Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds

Hanna Eckert and Jürgen Bajorath

Supplementary Tables 1, 2, and 3

Table 1. Molecular descriptors selected for the generation of PDR-FP.

Descriptor	Bits	Definition
Physical properties		
apol	6	Sum of atomic polarizabilities[1]
bpol	6	Difference of bonded atom polarizabilities[1]
density	6	Mass density (weight divided by VDW volume) [2]
logP(o/w)	6	Log of the octanol/water partition coefficient[3]
logs	6	Log of the solubility in water[4]
mr	6	Molar refractivity[5]
SlogP	6	Log of the octanol/water partition coefficient[6]
TPSA	6	Polar VDW surface area (connection table approximation) [7]
Weight	6	Molecular weight
vdw_area	6	VDW surface area (connection table approximation) [2]
vdw_vol	6	VDW volume (connection table approximation) [2]
Atom and bond counts		
a_aro	3	Number of aromatic atoms
a_heavy	6	Number of heavy (non-hydrogen) atoms
a_IC	6	Atom information content (total)
a_ICM	6	Atom information content (mean)
a_nC	4	Number of carbon atoms
a_nN	4	Number of nitrogen atoms
a_nO	5	Number of oxygen atoms
a_nS	2	Number of sulfur atoms
b_1rotN	6	Number of rotatable single bonds
b_1rotR	6	Fraction of rotatable single bonds
b_count	6	Number of bonds
b_double	3	Number of double bonds
b_single	6	Number of single bonds
rings	3	Number of rings
opr_brigid	7	Number of rigid bonds from[8]
Kier&Hall Connectivity and Kappa Shape indices		
chi0	6	Atomic connectivity index (order 0) [9,10]
chi0_C	6	Carbon connectivity index (order 0) [9,10]
chi1	6	Atomic connectivity index (order 1) [9,10]
chi1_C	6	Carbon connectivity index (order 1) [9,10]
Kier1	6	First kappa shape index[10]
Kier2	6	Second kappa shape index[10]
Kier3	6	Third kappa shape index[10]
KierA1	6	First alpha modified shape index[10]
KierA2	6	Second alpha modified shape index[10]
KierA3	6	Third alpha modified shape index[10]
KierFlex	6	Molecular flexibility index[10]
zagreb	6	Zagreb index[2]

Adjacency and distance matrix descriptors		
balabanJ	6	Balaban averaged distance sum connectivity[11]
diameter	5	Largest value in the distance matrix[12]
GCUT_PEOE_0	6	PEOE partial charge GCUT (0/3) [2]
GCUT_PEOE_1	6	PEOE partial charge GCUT (1/3) [2]
GCUT_PEOE_2	6	PEOE partial charge GCUT (2/3) [2]
GCUT_PEOE_3	6	PEOE partial charge GCUT (3/3) [2]
GCUT_SLOGP_0	6	logP GCUT (0/3) [6]
GCUT_SLOGP_1	6	logP GCUT (1/3) [6]
GCUT_SLOGP_2	6	logP GCUT (2/3) [6]
GCUT_SLOGP_3	6	logP GCUT (3/3) [6]
GCUT_SMR_0	6	Molar refractivity GCUT (0/3) [6]
GCUT_SMR_1	5	Molar refractivity GCUT (0/3) [6]
GCUT_SMR_2	5	Molar refractivity GCUT (0/3) [6]
GCUT_SMR_3	6	Molar refractivity GCUT (0/3) [6]
VDistEq	6	Vertex distance equality index[2]
VDistMa	6	Vertex distance magnitude index[2]
weinerPath	6	Weiner path number[13]
weinerPol	6	Weiner polarity number[13]
Subdivided surface areas		
SlogP_VSA0	5	Sum of v_i such that L_i is $\leq$ -0.40 [6]
SlogP_VSA1	6	Sum of v_i such that L_i is in (-0.40,-0.20] [6]
SlogP_VSA2	4	Sum of v_i such that L_i is in (-0.20, 0.00] [6]
SlogP_VSA3	3	Sum of v_i such that L_i is in (0.00, 0.10] [6]
SlogP_VSA4	3	Sum of v_i such that L_i is in (0.10, 0.15] [6]
SlogP_VSA5	4	Sum of v_i such that L_i is in (0.15, 0.20] [6]
SlogP_VSA7	6	Sum of v_i such that L_i is in (0.25, 0.30] [6]
SMR_VSA0	3	Sum of v_i such that R_i is in (0.00, 0.11] [6]
SMR_VSA1	6	Sum of v_i such that R_i is in (0.11, 0.26] [6]
SMR_VSA2	3	Sum of v_i such that R_i is in (0.26, 0.35] [6]
SMR_VSA3	6	Sum of v_i such that R_i is in (0.35, 0.39] [6]
SMR_VSA4	5	Sum of v_i such that R_i is in (0.39, 0.44] [6]
SMR_VSA5	6	Sum of v_i such that R_i is in (0.44, 0.49] [6]
Pharmacophore feature descriptor		
a_acc	3	Number of H-bond acceptor atoms
a_don	3	Number of H-bond donoar atoms
a_hyd	5	Number of hydrophobic atoms
vsa_acc	5	VDW surface area of H-bond acceptors
vsa_don	3	VDW surface area of H-bond donors
vsa_hyd	6	VDW surface area of hydrophobic atoms
vsa_other	6	VDW surface area of other atoms
Partial charge descriptors		
PEOE_PC+	6	Total positive partial charge [14]
PEOE_PC-	6	Total negative partial charge [14]
PEOE_RPC+	6	Relative positive partial charge [14]

PEOE_RPC-	6	Relative negative partial charge [14]
PEOE_VSA+4	2	Sum of v_i where q_i is in [0.20, 0.25)[14]
PEOE_VSA-5	4	Sum of v_i where q_i is in [-0.30, -0.25) [14]
PEOE_VSA-6	3	Sum of v_i where q_i is < -0.30 [14]
PEOE_VSA_FPNEG	6	Fractional polar negative VDW surface area[14]
PEOE_VSA_FPOL	6	Fractional polar VDW surface area[14]
PEOE_VSA_FPOS	6	Fractional positive VDW surface area[14]
PEOE_VSA_FPPOS	6	Fractional polar positive VDW surface area[14]
PEOE_VSA_HYD	6	Total hydrophobic VDW surface area[14]
PEOE_VSA_NEG	6	Total negative VDW surface area[14]
PEOE_VSA_PNEG	6	Total polar negative VDW surface area[14]
PEOE_VSA_POL	6	Total polar VDW surface area[14]
PEOE_VSA_POS	6	Total positive VDW surface area[14]
PEOE_VSA_PPOS	6	Total polar positive VDW surface area[14]

Descriptors implemented in PDR-FP are designated according to their MOE implementation[2]. "Bits" reports the number of bits that are used for equifrequent subdivision of descriptor value ranges. In the descriptor definition column, v_i is the van der Waals surface area of atom i and q_i represents the partial charge of atom i. L_i is the contribution to logP(o/w) of atom i, and R_i is the contribution to molar refractivity of atom i.

Table 2. Descriptor intervals implemented in PDR-FP.

Descriptor	Interval boundaries				
a_acc	3	4			
a_aro	12	13			
a_don	1	2			
a_heavy	20	23	25	27	30
a_hyd	13	15	17	19	
a_IC	55.445	63.619	70.555	77.503	85.921
a_ICM	1.4968	1.5864	1.6579	1.7336	1.8301
a_nC	16	19	22		
a_nN	2	3	4		
a_nO	2	3	4	5	
a_nS	1				
apol	40.835	46.703	51.479	56.281	62.177
b_1rotN	3	4	5	6	7
b_1rotR	0.1154	0.1500	0.1780	0.2060	0.2500
b_count	35	40	45	49	55
b_double	2	3			
b_single	22	26	30	34	39
balabanJ	1.3464	1.4597	1.5689	1.6870	1.8577
bpol	20.223	24.126	27.470	31.006	35.383
chi0	14.113	16.113	17.648	19.225	21.087
chi0_C	9.422	10.991	12.242	13.405	14.941
chi1	9.414	10.742	11.863	12.957	14.303
chi1_C	5.0328	6.1018	6.8602	7.6330	8.6574
density	0.7273	0.7560	0.7816	0.8116	0.8576
diameter	11	13	14	16	
GCUT_PEOE_0	-0.8595	-0.8413	-0.8286	-0.8124	-0.7998
GCUT_PEOE_1	-0.4437	-0.4232	-0.4098	-0.3964	-0.3826
GCUT_PEOE_2	0.0110	0.0490	0.0636	0.0737	0.0828
GCUT_PEOE_3	2.3524	2.4408	2.5227	2.6140	2.7312
GCUT_SLOGP_0	-1.4897	-1.1630	-1.0127	-0.9974	-0.9507
GCUT_SLOGP_1	-0.3534	-0.3252	-0.3036	-0.2832	-0.2639
GCUT_SLOGP_2	0.1059	0.1386	0.1516	0.1645	0.1833
GCUT_SLOGP_3	2.4479	2.5321	2.6030	2.6782	2.7777
GCUT_SMR_0	-0.5312	-0.5219	-0.5116	-0.4912	-0.4823
GCUT_SMR_1	-0.2161	-0.2054	-0.1968	-0.1805	
GCUT_SMR_2	0.1647	0.1839	0.1958	0.2360	
GCUT_SMR_3	2.6500	2.7361	2.8126	2.8951	2.9985
Kier1	15.3	17.7	19.4	21.3	23.5
Kier2	6.857	8.022	8.982	9.871	10.951
Kier3	3.6213	4.3795	5.0000	5.5868	6.3525
KierA1	11.898	13.717	15.198	16.710	18.564

KierA2	5.2301	6.1492	6.8918	7.6389	8.5375	
KierA3	2.7156	3.3068	3.8000	4.3020	4.9492	
KierFlex	3.0237	3.6822	4.2399	4.8266	5.5842	
logP(o/w)	1.9320	2.7470	3.3510	3.8970	4.5000	
logS	-6.3502	-5.5532	-4.9185	-4.2682	-3.4056	
mr	7.695	8.764	9.609	10.464	11.519	
opr_brigid	12	13	17	18	20	23
PEOE_PC+	1.5082	1.8050	2.0474	2.2943	2.6203	
PEOE_PC-	-2.5295	-2.2428	-2.0152	-1.7850	-1.5032	
PEOE_RPC+	0.1094	0.1232	0.1363	0.1518	0.1755	
PEOE_RPC-	0.1435	0.1614	0.1792	0.2014	0.2366	
PEOE_VSA+4	0.1					
PEOE_VSA-5	13.50	13.72	27.18			
PEOE_VSA-6	2.22	5.10				
PEOE_VSA_FPNEG	0.0504	0.0725	0.0915	0.1140	0.1461	
PEOE_VSA_FPOL	0.0963	0.1306	0.1605	0.1920	0.2362	
PEOE_VSA_FPOS	0.4339	0.4790	0.5170	0.5576	0.6124	
PEOE_VSA_FPPOS	0.0337	0.0464	0.0643	0.0802	0.1053	
PEOE_VSA_HYD	225.85	259.67	286.43	312.97	345.83	
PEOE_VSA_NEG	122.81	146.06	163.76	180.86	201.37	
PEOE_VSA_PNEG	16.208	24.394	32.017	39.316	49.993	
PEOE_VSA_POL	31.524	42.587	56.453	65.690	82.908	
PEOE_VSA_POS	129.90	155.06	176.77	199.77	229.71	
PEOE_VSA_PPOS	12.900	12.989	22.000	27.655	36.220	
rings	3	4				
SlogP	1.7161	2.6368	3.2807	3.8560	4.4925	
SlogP_VSA0	0.01	14.10	22.00	40.00		
SlogP_VSA1	8.400	16.786	26.005	37.020	53.859	
SlogP_VSA2	11.000	23.868	47.545			
SlogP_VSA3	1.0	39.2				
SlogP_VSA4	5.92	27.08				
SlogP_VSA5	0.01	16.80	37.75			
SlogP_VSA7	70.57	92.33	123.51	143.47	167.06	
SMR_VSA0	23.87	58.72				
SMR_VSA1	3.10	16.60	22.00	34.97	50.75	
SMR_VSA2	0.20	17.99				
SMR_VSA3	4.736	7.911	10.915	14.739	20.840	
SMR_VSA4	2.740	5.514	17.659	25.912		
SMR_VSA5	88.20	110.76	141.14	158.79	183.78	
TPSA	47.560	60.440	71.150	82.430	97.300	
VDistEq	3.1314	3.3298	3.4693	3.5938	3.7335	
VDistMa	8.2985	8.6583	8.9500	9.1954	9.4879	
vdw_area	277.09	313.78	343.39	373.14	409.29	
vdw_vol	361.18	411.63	451.90	492.42	542.43	
vsa_acc	19.249	28.433	35.320	46.413		

vsa_don	1	6			
vsa_hyd	186.84	217.04	240.32	263.57	292.26
vsa_other	25.899	34.807	43.033	51.046	62.009
Weight	284.37	325.36	357.20	390.42	429.88
weinerPath	813	1191	1551	1976	2578
weinerPol	27	33	38	43	49
zagreb	99	115	127	139	155

For each descriptor, rounded interval boundaries are reported that were used to divide descriptor value ranges into equifrequent intervals. The individual number of intervals (between two and seven) chosen for binning was dependent one the database value distribution. For n interval boundaries b_1, b_2, …, b_n, the following $n + 1$ intervals were defined: $[-\infty, b_1)$, $[b_1, b_2)$, …, $[b_{n-1}, b_n)$, $[b_n, \infty]$.

Table 3. Tc similarity statistic based on PDR-FP.

Code	Biological activity	NC	min Tc	max Tc	av Tc	stdDev Tc
ACE	antihypertensive (ACE inhibitors)	15	0.177	0.722	0.367	0.122
COX	cyclooxygenase-2 (Cox-2) inhibitors	16	0.107	0.824	0.307	0.128
HIV	HIV protease inhibitors	15	0.292	0.824	0.532	0.106
BK2	bradykinin BK2 antagonist	22	0.039	0.706	0.421	0.137
ETA	endothelin ETA antagonist	28	0.051	0.661	0.342	0.124
SQS	squalene synthetase inhibitors	29	0.063	0.755	0.367	0.154
GLU	glucagon receptor antagonists	33	0.045	0.691	0.329	0.153
ULD	upregulator of LDL Receptor	21	0.028	0.676	0.331	0.168
SQE	squalene epoxidase inhibitors	25	0.075	0.806	0.294	0.150
CAM	cell adhesion molecule antagonists	25	0.094	0.661	0.351	0.131
F7I	factor VIIa inhibitors	23	0.114	0.789	0.384	0.122
GLY	glycoprotein IIb/IIIa receptor antagonists	25	0.094	0.722	0.319	0.135
NK2	neurokinin NK2 antagonists	25	0.274	0.755	0.473	0.078
REN	renin inhibitors	22	0.368	0.806	0.556	0.086
TBI	thrombin inhibitors	48	0.102	0.789	0.419	0.119

Reported are Tc similarity statistics produced in complete pairwise comparisons with PDR-FP for all 15 compound activity classes: “min Tc”, “max Tc”, and “av Tc” stand for the minimum, maximum, and average Tc value, respectively, and “stdDev” reports the standard deviation of the Tc values. In addition, “Code” contains the class code and “NC” the number of compounds per class.

References

1. *CRC Handbook of Chemistry and Physics*; CRC Press, 1994.
2. *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3, 2005.
3. Labute, P. MOE LogP(Octanol/Water) Model. *unpublished*, **1998**. Source code in `$MOE/lib/svl/quasar.svl/q_logp.svl`.
4. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach; **2004**, 266-275.
5. Labute, P. MOE Molar Refractivity Model. *unpublished*, **1998**. Source code in `$MOE/lib/svl/quasar.svl/q_mref.svl`.
6. Wildman, S. A.; Crippen, G. M. Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.
7. Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714-3717.
8. Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comp. Aid. Mol. Des.* **2000**, *14*, 251-264.
9. Hall, L. H.; Kier, L. B. The nature of structure-activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307.
10. Hall, L.H.; Kier, L.B. The molecular connectivity chi indices and kappa shape indices in structure-property modeling. *Rev. Comput. Chem.*, K. B. Lipkowitz and D. B. Boyd, Eds. New York: VCH, **1991**, 367-422.

11. Balaban, A.T. Highly discriminating distance-based topological index. *Chemical Physics Letters* **1982**, *89*, No. 5, 399-404.

12. Petitjean, M. Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331-337.

13. Balaban, A.T. Five new topological indices for the branching of tree-like graphs. *Theoretica Chimica Acta* **1979**, *53*, 355-375.

14. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219.