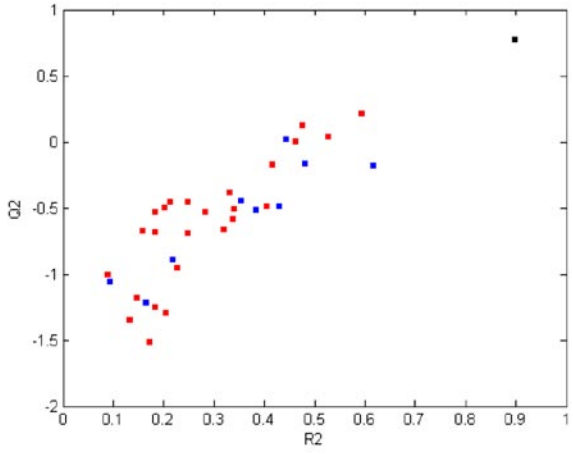

Figure F17. The **y**-randomization plot for the PLS model on data subset 3: black square - the real model, blue squares - 10 randomized models, red squares - 25 randomized models.

Table T32. Comparative statistics of 10 and 25 **y**-randomizations for the PLS model on data subset 3.

Parameter[a]	10 iterations	25 iterations
Maximum (Q^2_{yrand})	0.020	0.218
Maximum (R^2_{yrand})	**0.616**	**0.592**
Standard deviation (Q^2_{yrand})	**0.418**	**0.467**
Standard deviation (R^2_{yrand})	**0.162**	**0.133**
Minimum model-random. diff. (Q^2_{yrand})[b]	**1.82**	**1.20**
Minimum model-random. diff. (R^2_{yrand})[b]	**1.74**	**0.51**
Confidence level for min. diff. (Q^2_{yrand})[c]	**0.069**	**0.230**
Confidence level for min. diff. (R^2_{yrand})[c]	**0.082**	**0.610**
Randomizations %, conf. level > 0.0001 (Q^2_{yrand})[d]	**70%**	**80%**
Randomizations %, conf. level > 0.0001 (R^2_{yrand})[d]	**70%**	**100%**
y-Randomization intercept (r_{yrand} *vs.* Q^2_{yrand})[e]	-0.900	-0.958
y-Randomization intercept (r_{yrand} *vs.* R^2_{yrand})[e]	0.191	0.148

[a]Statistical parameters are calculated for Q^2 from from **y**-randomization (Q^2_{yrand}) and R^2 from **y**-randomization (R^2_{yrand}). Values typed bold represent obvious critical cases. [b]Minimum model-randomizations difference: the difference between the real model (Table 1) and the best **y**-randomization in terms of correlation coefficients Q^2_{yrand} or R^2_{yrand}, expressed in units of the standard deviations of Q^2_{yrand} or R^2_{yrand}, respectively. The best **y**-randomization is defined by the highest Q^2_{rand} or R^2_{rand}. [c]Confidence level for normal distribution of the minimum difference between the real and randomized models. [d]Percentage of randomizations characterized by the difference between the real and randomized models (in terms of Q^2_{yrand} or R^2_{yrand}) at confidence levels > 0.0001. [e]Intercepts obtained from two **y**-randomization plots for each regression model proposed. Q^2_{yrand} or R^2_{yrand} is the vertical axis, whilst the horizontal axis is the absolute value of the correlation coefficient r_{yrand} between the original and randomized vectors **y**. The randomization plots are completed with the data for the real model (r_{yrand} = 1.000, Q^2 or R^2).

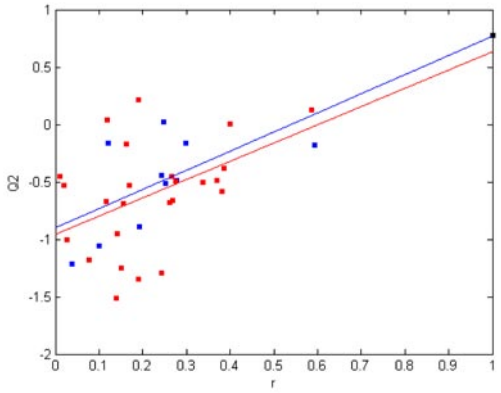

Figure F18. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the PLS model on data subset 3: $Q^2 = -0.900 + 1.667\ r$ and $Q^2 = -0.958 + 1.590\ r$, respectively.

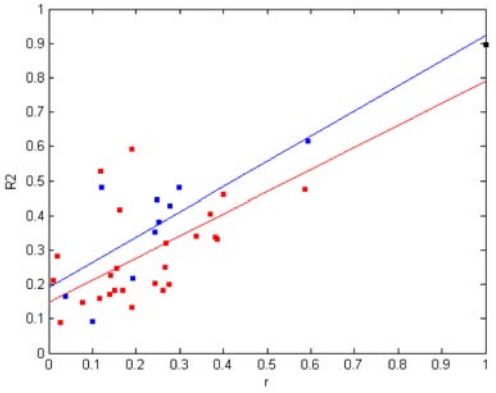

Figure F19. The plot for determining the linear regression equations for 10 (blue) and 25 (red) models from **y**-randomizations of the PLS model on data subset 3: $R^2 = -0.191 + 0.732\ r$ and $R^2 = 0.148 + 0.642\ r$, respectively.

SAMPLES REORDERING EFFECT ON LEAVE-*N*-OUT CROSSVALIDATION

The effect of sample randomization can be noticed when a model is validated by means of two LNO modes: a single LNO using the original (not reordered) data set, and a multiple LNO with the reordered (randomized) data. This is illustrated in the LNO plots for data sets 1 (Figure F20), 2 (Figure F3), 3 (Figure F8) and 4 (Figure F12), and subset 3 (Figure F15). Green curves are from single LNO for original data, compared to red curves for multiple LNO. The green curves do not show randomization effect on LNO for data set 1, probably due to relatively large number of samples and certain statistically insignificant **x** - **y** relationships, and also for subset 3. Data sets 2, 3 and 4, on the contrary, show green curves above the red curves from $N = 1$ to critical N at which Q^2_{LNO} is still stable.

The randomization effect on LNO is more obvious when Pearson correlation coefficients between all variables (descriptors and **y**) and the row index of **X** (*i.e.*, OrdNum, the ordinal number or position of a sample in a data set), are calculated for the original data, randomized data prior to LNO, new randomized data sets for single and multiple LNO (Tables T33 - T37 and Figure F21). This analysis shows significant correlations (absolute values of correlation coefficients over 0.30 are highlighted) between OrdNum and variables in two cases: a) small to moderate original data sets 2, 3, and 4 and subset 3 (15 - 56 samples); and b) small to moderate randomized data sets 3 and 4 and subset 3 (15 - 50 samples). The smaller the data set, the higher is the probability for chance correlation between OrdNum and variables, what results in virtually higher Q^2_{LNO} than in reality. By randomizing the original data, and especially by performing multiple LNO with several runs, one may reduce this chance correlation even for small data sets.

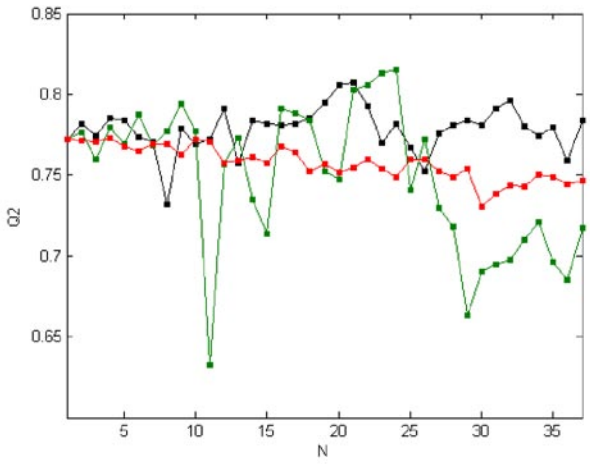

Figure F20. Comparative leave-*N*-out crossvalidation plot for the MLR model on data set 1. Black - single LNO, red - multiple LNO (10 times), and green - single LNO using the original (not randomized) data set.

Table T33. Correlation coefficients between the row index[a] of data set 1 and all variables[b] in single[c] and multiple[d] leave-*N*-out (LNO) crossvalidation runs.

Descriptor	Original	S-LNO	M-LNO-1	M-LNO-2	M-LNO-3	M-LNO-4	M-LNO-5	M-LNO-6	M-LNO-7	M-LNO-8	M-LNO-9	M-LNO-10
LogK_{ow}	–0.2076	0.0832	0.1072	0.0978	0.1571	–0.0886	0.0117	0.0431	0.1765	0.2014	–0.1055	0.1112
pK_a	–0.0761	–0.0937	–0.1219	0.0976	0.1256	–0.0248	0.0111	0.2204	0.1258	0.0695	0.0613	0.0335
E_{LUMO}	–0.1343	–0.0979	–0.1235	0.0970	0.1433	–0.0132	0.1278	0.0768	–0.0115	–0.0930	–0.0901	0.1255
E_{HOMO}	–0.1725	–0.0405	–0.1516	0.0210	0.0925	–0.0614	0.1131	0.1201	–0.0000	–0.0535	–0.0163	0.0931
N_{hdon}	0.0973	0.0680	–0.0453	0.0493	–0.1707	0.2014	0.0360	–0.0960	–0.2787	–0.0747	–0.0827	–0.0493
y	–0.2040	0.0317	0.0988	0.1269	0.1719	–0.1107	0.0226	0.0283	0.2049	0.1877	–0.0885	0.0766
AAR[e]	0.1486	0.0692	0.1080	0.0816	0.1435	0.0833	0.0537	0.0974	0.1329	0.1133	0.0741	0.0815

[a]The row index for **X** or **y** is the OrdNum, the ordinal number or position of a sample in the data set. [b]Molecular and other descriptors and the dependent variable *y*. [c]Single LNO using **X** with originally ordered (Original) and randomly reordered (S-LNO) samples. [d]Ten multiple LNO (M-LNO) run using **X** with randomly re-ordered samples. [e]Average of absolute correlation coefficients.

Table T34. Correlation coefficients[a] between the row index[b] of data set 2 and all variables[c] in single[d] and multiple[e] leave-*N*-out (LNO) crossvalidation runs.

Descriptor	Original	S-LNO	M-LNO-1	M-LNO-2	M-LNO-3	M-LNO-4	M-LNO-5	M-LNO-6	M-LNO-7	M-LNO-8	M-LNO-9	M-LNO-10
CYP51-g	–0.1755	0.0577	–0.1755	0.1755	0.1325	0.1599	0.0193	–0.1046	–0.0834	–0.0427	–0.0600	0.0711
CYP51-e	–0.1805	0.0696	–0.1642	0.1853	0.1165	0.1617	0.0223	–0.1024	–0.0703	–0.0323	–0.0628	0.0876
PMR1-t	–0.0267	–0.0829	0.1481	0.0639	–0.0831	0.1226	0.2091	0.1425	–0.1371	0.0362	–0.1062	0.0850
CYP51-e*Npi	0.1150	–0.1017	–0.1012	0.1380	0.1302	0.1437	–0.0248	–0.1104	–0.0887	–0.1016	–0.0347	0.0585
PCR*Npi	**0.3409**	–0.1957	–0.0578	0.0885	0.1172	0.0856	0.0026	–0.0663	–0.1499	–0.0892	0.0116	0.0363
PMR1-e*Lpi	0.1778	0.0039	0.0032	0.0689	–0.1401	0.0835	0.1316	0.0817	–0.0244	0.0362	–0.1283	–0.1462
CYP51-e*Lpi	0.2141	–0.0161	–0.1623	0.0593	0.0994	0.0598	0.0476	–0.0799	0.0109	–0.1216	0.0018	–0.0637
PCR*Lpi	**0.4860**	–0.0704	–0.1481	0.0131	0.0415	–0.0117	0.0812	–0.0202	–0.0018	–0.1113	0.0360	–0.1308
y	0.0143	–0.0038	0.0932	–0.2035	0.0764	–0.1299	–0.1297	–0.0601	0.1572	–0.0695	0.1218	–0.0827
AAR[f]	0.1923	0.0669	0.1171	0.1107	0.1041	0.1065	0.0742	0.0853	0.0804	0.0712	0.0626	0.0847

[a]Correlation coefficients with absolute value greater than 0.30 are typed bold and indicate chance correlation. [b]The row index for **X** or **y** is the OrdNum, the ordinal number or position of a sample in the data set. [c]Molecular and other descriptors and the dependent variable *y*. [d]Single LNO using **X** with originally ordered (Original) and randomly reordered (S-LNO) samples. [e]Ten multiple LNO (M-LNO) run using **X** with randomly re-ordered samples. [f]Average of absolute correlation coefficients.

Table T35. Correlation coefficients[a] between the row index[b] of data set 3 and all variables[c] in single[d] and multiple[e] leave-*N*-out (LNO) crossvalidation runs.

Descriptor	Original	S-LNO	M-LNO-1	M-LNO-2	M-LNO-3	M-LNO-4	M-LNO-5	M-LNO-6	M-LNO-7	M-LNO-8	M-LNO-9	M-LNO-10
E_e	**0.5076**	0.0949	0.1407	0.0575	–0.1888	0.0635	0.0045	–0.0507	0.1439	0.1214	0.1908	–0.1330
E_{CC}	**0.3455**	0.0906	0.1727	0.0167	–0.1716	0.0840	0.0846	–0.0124	0.1594	0.0196	0.2338	–0.0371
Δ_{HL}	**0.6466**	–0.0210	**0.3903**	0.0605	–0.1444	–0.0356	0.2341	0.0818	0.0718	0.0414	0.1616	–0.2051
σ_b	**–0.3140**	–0.0831	–0.1223	–0.0013	0.1629	–0.1262	–0.1400	0.0020	–0.1047	–0.0144	–0.2263	–0.0052
σ_r	**0.5748**	0.1643	**0.3680**	0.1362	–0.2463	–0.0196	0.1012	0.0084	0.1725	0.0745	0.1815	–0.2277
D_{CC}	**–0.3603**	–0.0880	–0.1727	–0.0157	0.1691	–0.1063	–0.0894	0.0176	–0.1446	–0.0101	–0.2385	0.0382
Q_{C2mul}	**–0.4845**	–0.1368	–0.1830	–0.0760	0.2348	–0.1148	–0.0253	0.0306	–0.2149	–0.0868	–0.1665	0.1153
Q_{Omul}	**–0.4416**	–0.1062	–0.2009	–0.0430	0.2013	–0.1228	–0.1201	0.0078	–0.1432	–0.0805	–0.2272	0.0694
y	**–0.4989**	–0.0814	–0.2411	–0.0496	0.2182	–0.0617	–0.1332	–0.0034	–0.1767	–0.1379	**–0.3076**	0.1470
AAR[f]	**0.4638**	0.0963	0.2213	0.0507	0.1930	0.0816	0.1036	0.0239	0.1480	0.0652	0.2149	0.1087

[a]Correlation coefficients with absolute value greater than 0.30 are typed bold and indicate chance correlation. [b]The row index for **X** or **y** is the OrdNum, the ordinal number or position of a sample in the data set. [c]Molecular and other descriptors and the dependent variable *y*. [d]Single LNO using **X** with originally ordered (Original) and randomly reordered (S-LNO) samples. [e]Ten multiple LNO (M-LNO) run using **X** with randomly re-ordered samples. [f]Average of absolute correlation coefficients.

Table T36. Correlation coefficients[a] between the row index[b] of data subset 3 and all variables[c] in single[d] and multiple[e] leave-*N*-out (LNO) crossvalidation runs.

Descriptor	Original	S-LNO	M-LNO-1	M-LNO-2	M-LNO-3	M-LNO-4	M-LNO-5	M-LNO-6	M-LNO-7	M-LNO-8	M-LNO-9	M-LNO-10
E_e	**0.6074**	–0.1385	–0.0138	**–0.3049**	**0.3157**	–0.1725	–0.2515	**0.4093**	–0.1385	–0.1385	–0.2413	**0.6744**
E_{CC}	**0.4509**	–0.0597	–0.2841	–0.1276	**0.4816**	0.0622	0.0004	0.2778	–0.0597	–0.0597	**–0.3364**	**0.4872**
Δ_{HL}	**0.5795**	–0.2181	–0.1522	–0.2237	0.1677	–0.0861	–0.1166	**0.4812**	–0.2181	–0.2181	**–0.4683**	**0.4326**
σ_b	**–0.4701**	0.0886	0.2385	0.2657	**–0.5246**	–0.1295	–0.0477	–0.2793	0.0886	0.0886	0.2317	**–0.4565**
σ_r	**0.5803**	0.1950	**–0.4131**	0.0696	**0.3157**	–0.0696	–0.2878	**0.5013**	0.1950	0.1950	**–0.6777**	**0.6267**
D_{CC}	**–0.4584**	0.0673	0.2586	0.1787	**–0.4668**	–0.0778	0.0021	**–0.3028**	0.0673	0.0673	**0.3091**	**–0.5068**
Q_{C2mul}	**–0.5163**	0.0910	0.2006	0.1692	**–0.5001**	–0.0444	0.0560	–0.2107	0.0910	0.0910	0.2682	**–0.6044**
Q_{Omul}	**–0.5114**	0.0611	0.2646	0.1746	**–0.4950**	–0.1096	0.0243	**–0.3742**	0.0611	0.0611	**0.3040**	**–0.5541**
y	**–0.5971**	0.0973	0.2155	0.1517	**–0.4268**	–0.0875	0.1800	**–0.3810**	**0.3282**	**0.5054**	**0.3307**	**–0.6955**
AAR[f]	**0.5302**	0.1130	0.2268	0.1851	**0.4104**	0.0933	0.1074	**0.3575**	0.1386	0.1583	**0.3519**	**0.5598**

[a]Correlation coefficients with absolute value greater than 0.30 are typed bold and indicate chance correlation. [b]The row index for **X** or **y** is the OrdNum, the ordinal number or position of a sample in the data set. [c]Molecular and other descriptors and the dependent variable *y*. [d]Single LNO using **X** with originally ordered (Original) and randomly reordered (S-LNO) samples. [e]Ten multiple LNO (M-LNO) run using **X** with randomly re-ordered samples. [f]Average of absolute correlation coefficients.

Table T37. Correlation coefficients[a] between the row index[b] of data set 4 and all variables[c] in single[d] and multiple[e] leave-*N*-out (LNO) crossvalidation runs.

Descriptor	Original	S-LNO	M-LNO-1	M-LNO-2	M-LNO-3	M-LNO-4	M-LNO-5	M-LNO-6	M-LNO-7	M-LNO-8	M-LNO-9	M-LNO-10
ClogP	**0.3623**	–0.0646	0.1501	–0.2715	**0.3010**	0.1021	–0.0234	0.1319	0.0371	–0.0141	**–0.5656**	0.0694
MgVol	**0.5268**	–0.2474	**–0.5023**	0.0793	**0.4343**	**0.4078**	0.2134	0.0831	–0.1529	0.1416	**–0.3588**	0.2681
$B1_{X.2}$	0.0482	–0.1585	0.0012	–0.1826	**–0.3049**	–0.0042	0.0657	0.0970	0.1525	–0.1741	0.2139	–0.2139
y	–0.0244	0.2429	**0.6429**	–0.2265	0.0638	–0.0759	–0.1541	0.0986	0.1250	–0.0460	**–0.5194**	–0.1308
AAR[f]	0.2404	0.1784	**0.3241**	0.1900	0.2760	0.1475	0.1142	0.1026	0.1169	0.0939	**0.4144**	0.1706

[a]Correlation coefficients with absolute value greater than 0.30 are typed bold and indicate chance correlation. [b]The row index for **X** or **y** is the OrdNum, the ordinal number or position of a sample in the data set. [c]Molecular and other descriptors and the dependent variable *y*. [d]Single LNO using **X** with originally ordered (Original) and randomly reordered (S-LNO) samples. [e]Ten multiple LNO (M-LNO) run using **X** with randomly re-ordered samples. [f]Average of absolute correlation coefficients.

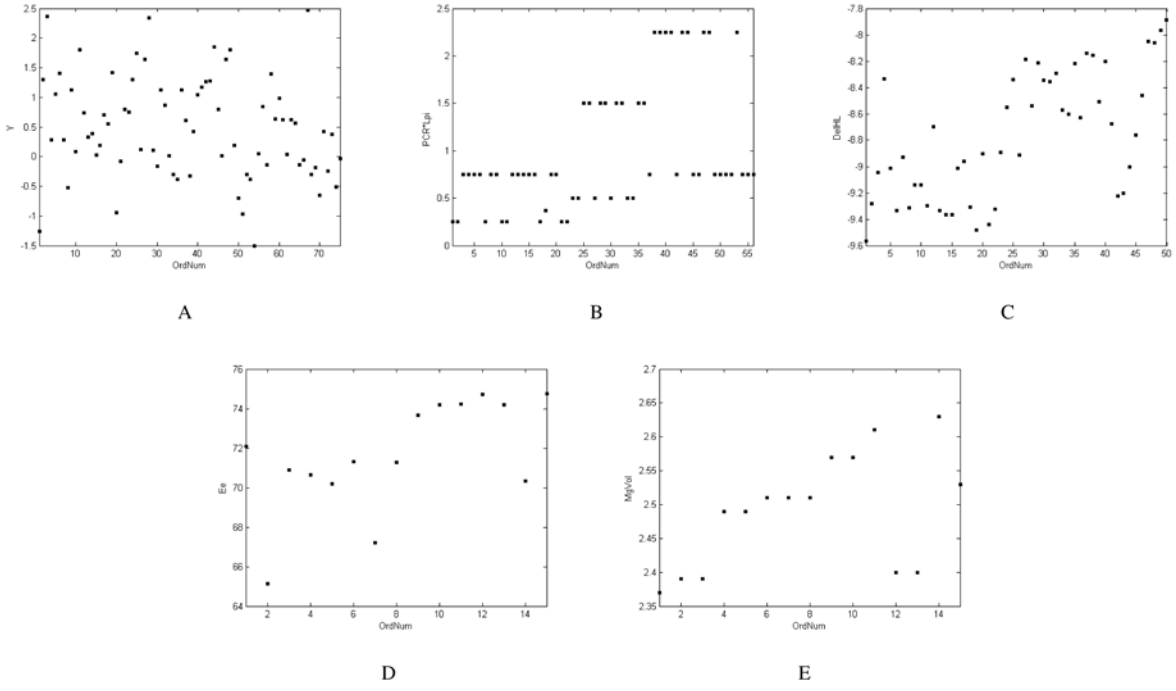

Figure F21. Representative chance correlations between the row index (OrdNum) and variables in the data sets with originally ordered samples: A) data set 1, B) data set 2, C) data set 3, D) data subset 3, and E) data set 4

SAMPLES REORDERING EFFECT ON y-RANDOMIZATION

When the same correlation analysis from LNO is performed for 10 + 25 data sets from **y**-randomization, which were obtained first by reordering (randomizing) the samples in the whole set and *a posteriori* randomizing only **y**, similar conclusions can be drawn (Table T38). Larger data sets 1 and 2 (75 and 50 samples) do not show significant correlation coefficients, data set 3 (50 samples) possesses one critical correlation, whilst data set 4 and subset 3 (15 samples) have ten correlation coefficients above 0.30.

When using the original instead of reordered data for 10 **y**-randomizations, the results show that there is no effect for larger sets 1 (Figure F22) and 2 (Figure F23), but is visible for smaller set 3 (Figure F24), subset 3 (Figure F25) and set 4 (Figure F26). The effect can be noticed as a shift of Q^2 and R^2 towards higher values relative to the normal **y**-randomization procedure.

Table T38. Correlation coefficients between **y** and OrdNum[a] in **y**-randomization.[b*]

	Set 1	Set 2	Set 3	Subset 3	Set 4
Original[c]	–0.2040	0.0143	**–0.4989**	**–0.5971**	–0.0244
Re-ordered[d]	0.0317	–0.0038	–0.0814	0.0973	0.2429
y-R10-1	0.1322	–0.1080	–0.2216	–0.0358	0.2452
y-R10-2	–0.1867	0.0140	–0.1235	0.0261	–0.0655
y-R10-3	0.1668	0.0479	–0.0663	–0.2572	**0.3567**
y-R10-4	0.0899	0.2588	0.0391	–0.0428	–0.1477
y-R10-5	–0.2235	0.2478	0.2140	–0.0098	0.0957
y-R10-6	0.0499	0.1844	–0.2932	**–0.5810**	0.1144
y-R10-7	0.0165	–0.0529	–0.1966	**0.3282**	0.1236
y-R10-8	–0.2058	0.0268	0.2203	**0.5054**	**0.3584**
y-R10-9	0.0441	–0.0935	–0.1300	–0.1778	0.1247
y-R10-10	–0.2310	–0.0352	0.1378	0.1291	–0.1359
y-R25-1	–0.1770	0.0550	0.0998	**–0.4141**	–0.0897
y-R25-2	–0.2189	–0.0423	–0.0054	0.0463	0.1377
y-R25-3	–0.1343	–0.0963	–0.1591	–0.1252	**0.4300**
y-R25-4	0.1972	–0.1300	0.0004	**–0.3551**	0.0362
y-R25-5	0.1405	–0.0487	0.1384	**–0.5050**	**–0.3774**
y-R25-6	–0.1792	0.0220	0.0531	0.0406	**–0.3354**
y-R25-7	–0.0087	–0.0124	–0.1452	**–0.5235**	0.1466
y-R25-8	–0.0691	–0.1661	**0.2999**	–0.1264	–0.0635
y-R25-9	0.0850	0.1182	–0.0999	0.2671	–0.1247
y-R25-10	0.2185	0.1239	–0.2131	0.1948	**0.3627**
y-R25-11	–0.1041	0.0767	0.1457	0.2815	0.1296
y-R25-12	–0.0937	0.1939	0.0193	–0.0004	**–0.3150**
y-R25-13	–0.1239	–0.0477	–0.1114	–0.1422	0.2440
y-R25-14	–0.0398	–0.0852	0.1172	0.2377	**0.5087**
y-R25-15	0.0653	–0.0496	–0.1009	**0.3419**	**–0.3966**
y-R25-16	–0.0387	0.0372	–0.2133	0.1810	–0.0957
y-R25-17	0.0280	–0.1866	–0.2335	0.0156	0.2558
y-R25-18	–0.2136	–0.0252	–0.0933	0.2555	–0.0463
y-R25-19	–0.1357	–0.0046	–0.1250	**–0.6297**	–0.0092
y-R25-20	0.1065	–0.0857	0.0651	**–0.3976**	–0.2409
y-R25-21	–0.0856	–0.0338	–0.0650	0.2878	–0.2365
y-R25-22	–0.0341	–0.0072	0.2350	**0.5616**	–0.2863
y-R25-23	–0.0363	–0.0059	0.2447	0.0034	0.2891
y-R25-24	–0.0628	0.0814	–0.2725	0.1871	**0.3090**
y-R25-25	–0.0180	0.0359	–0.0407	–0.0548	–0.1917
yR-AbsAver[e]	0.1132	0.0812	0.1411	0.2363	0.2122

[a]Pearson correlation coefficients are calculated for all 10 (y-R10) and 25 (y-R25) **y**-randomizations. [b]OrdNum, the ordinal number or position of a sample in the data set, is the row index for **X** or **y** is the OrdNum. [c]Originally ordered samples in a data set. [d]Re-ordered samples in a data set. [e]Average absolute value for all 10 + 25 **y**-randomizations. [*]All correlation coefficients greater than 0.30 are typed bold.

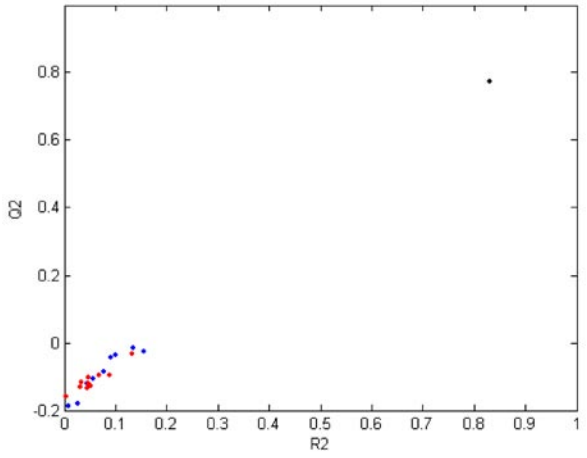

Figure F22. The effect of sample reordering to **y**-randomization plot for the MLR model on data set 1. Black ball - the real model, blue balls - 10 randomized models with re-ordered samples in **X**, and red balls - 10 randomized models with originally ordered samples in **X**. The studied effect cannot be observed.

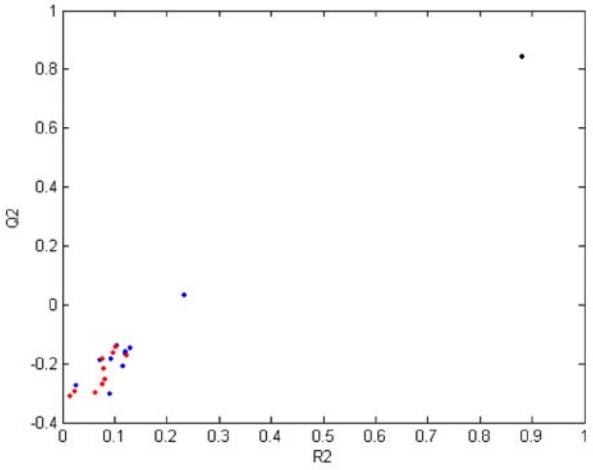

Figure F23. The effect of sample reordering to **y**-randomization plot for the PLS model on data set 2. Black ball - the real model, blue balls - 10 randomized models with re-ordered samples in **X**, and red balls - 10 randomized models with originally ordered samples in **X**. The studied effect cannot be observed.

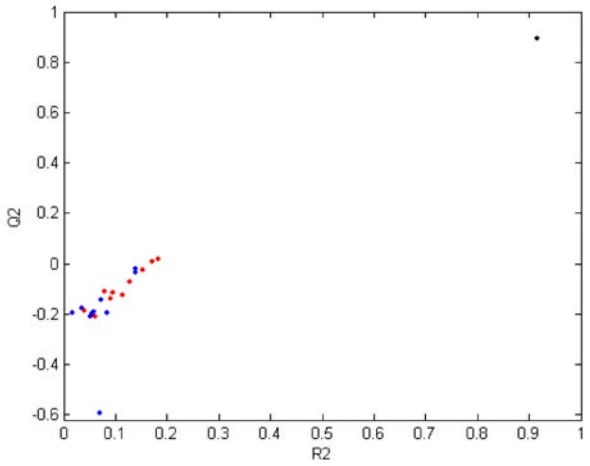

Figure F24. The effect of sample reordering to **y**-randomization plot for the PLS model on data set 3. Black ball - the real model, blue balls - 10 randomized models with re-ordered samples in **X**, and red balls - 10 randomized models with originally ordered samples in **X**. The studied effect can be noticed as the systematic shift of Q^2 and R^2 towards higher values.

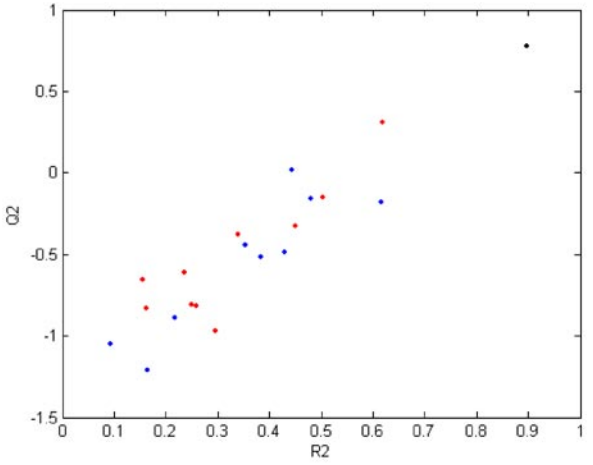

Figure F25. The effect of sample reordering to **y**-randomization plot for the PLS model on data subset 3. Black ball - the real model, blue balls - 10 randomized models with re-ordered samples in **X**, and red balls - 10 randomized models with originally ordered samples in **X**. The studied effect can be noticed as the systematic shift of Q^2 and R^2 towards higher values.

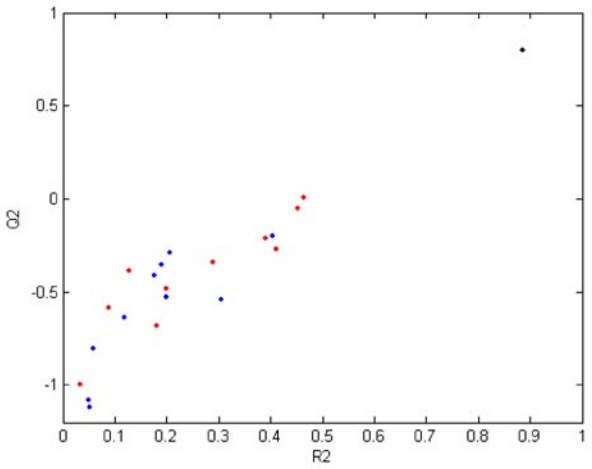

Figure F26. The effect of sample reordering to **y**-randomization plot for the MLR model on data set 4. Black ball - the real model, blue balls - 10 randomized models with re-ordered samples in **X**, and red balls - 10 randomized models with originally ordered samples in **X**. The studied effect can be noticed as the systematic shift of Q^2 and R^2 towards higher values.

ANALYSIS OF x - y RELATIONSHIPS

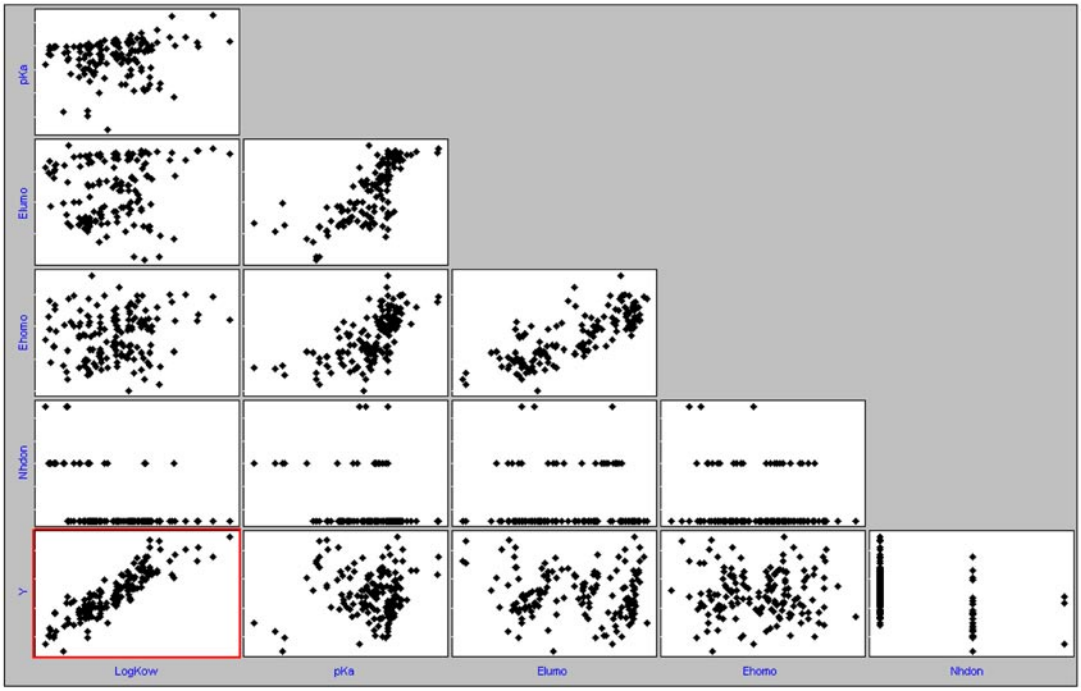

Figure F27. The scatterplots for the complete data set 1 (153 samples). The best **x** - **y** correlation profile is marked in red.

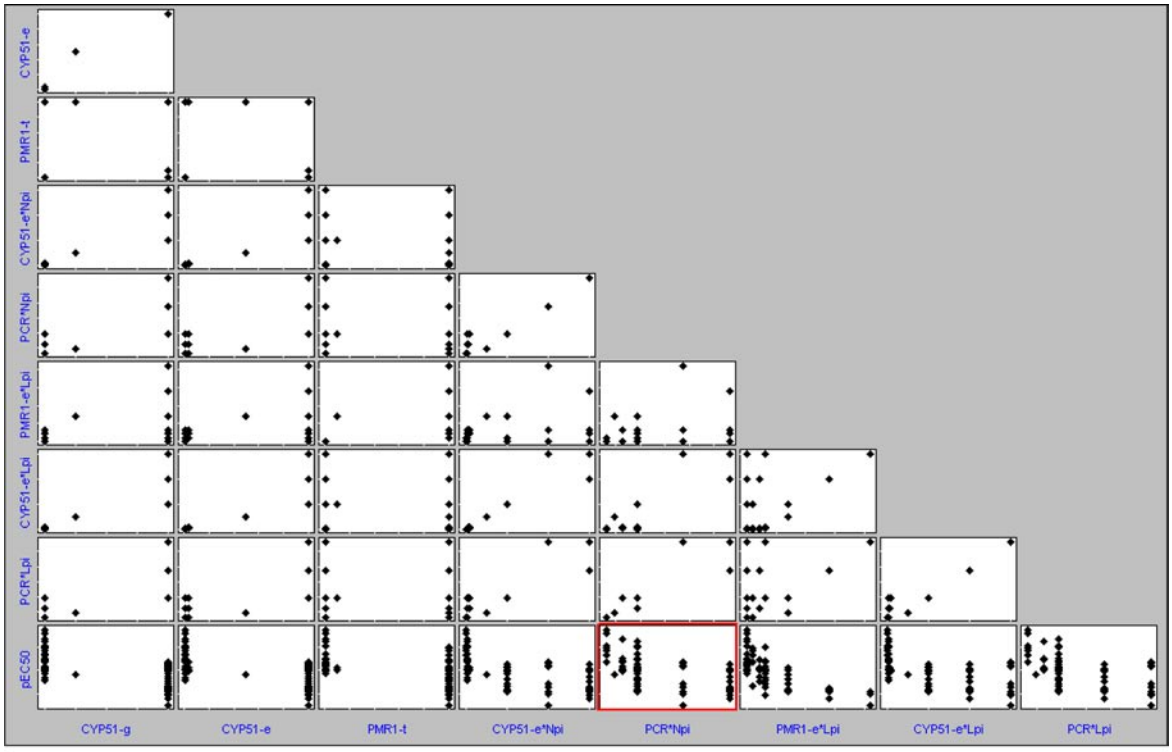

Figure F28. The scatterplots for the complete data set 2 (86 samples). The best **x** - **y** correlation profile is marked in red.

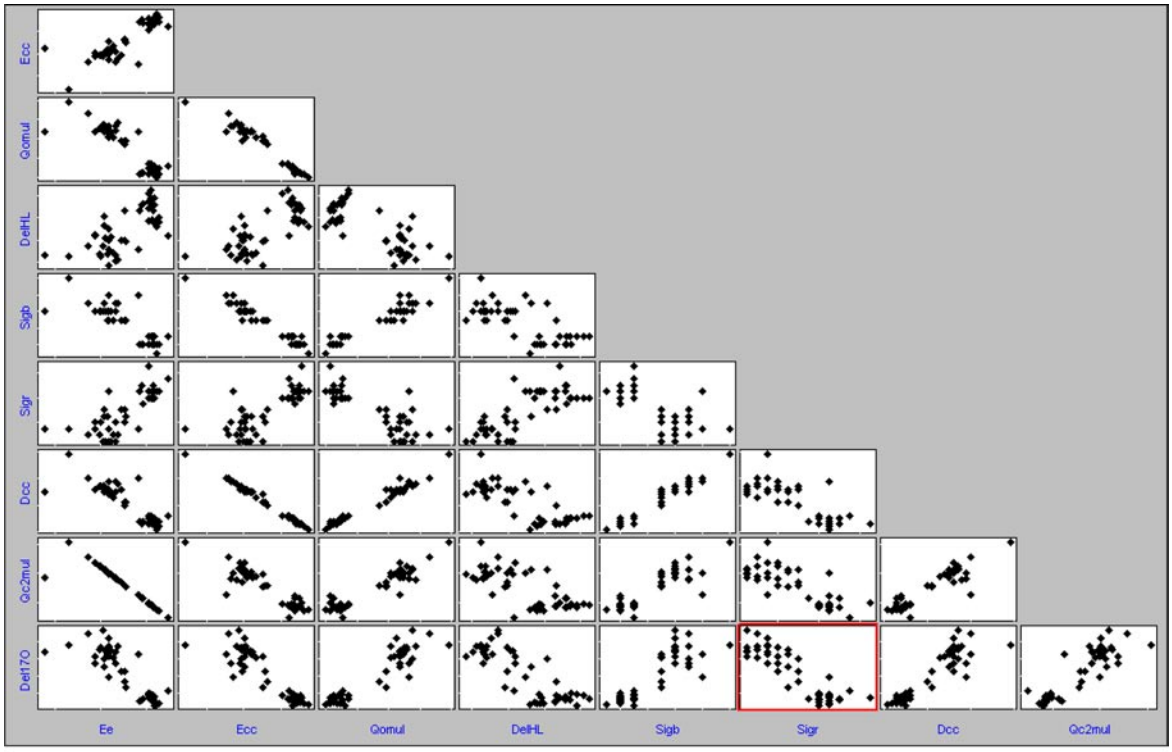

Figure F29. The scatterplots for the complete data set 3 (50 samples). The best **x** - **y** correlation profile is marked in red.

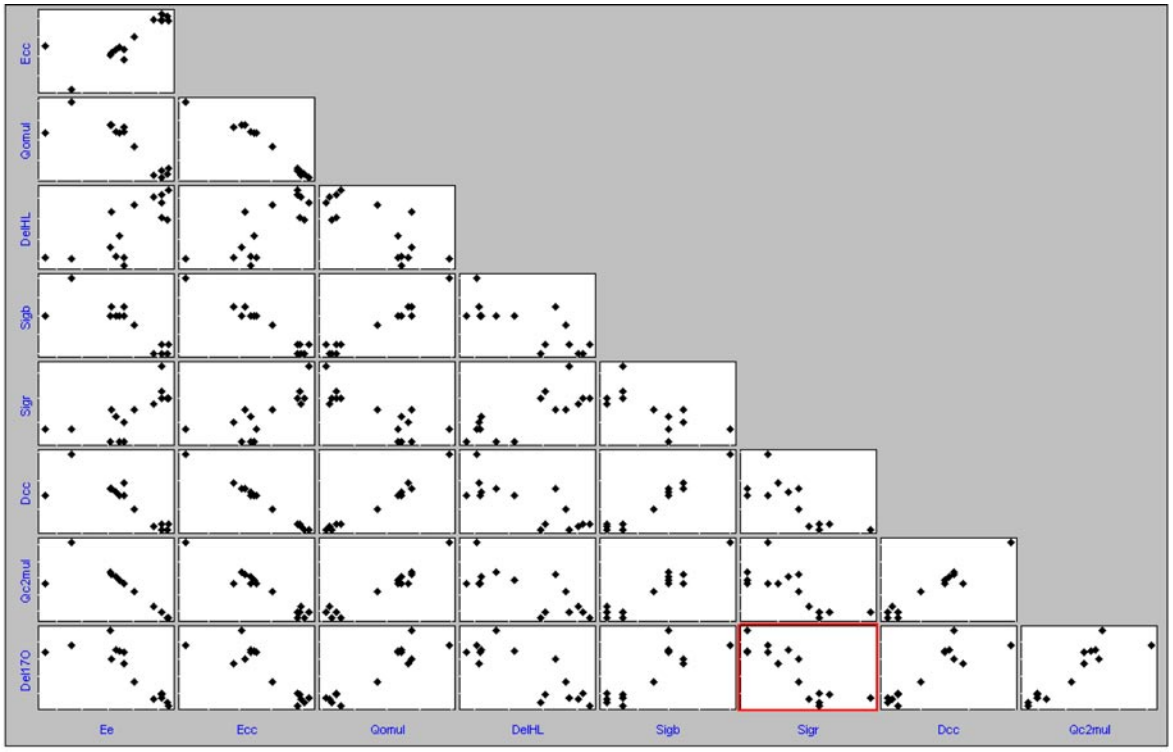

Figure F30. The scatterplots for the complete data subset 3 (15 samples). The best **x** - **y** correlation profile is marked in red.

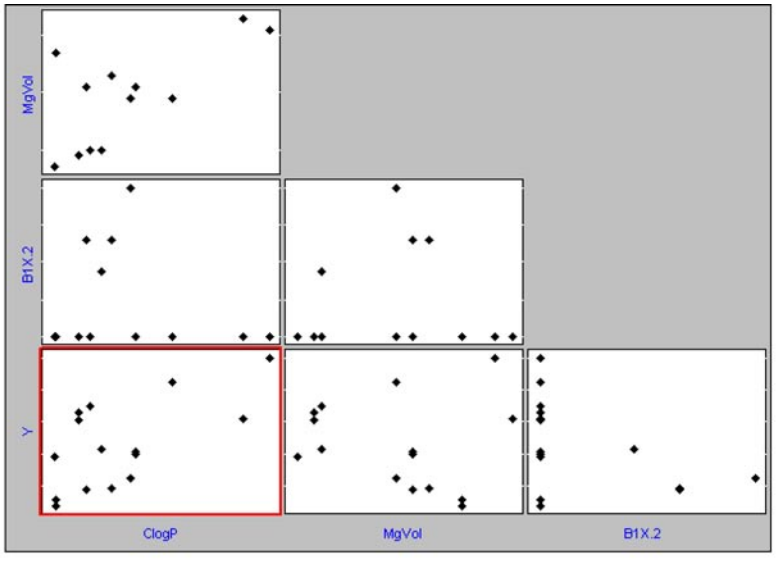

Figure F31. The scatterplots for the complete data set 4 (15 samples). The best **x** - **y** correlation profile is marked in red.

Table T39. Statistics for univariate descriptor - dependent variable (**x** - **y**) relationships in all the data sets and subsets studied.

Data set	Subset	n[a]	Descriptor[b]	R[c]	a (σ_a)[d]	b (σ_b)[d]	$t_{a,n-1}$ (p_a)[e]	$t_{b,n-1}$ (p_b)[e]	R^2[f]	$F_{1,n-2}$ (p)[g]
1	Complete set	153	LogK_{ow} [+]	0.881	–1.02(7)	0.63(3)	–14.0 (<0.0001)	22.8 (<0.0001)	0.776	522.0 (<0.0001)
			pK_a [+]	**0.025**	**0.4(4)**	**0.01(4)**	**0.9 (0.36)**	**0.3 (0.76)**	**0.0006**	**0.1 (0.76)**
			E_{LUMO} [–]	**–0.100**	**0.49(6)**	**–0.2(2)**	7.6 (<0.0001)	**1.2 (0.22)**	**0.010**	**1.5 (0.22)**
			E_{HOMO} [+]	**–0.006**	**0.4(20)**	**–0.02(22)**	**0.2 (0.86)**	**0.07 (0.94)**	**0.00003**	**0.005 (0.94)**
			N_{hdon} [+]	–0.410	1.4(2)	–0.7(1)	8.1 (<0.0001)	5.5 (<0.0001)	0.168	30.5 (<0.0001)
	Training set	75	LogK_{ow} [+]	0.890	–1.1(1)	0.64(4)	10.3 (<0.0001)	16.6 (<0.0001)	0.792	277.0 (<0.0001)
			pK_a [+]	**0.082**	**0.1(6)**	**0.04(6)**	**0.2 (0.86)**	**0.7 (0.48)**	**0.007**	**0.5 (0.48)**
			E_{LUMO} [–]	**–0.106**	**0.5(1)**	**–0.2(3)**	4.9 (<0.0001)	**0.9 (0.36)**	**0.011**	**0.8 (0.36)**
			E_{HOMO} [+]	**0.021**	**1.0(29)**	**0.3(3)**	**0.4 (0.73)**	**0.2 (0.86)**	**0.0004**	**0.03 (0.86)**
			N_{hdon} [+]	–0.408	1.3(2)	–0.7(2)	5.8 (<0.0001)	3.9 (0.0002)	0.170	14.9 (0.0002)
	External validation set	78	LogK_{ow} [+]	0.871	–0.97(10)	0.62(4)	9.3 (<0.0001)	15.5 (<0.0001)	0.759	238.8 (<0.0001)
			pK_a [+]	**–0.047**	**0.7(6)**	**–0.03(6)**	**1.3 (0.21)**	**0.4 (0.68)**	**0.002**	**0.2 (0.68)**
			E_{LUMO} [–]	**–0.093**	0.49(9)	**–0.2(2)**	5.8 (<0.0001)	**0.8 (0.42)**	**0.009**	**0.7 (0.42)**
			E_{HOMO} [+]	**–0.039**	**–0.4(28)**	**–0.1(3)**	**0.2 (0.87)**	**0.3 (0.74)**	**0.002**	**0.1 (0.74)**
			N_{hdon} [+]	–0.408	1.3(2)	–0.7(2)	5.8 (<0.0001)	3.9 (0.0002)	0.166	15.2 (0.0002)
2	Complete set	86	CYP51-g [–]	–0.722	6.9(1)	–0.27(3)	66.4 (<0.0001)	9.6 (<0.0001)	0.522	91.6 (<0.0001)
			CYP51-e [–]	–0.726	6.61(8)	–0.011(1)	80.7 (<0.0001)	9.7 (<0.0001)	0.528	93.8 (<0.0001)
			PMR1-t [–]	–0.502	6.6(1)	–0.008(2)	53.8 (<0.0001)	5.3 (<0.0001)	0.252	28.3 (<0.0001)
			CYP51-e*Npi [+]	–0.624	6.45(8)	–0.004(1)	76.0 (<0.0001)	7.3 (<0.0001)	0.389	53.5 (<0.0001)
			PCR*Npi [–]	–0.556	6.7(1)	–0.6(1)	54.2 (<0.0001)	6.1 (<0.0001)	0.309	37.5 (<0.0001)
			PMR1-e*Lpi [–]	–0.678	6.39(7)	–0.09(1)	87.3 (<0.0001)	8.4 (<0.0001)	0.459	71.3 (<0.0001)
			CYP51-e*Lpi [+]	–0.634	6.45(8)	–0.004(1)	76.9 (<0.0001)	7.5 (<0.0001)	0.402	56.4 (<0.0001)
			PCR*Lpi [–]	–0.564	6.7(1)	–0.6(1)	55.2 (<0.0001)	6.3 (<0.0001)	0.318	39.1 (<0.0001)
	Training set	56	CYP51-g [–]	–0.721	6.9(1)	–0.27(4)	50.7 (<0.0001)	7.6 (<0.0001)	0.519	58.3 (<0.0001)
			CYP51-e [–]	–0.726	6.6(1)	–0.011(2)	61.6 (<0.0001)	7.8 (<0.0001)	0.528	60.3 (<0.0001)
			PMR1-t [–]	–0.461	6.5(2)	–0.008(2)	39.6 (<0.0001)	3.8 (0.0003)	0.213	14.6 (<0.0001)
			CYP51-e*Npi [+]	–0.631	6.5(1)	–0.004(1)	58.7 (<0.0001)	6.0 (<0.0001)	0.398	35.7 (<0.0001)
			PCR*Npi [–]	–0.568	6.7(2)	–0.7(1)	42.4 (<0.0001)	5.1 (<0.0001)	0.322	25.7 (<0.0001)
			PMR1-e*Lpi [–]	–0.690	6.40(9)	–0.08(1)	68.9 (<0.0001)	7.0 (<0.0001)	0.476	49.0 (<0.0001)
			CYP51-e*Lpi [+]	–0.661	6.5(1)	–0.005(1)	60.9 (<0.0001)	6.5 (<0.0001)	0.437	42.0 (<0.0001)
			PCR*Lpi [–]	–0.590	6.7(2)	–0.7(1)	43.9 (<0.0001)	5.4 (<0.0001)	0.348	28.8 (<0.0001)
	External validation set	30	CYP51-g	–0.729	6.8(2)	–0.25(4)	42.4 (<0.0001)	5.6 (<0.0001)	0.531	31.7 (<0.0001)
			CYP51-e	–0.729	6.6(1)	–0.010(2)	51.6 (<0.0001)	5.6 (<0.0001)	0.531	31.7 (<0.0001)
			PMR1-t	–0.587	6.6(2)	–0.009(2)	37.3 (<0.0001)	3.8 (0.0006)	0.344	14.7 (0.0007)
			CYP51-e*Npi	–0.609	6.4(1)	–0.004(1)	47.3 (<0.0001)	4.1 (0.0003)	0.371	16.5 (0.0004)
			PCR*Npi	–0.529	6.6(2)	–0.6(2)	32.9 (<0.0001)	3.3 (0.003)	0.280	10.9 (0.003)
			PMR1-e*Lpi	–0.656	6.4(1)	–0.10(2)	51.8 (<0.0001)	4.6 (<0.0001)	0.430	21.1 (<0.0001)
			CYP51-e*Lpi	–0.579	6.4(1)	–0.004(1)	46.1 (<0.0001)	3.8 (0.0008)	0.336	14.1 (0.0008)
			CYP51-g	–0.510	6.6(2)	–0.5(2)	32.7 (<0.0001)	3.1 (0.004)	0.260	9.8 (0.004)
3	Complete set	50	E_e [–]	–0.856	1289(65)	–10.4(9)	19.8 (<0.0001)	11.5 (<0.0001)	0.733	131.7 (<0.0001)
			E_{CC} [–]	–0.892	8592(589)	–66(5)	14.6 (<0.0001)	13.7 (<0.0001)	0.796	187.0 (<0.0001)
			Δ_{HL} [–]	–0.827	111(42)	–49(5)	2.6 (0.01)	10.2 (<0.0001)	0.683	103.6 (<0.0001)
			σ_b [+]	0.862	–2650(69)	11437(971)	3.9 (0.0003)	11.8 (<0.0001)	0.743	138.7 (<0.0001)
			σ_r [–]	–0.891	602(5)	–7799(575)	123.9 (<0.0001)	13.6 (<0.0001)	0.793	183.9 (<0.0001)
			D_{CC} [+]	0.907	–5131(380)	3830(257)	13.5 (<0.0001)	14.9 (<0.0001)	0.823	222.4 (<0.0001)
			Q_{C2mul} [+]	0.892	641(8)	402(29)	85.0 (<0.0001)	13.6 (<0.0001)	0.795	185.9 (<0.0001)
			Q_{Omul} [+]	0.928	938(23)	1205(70)	40.8 (<0.0001)	17.3 (<0.0001)	0.862	298.8 (<0.0001)
	Training set	40	E_e [–]	–0.851	1250(71)	–10(1)	17.7 (<0.0001)	10.0 (<0.0001)	0.725	100.0 (<0.0001)
			E_{CC} [–]	–0.874	8295(700)	–63(6)	11.9 (<0.0001)	11.1 (<0.0001)	0.764	122.8 (<0.0001)
			Δ_{HL} [–]	–0.849	132(42)	–47(5)	3.2 (0.003)	9.9 (<0.0001)	0.721	98.1 (<0.0001)
			σ_b [+]	0.846	–236(80)	11020(1127)	3.0 (0.09)	9.8 (<0.0001)	0.716	95.7 (<0.0001)

Table T39. continuation

Data set	Subset	n[a]	Descriptor[b]	R[c]	a (σ_a)[d]	b (σ_b)[d]	$t_{a,n-1}$ (p_a)[e]	$t_{b,n-1}$ (p_b)[e]	R^2[f]	$F_{1,n-2}$ (p)[g]
3	Training set	40	σ_r [–]	–0.891	601(5)	–7674(635)	115.9 (<0.0001)	12.1 (<0.0001)	0.794	146.1 (<0.0001)
			D_{CC} [+]	0.894	–4954(448)	3710(302)	11.1 (<0.0001)	12.3 (<0.0001)	0.799	150.7 (<0.0001)
			Q_{C2mul} [+]	0.892	638(8)	392(32)	79.1 (<0.0001)	12.1 (<0.0001)	0.795	147.3 (<0.0001)
			Q_{Omul} [+]	0.915	921(30)	1152(82)	34.1 (<0.0001)	14.0 (<0.0001)	0.838	196.5 (<0.0001)
	External validation set	10	E_e	–0.901	1548(172)	–14(2)	9.0 (<0.0001)	5.9 (0.0002)	0.812	34.7 (0.0004)
			E_{CC}	–0.945	9715(1129)	–75(9)	8.6 (<0.0001)	8.1 (<0.0001)	0.892	66.1 (<0.0001)
			Δ_{HL}	–0.779	**11(149)**	–60(17)	**0.1 (0.95)**	3.5 (0.007)	0.607	12.4 (0.008)
			σ_b	0.905	–367(150)	12904(2146)	2.4 (0.04)	6.0 (0.0002)	0.819	36.2 (0.0003)
			σ_r	–0.889	610(15)	–8437(1537)	41.7 (<0.0001)	5.5 (0.0004)	0.790	30.1 (0.0006)
			D_{CC}	0.946	–5817(773)	4295(522)	7.5 (<0.0001)	8.2 (<0.0001)	0.895	67.6 (<0.0001)
			Q_{C2mul}	0.901	661(22)	464(79)	30.0 (<0.0001)	5.9 (0.0002)	0.812	34.6 (0.0004)
			Q_{Omul}	0.978	1021(37)	1444(109)	27.8 (<0.0001)	13.3 (<0.0001)	0.957	176.0 (<0.0001)
4	Complete set	15	ClogP [+]	0.632	3(1)	0.9(3)	3.0 (0.009)	2.9 (0.011)	0.400	8.7 (0.01)
			MgVol [–]	**–0.107**	**8(5)**	**–0.7(18)**	**1.8 (0.10)**	**0.4 (0.70)**	**0.012**	**0.2 (0.70)**
			B1$_{X,2}$ [–]	–0.451	7.4(6)	**–0.9(5)**	12.1 (<0.0001)	**1.8 (0.09)**	0.203	**3.3 (0.09)**
	Training set	13	ClogP [+]	0.698	3(1)	1.0(3)	2.7 (0.02)	3.2 (0.007)	0.487	10.4 (0.008)
			MgVol [–]	–0.041	7(5)	–0.3(21)	1.4 (0.20)	0.1 (0.89)	0.002	0.02 (0.89)
			B1$_{X,2}$ [–]	–0.449	7.4(7)	**–0.9(6)**	10.9 (<0.0001)	**1.7 (0.12)**	0.201	**2.8 (0.12)**
Sub-1	Complete set	15	LogK_{ow} [+]	0.861	–1.0(3)	0.6(1)	3.6 (0.003)	6.1 (<0.0001)	0.741	37.1 (<0.0001)
			pK_a [+]	**0.115**	**–0.04(130)**	**0.06(14)**	**0.03 (0.97)**	**0.4 (0.68)**	**0.013**	**0.2 (0.68)**
			E_{LUMO} [–]	**–0.174**	**0.4(3)**	**–0.4(6)**	**1.7 (0.1)**	**0.6 (0.53)**	**0.030**	**0.4 (0.53)**
			E_{HOMO} [–]	**0.017**	**0.9(69)**	**0.05(75)**	**0.13 (0.90)**	**0.06 (0.95)**	**0.0003**	**0.004 (0.95)**
			N_{hdon} [–]	–0.416	1.4(6)	**–0.7(4)**	2.3 (0.04)	**1.6 (0.12)**	0.173	**2.7 (0.12)**
Sub-3	Complete set	15	E_e [–]	–0.816	1166(122)	-9(2)	9.5 (<0.0001)	5.1 (0.0002)	0.666	25.9 (0.0002)
			E_{CC} [–]	–0.859	7650(1176)	–58(10)	6.5 (<0.0001)	6.0 (<0.0001)	0.737	36.5 (<0.0001)
			Δ_{HL} [–]	–0.827	**128(79)**	–48(9)	**1.6 (0.13)**	5.3 (0.0001)	0.683	28.0 (0.0001)
			σ_b [+]	0.846	**–219(134)**	10771(1884)	**1.6 (0.12)**	5.7 (<0.0001)	0.716	32.7 (<0.0001)
			σ_r [–]	–0.855	601(10)	–7417(1248)	58.5 (<0.0001)	5.9 (<0.0001)	0.731	35.3 (<0.0001)
			D_{CC} [+]	0.872	–4532(791)	3426(534)	5.7 (<0.0001)	6.4 (<0.0001)	0.760	41.2 (<0.0001)
			Q_{C2mul} [+]	0.894	643(14)	412(57)	45.6 (<0.0001)	7.2 (<0.0001)	0.800	51.9 (<0.0001)
			Q_{Omul} [+]	0.920	913(43)	1129(133)	21.0 (<0.0001)	8.5 (<0.0001)	0.847	72.1 (<0.0001)
	Training set	13	E_e [–]	–0.814	1151(131)	–9(2)	8.8 (<0.0001)	4.6 (0.0006)	0.662	21.5 (0.0007)
			E_{CC} [–]	–0.857	7499(1264)	–57(10)	5.9 (<0.0001)	5.5 (0.0001)	0.734	30.3 (0.0002)
			Δ_{HL} [–]	–0.826	**123(87)**	–47(10)	**1.4 (0.18)**	4.9 (0.0004)	0.682	23.6 (0.0005)
			σ_b [+]	0.841	**–229(150)**	10891(2111)	**1.5 (0.15)**	5.2 (0.0002)	0.708	26.6 (0.0003)
			σ_r [–]	–0.844	603(12)	–7592(1452)	49.7 (<0.0001)	5.2 (0.0002)	0.713	27.3 (0.0003)
			D_{CC} [+]	0.868	–4432(858)	3358(579)	5.1 (0.0002)	5.8 (<0.0001)	0.754	33.6 (0.0001)
			Q_{C2mul} [+]	0.889	641(16)	408(63)	41.2 (<0.0001)	6.4 (<0.0001)	0.789	41.2 (<0.0001)
			Q_{Omul} [+]	0.918	910(48)	1124(147)	19.1 (<0.0001)	7.7 (<0.0001)	0.843	58.9 (<0.0001)

[a]Number of samples in a given data set or its subset.

[b]Statistically not significant relationships are typed bold. Molecular descriptors which are characterized by most or all of such relationships are also typed bold. This means that the data set containing such descriptors is not statistically justified. Signs "+" and "–" in square brackets for a particular descriptor denote its positive or negative regression coefficient in the regression model.

[c]Pearson correlation coefficient between a descriptor and the dependent variable **y**.

[d]Regression coefficients a and b from a linear regression equation for **y** and descriptor **x**: $\mathbf{y} = a + b\,\mathbf{x}$. Statistical errors on the coefficients a and b are σ_a and σ_b, respectively. The values of the coefficients are rounded to significant figures and the respective errors are given in brackets.

[e]Student t-test parameters for the regression coefficients a and b are $t_{a,n-1} = a/\sigma_a$ and $t_{b,n-1} = b/\sigma_b$, respectively. The parameters are for n number of samples, *i.e.*, n-1 degrees of freedom. Corresponding probabilities p_a and p_b, *i.e.*, probabilities that the regression coefficients are not statistically significant (statistically not different from zero), are given in brackets, rounded to one or two significant figures.

[f]Explained fitted variance, defined in Table 1 as the coefficient of multiple determination. In the special case of linear regression, it may be considered as the coefficient of (univariate) determination.

[g]F-value from F-test and the corresponding probability in brackets, *i.e.*, the probability that the obtained linear regression equation is not statistically significant (obtained by chance). F-values are given for one variable and n-2 degrees of freedom. F-value is rounded to one digit and its probability is given for one or two significant figures.

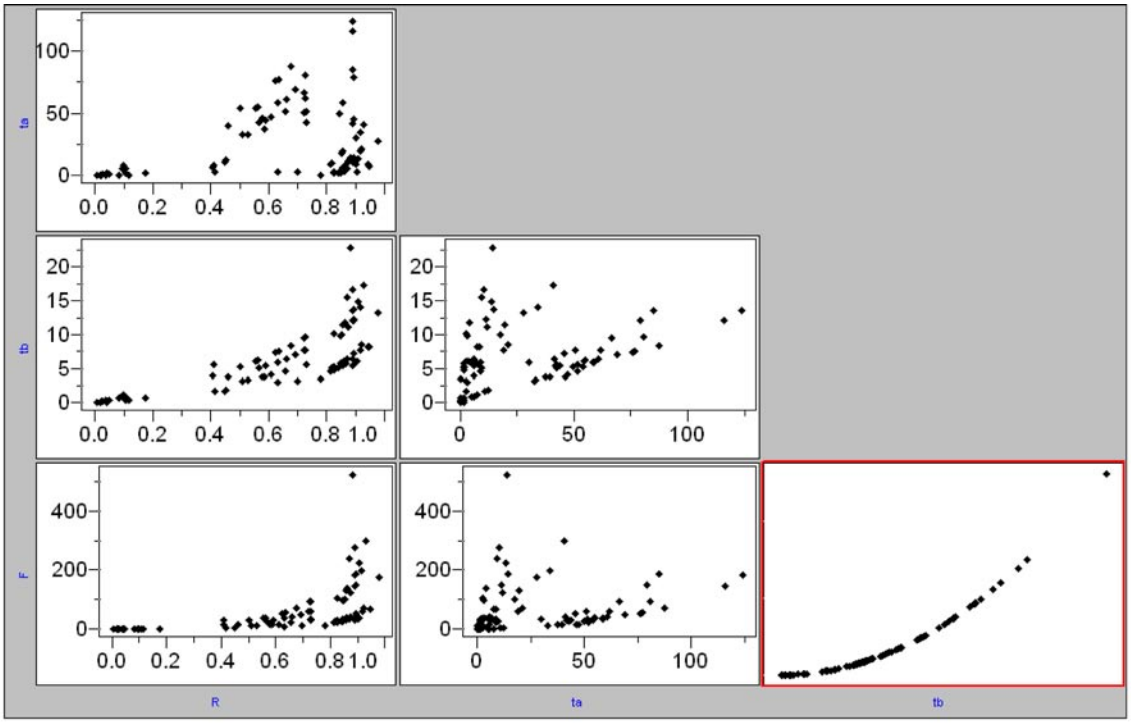

Figure F32. The scatterplots illustrating some of the relationships between statistical parameters from Table T33. The analytical relationship F - t_b is marked in red.