

Supporting Information

Similarity Searching for Potent Compounds Using Feature Selection

Martin Vogt and Jürgen Bajorath

¹Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany.

Supplementary Table S1

Supplementary Table S2

Table S1. ChEMBL compound data sets.

ChEMBL Target id	Target name	Target accession no. (SWISS-PROT)	Size
11	Thrombin	P00734	787
15	Carbonic anhydrase II	P00918	1498
43	Beta-2 adrenergic receptor	P07550	291
47	Muscarinic acetylcholine receptor M2	P08172	345
61	Muscarinic acetylcholine receptor M1	P11229	300
72	Dopamine D2 receptor	P14416	1765
86	Monoamine oxidase A	P21397	171
93	Acetylcholinesterase	P22303	111
100	Norepinephrine transporter	P23975	1054
104	Monoamine oxidase B	P27338	135
114	Adenosine A1 receptor	P30542	2024
121	Serotonin transporter	P31645	1367
129	Mu opioid receptor	P35372	1501
130	Dopamine D3 receptor	P35462	1301
136	Delta opioid receptor	P41143	1249
155	Dopamine transporter	Q01959	863
165	HERG	Q12809	846
194	Coagulation factor X	P00742	1191
214	Muscarinic acetylcholine receptor M4	P08173	211
215	Muscarinic acetylcholine receptor M5	P08912	194
219	Muscarinic acetylcholine receptor M3	P20309	491
226	Beta-3 adrenergic receptor	P13945	181
252	Adenosine A2a receptor	P29274	2153
259	Cannabinoid CB2 receptor	P34972	1565
278	Adenosine A2b receptor	P29275	903
280	Adenosine A3 receptor	P33765	1827
10142	Melanocortin receptor 4	P32245	1287
10186	Carbonic anhydrase VI	P23280	202
10193	Carbonic anhydrase I	P00915	1439
10378	Cathepsin B	P07858	196
10495	Cathepsin K	P43235	271
10498	Cathepsin L	P07711	206
10532	Butyrylcholinesterase	P06276	159
10911	Phenylethanolamine N-methyltransferase	P11086	144
11003	Melanocortin receptor 3	P41968	373
11006	Melanocortin receptor 5	P33032	273
11016	Carbonic anhydrase VB	Q9Y2D0	177

11042	Carbonic anhydrase VA	P35218	244
11060	Carbonic anhydrase VII	P43166	247
11063	Carbonic anhydrase XIV	Q9ULX7	331
11109	Matrix metalloproteinase 3	P08254	200
11280	Metabotropic glutamate receptor 5	P41594	117
11290	Histamine H4 receptor	Q9H3N8	643
11534	Cathepsin S	P25774	371
11627	Acyl coenzyme A:cholesterol acyltransferase	P23141	137
12209	Carbonic anhydrase XII	O43570	683
12896	Carbonic anhydrase IV	P22748	276
12952	Carbonic anhydrase IX	Q16790	1099
13061	Protein-tyrosine phosphatase 1B	P18031	168

All 49 activity classes are listed. Size refers to the number of compounds per class.

Table S2. Potency range-dependent compound subsets.

Target id	No. of compounds			
	Ki >= 7	7 > Ki >= 6	6 > Ki >= 5	5 > Ki
11	787	312	179	218
15	1498	667	351	214
43	291	154	73	38
47	345	165	97	55
61	300	141	84	47
72	1765	799	646	263
86	171	26	46	49
93	111	55	14	14
100	1054	417	308	251
104	135	32	33	42
114	2024	612	852	519
121	1367	804	303	199
129	1501	816	462	203
130	1301	753	349	168
136	1249	608	331	247
155	863	280	224	295
165	846	104	248	385
194	1191	841	135	163
214	211	86	63	32
215	194	62	47	53
219	491	314	96	49
226	181	64	48	50
252	2153	971	655	462
259	1565	868	386	274
278	903	514	252	116
280	1827	902	551	308
10142	1287	604	363	242
10186	202	26	39	70
10193	1439	334	237	548
10378	196	42	21	71
10495	271	102	77	67
10498	206	54	40	54
10532	159	36	39	47
10911	144	20	36	49
11003	373	34	86	88
11006	273	27	55	84
11016	177	64	19	61
11042	244	59	33	66

11060	247	133	20	40
11063	331	183	65	59
11109	200	97	47	30
11280	117	35	34	28
11290	643	336	202	72
11534	371	181	90	55
11627	137	33	45	39
12209	683	498	60	98
12896	276	28	48	89
12952	1099	618	191	187
13061	168	30	33	65

For each activity class, the distribution of compounds over subsets covering four different potency ranges is reported.