**Emerging chemical patterns: a new methodology for molecular classification and
compound selection**

Jens Auer and Jürgen Bajorath

Supplementary Table 1

Supplementary Figures 1 and 2

Table 1. Molecular descriptors.

Descriptor	Definition
<i>Physical properties</i>	
logP(o/w)	Log of the octanol/water partition coefficient ¹
TPSA	Polar VDW surface area (connection table approximation) ²
<i>Atom Counts and Bond Counts</i>	
a_ICM	Atom information content (mean)
a_nBR	Number of bromine atoms
a_nCl	Number of chlorine atoms
a_nF	Number of fluorine atoms
a_nI	Number of iodine atoms
a_nN	Number of nitrogen atoms
a_nO	Number of oxygen atoms
a_nP	Number of phosphor atoms
a_nS	Number of sulfur atoms
b_1rotR	Fraction of rotatable bonds
b_1rotN	Number of rotatable bonds
b_double	Number of double bonds
b_triple	Number of triple bonds
<i>Kier&Hall Connectivity and Kappa Shape Indices</i>	
chi0v_C	Carbon valence connectivity index (order 0). ^{3,4}
<i>Adjacency and distance matrix descriptors</i>	
balabanJ	Balaban averaged distance sum connectivity ⁵
petitjean	Value of (diameter - radius) / diameter. ⁶
VDistEQ	Vertex distance equality index ⁷
<i>Subdivided surface areas</i>	
SlogP_VSA0	Sum of v_i such that $L_i \leq -0.40$ ⁸
SlogP_VSA1	Sum of v_i such that L_i is in $(-0.40, -0.20]$ ⁸
SlogP_VSA2	Sum of v_i such that L_i is in $(-0.20, 0.00]$ ⁸
SlogP_VSA3	Sum of v_i such that L_i is in $(0.00, 0.10]$ ⁸
SlogP_VSA4	Sum of v_i such that L_i is in $(0.10, 0.15]$ ⁸
SlogP_VSA5	Sum of v_i such that L_i is in $(0.15, 0.20]$ ⁸
SlogP_VSA7	Sum of v_i such that L_i is in $(0.25, 0.30]$ ⁸
SlogP_VSA8	Sum of v_i such that L_i is in $(0.30, 0.40]$ ⁸
SlogP_VSA9	Sum of v_i such that $L_i > 0.40$ ⁸
SMR_VSA0	Sum of v_i such that R_i is in $(0.00, 0.11]$ ⁸
SMR_VSA1	Sum of v_i such that R_i is in $(0.11, 0.26]$ ⁸
SMR_VSA2	Sum of v_i such that R_i is in $(0.26, 0.35]$ ⁸
SMR_VSA3	Sum of v_i such that R_i is in $(0.35, 0.39]$ ⁸

SMR_VSA4	Sum of v_i such that R_i is in $(0.39, 0.44]^8$
SMR_VSA5	Sum of v_i such that R_i is in $(0.44, 0.49]^8$
SMR_VSA6	Sum of v_i such that R_i is in $(0.485, 0.56]^8$
<i>Pharmacophore feature descriptor</i>	
a_acc	Number of hydrogen bond acceptor atoms.
a_don	Number of hydrogen bond donor atoms.
a_base	Number of basic atoms.
vsa_acid	VDW surface areas of acidic atoms.
vsa_base	VDW surface areas of basic atoms.
vsa_don	VDW surface areas of H-bond donors
vsa_other	VDW surface area of other atoms.
vsa_pol	VDW surface areas of polar atoms.
<i>Partial charge descriptors</i>	
PEOE_RPC-	Relative negative partial charge ⁹
PEOE_VSA_FNEG	Fractional negative VDW surface area. ⁹
PEOE_VSA_FHYD	Fractional hydrophobic VDW surface area. ⁹
PEOE_VSA+0	Sum of v_i where q_i is in $[0.00, 0.05)^9$
PEOE_VSA+1	Sum of v_i where q_i is in $[0.05, 0.10)^9$
PEOE_VSA+2	Sum of v_i where q_i is in $[0.10, 0.15)^9$
PEOE_VSA+3	Sum of v_i where q_i is in $[0.15, 0.20)^9$
PEOE_VSA+4	Sum of v_i where q_i is in $[0.20, 0.25)^9$
PEOE_VSA+5	Sum of v_i where q_i is in $[0.20, 0.30)^9$
PEOE_VSA+6	Sum of v_i where $q_i \geq 0.30$ ⁹
PEOE_VSA-0	Sum of v_i where q_i is in $[-0.05, 0.00)^9$
PEOE_VSA-1	Sum of v_i where q_i is in $[-0.10, -0.05)^9$
PEOE_VSA-2	Sum of v_i where q_i is in $[-0.15, -0.10)^9$
PEOE_VSA-3	Sum of v_i where q_i is in $[-0.20, -0.15)^9$
PEOE_VSA-4	Sum of v_i where q_i is in $[-0.25, -0.20)^9$
PEOE_VSA-5	Sum of v_i where q_i is in $[-0.30, -0.25)^9$
PEOE_VSA-6	Sum of v_i where $q_i < -0.30$ ⁹

Figure legends

Figure 1. Simulated lead optimization trials with binary QSAR. Average potencies and compound numbers over 500 calculations of ten iterations each are reported for training sets of 5 and 10 compounds: (a) BZR, (b) DHFR, (c) GSK3, (d) HIVPROT. Box plot representations are according to **Figure 1** of the manuscript.

Figure 2. Simulated lead optimization trials with decision tree. Average potencies and compound numbers over 500 calculations of ten iterations each are reported for training sets of 5 and 10 compounds: (a) BZR, (b) DHFR, (c) GSK3, (d) HIVPROT. Box plot representations are according to **Figure 1** of the manuscript.

Figure 1a

Figure 1b

Figure 1c

Figure 1d

Figure 2a.

Figure 2b.

Figure 2c.

Figure 2d.

References

1. Labute, P. MOE LogP(Octanol/Water) Model. *unpublished*, **1998**. Source code in \$MOE/lib/svl/quasar.svl/q_logp.svl.
2. 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.
3. 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.
4. 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.
5. Balaban, A.T. Highly discriminating distance-based topological index. *Chemical Physics Letters* **1982**, *89*, No. 5, 399-404.
6. 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.
7. MOE (*Molecular Operating Environment*); Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3, 2005.
8. Wildman, S. A.; Crippen, G. M. Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.
9. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219.