Computer Chemistry

Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons**

*Tobias Fink, Heinz Bruggesser, and Jean-Louis Reymond**

The development of modern medicine largely depends on the continuous discovery of new drug molecules for treating diseases.[1] One striking feature of these drugs is their relatively small molecular weight (M_W), which averages only 340 Da.[2] Recently, drug discovery has focused on even smaller building blocks with M_W of 160 Da or less to be used as lead structures that can be optimized for biological activity by adding substituents[3] At that size it becomes legitimate to ask how many such very small molecules would be possible in total within the boundaries of synthetic organic chemistry? To address this question we have generated a database (GDB) containing all possible organic structures with up to 11 main atoms under constraints defining chemical stability and synthetic feasibility. The database contains 13.9 million molecules with an average M_W of 153 Da, and opens an unprecedented window on the small-molecule chemical universe.

Estimates have been proposed for the total number of organic molecules to be in the range of 10^{18}–10^{200} compounds.[4] Systematic analysis of a database of screening compounds at Novartis identified 849 574 different substituents with 12 main atoms or less, which illustrates what synthetic chemistry has achieved so far.[5,6] To gain an insight into the size and composition of the entire small-molecule chemical universe, we set out to generate all possible organic molecules up to $M_W \approx 160$ by computer simulation from first principles.

Several programs exist to enumerate molecular structures corresponding to a given elemental formula,[7] but they have never been implemented to carry out an exhaustive listing and their adaptation to this task would be quite cumbersome. The database was therefore constructed by a new approach, starting with a collection of mathematical graphs corresponding to saturated hydrocarbons, which were diversified to

[*] T. Fink, Prof. Dr. J.-L. Reymond
Department of Chemistry and Biochemistry
University of Berne
Freiestrasse 3, 3012 Berne (Switzerland)
Fax: (+41) 31-631-8057
E-mail: jean-louis.reymond@ioc.unibe.ch

Dr. H. Bruggesser
Institute of Mathematics
University of Berne
Sidlerstrasse 5, 3012 Berne (Switzerland)

[**] This work was financially supported by the University of Berne and the Swiss National Science Foundation. The authors thank Dr. Peter Ertl and Dr. Bernhard Rohde at Novartis for helpful discussions, Molinspiration for providing access to their virtual screening toolkit, and Molecular Networks GmbH for use of the 3D-coordinates generation program CORINA.

DOI: 10.1002/ange.200462457

molecules by systematically introducing bond unsaturations and atom types using an in-house developed application written in Java.

An exhaustive library of graphs with up to 11 nodes and a maximum node connectivity of four was produced by the program NAUTY.[8] The vast majority of these graphs (99.8%) contained three- and four-membered rings and was excluded to avoid generating a database consisting almost exclusively of such small rings.[9] Further restrictions included the elimination of nonplanar graphs and tricyclic bridgeheads.[10] Graph symmetry was determined in each graph using a known algorithm.[11] Bond unsaturations were then introduced combinatorially, followed by all possible atom-type combinations by introducing carbon, nitrogen, oxygen, and fluorine (as a model halogen) at each node. The resulting collection was finally reduced by applying filters for functional groups to eliminate unstable atom-type and bond-type combinations.[12] Each compound was stored as its USMILES representation of the structural formula.[13] The three-dimensional structure of each compound was finally determined using CORINA, thereby generating all possible stereoisomers.[14] A quantitative overview of the GDB construction process is shown in Table 1 and Table 2.

For comparison purposes a reference database (Rdb) of known compounds with up to 11 main atoms was assembled from ChemACX (21698 compounds (cpds)), ChemACX-SC (10735 cpds), NCI open database (19438 cpds), and the Merck Index (1540 cpds), resulting in 36227 unique structures.[16,17] 52% of these reference compounds were present in GDB. The remaining compounds contained features which had been specifically excluded, such as elements other than C, H, N, O, or halogen (23.5%), unstable functional groups (e.g. acyl halides, peroxides, 12.0%), 3- or 4-membered rings (5.4%), triple bonds, allenes, and bridgehead olefins (4.0%), or charges (e.g. quaternary ammonium centers, 3.0%). The composition of Rdb by molecular size and graph type is shown in Table 1.

Table 2: Stereochemical composition of GDB. The 13.9 million structures in GDB give rise to approximately 44 million stereoisomers as generated by CORINA.[14]

Contribution to GDB [%]	Stereochemical category[a]
24	No stereoisomers
18	*E*/*Z* isomers[b]
22	Two stereoisomers[c]
21	Multiple stereoisomers[d]
15	Mixed[e]

[a] Stereoisomeric diversity is mainly achieved by molecules of 11 atoms which make up 86% of GDB. [b] Single or multiple *E*/*Z* isomeric pairs and no stereogenic nonplanar centers (e.g. 2,4-hexadiene). [c] Enantiomeric pairs (e.g. 2-butanol) and achiral *syn*/*anti* pairs (e.g. 1,4-dimethyl cyclohexane) not having *E*/*Z* isomerism. [d] Structures with multiple independent stereogenic nonplanar centers, including *meso*-isomers, but no *E*/*Z* isomers (e.g. 2,3-butanediol). [e] Structures containing both nonplanar stereogenic centers and *E*/*Z* isomers (e.g. 3-penten-2-ol).

The 1830 graphs used for GDB generated between 4 and 79236 different compounds per graph. There were 103 different ring types in these graphs,[18] 50 of which did not appear in Rdb, although at least one Chemical Abstract System (CAS) entry could be found in each case for the parent hydrocarbon. In fact Rdb used only 1174 of the GDB graphs, but contained an additional 871 graphs with small rings and pentavalent nodes not used for GDB (Table 1, Figure 1). Analysis by compound type showed that heterocycles were most abundant in GDB, while aromatics were almost insignificant. By contrast, Rdb contained a relatively large proportion of acyclics and aromatics. Furthermore, GDB contained a much higher proportion of fused heterocycles than Rdb, but a smaller proportion of heteroaromatics. GDB-compounds had an average M_W of 153.2, and 87% of them had $M_W < 160$ (Table 3, Figure 2).

The databases were analyzed in terms of physicochemical and topological descriptors relevant for drug properties, including M_W, octanol/water partition coefficient (logP),[19]

Table 1: Overview of GDB and Rdb databases.

Parameter	Atoms											Total
	1	2	3	4	5	6	7	8	9	10	11	
GDB												
Graphs total[a]	1	1	2	6	21	78	353	1929	12207	89402	739335	**843335**
Maximal one 3- or 4-membered ring	1	1	2	4	9	22	64	215	769	3098	13808	**17993**
No 3- or 4-membered rings	1	1	1	2	4	7	16	41	119	394	1497	**2083**
Planar and no tricyclic bridgeheads	1	1	1	2	4	7	16	41	116	369	1272	**1830**
Molecules passed filters[b]	4	7	16	62	251	1252	6812	40942	258852	1719366	11864872	**13892436**[c]
Rdb												
Molecules[d]	7	43	127	277	612	1378	2492	4304	6257	8933	11797	**36227**
Graphs from molecules[e]	1	1	2	5	11	27	66	147	270	528	988	**2046**
Maximal one 3- or 4-membered ring	1	1	1	4	9	22	52	125	255	500	967	**1937**
No 3- or 4-membered rings	1	1	1	2	4	8	18	44	111	302	683	**1175**
Planar and no tricyclic bridgeheads	1	1	1	2	4	8[f]	18[f]	44[f]	111	302	682[g]	**1174**

[a] As generated by the program NAUTY.[8] [b] After application of filters.[12] 0.2% of molecules passed the filters. [c] The logarithm of the number of molecules in GDB increases as a quadratic function of the number of atoms, giving 145 million compounds for 12 atoms and 3×10^{25} for 25 atoms ($R^2 = 0.999$). [d] There are more molecules with four main atoms or less in Rdb because more elements types are used. [e] Number of different graphs used in the molecules. [f] Rdb contains pentavalent phosphorus derivatives (e.g. $MePCl_4$) which correspond to graphs with one node of connectivity 5 not used in GDB. [g] Tricyclo[3.3.3.0]undecane is the only graph for a stable tricyclic bridgehead molecule with up to 11 main atoms but was excluded from GDB.[15] For clarity, totals are highlighted in bold.

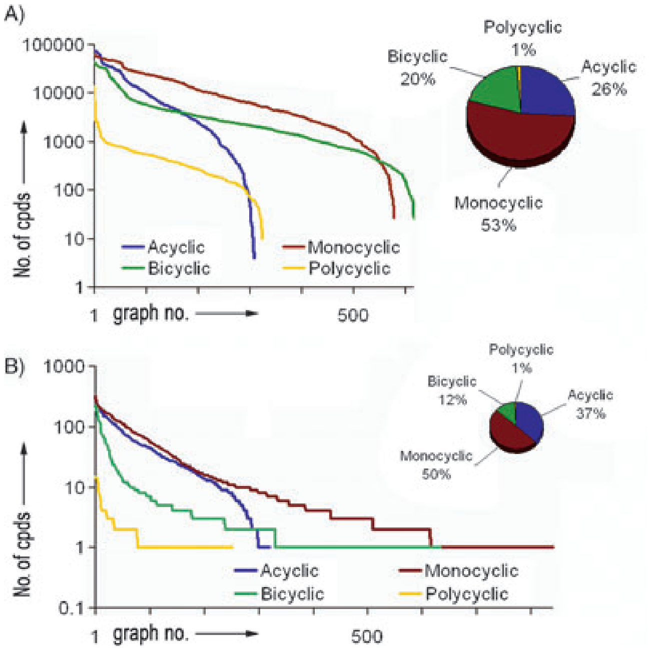


Figure 1. Composition of A) GDB and B) Rdb databases by graph type. Graphs were ordered in descending number of compounds per graph. The graphs giving the most compounds in GDB correspond to 4-ethylnonane (79236 cpds, acyclic), 1-ethyl-3-propylcyclohexane (60337 cpds, monocyclic), 5-ethyloctahydroindene (42682 cpds, bicyclic), and decahydrocyclopenta[α]pentalene (13882 cpds, polycyclic). For Rdb, the graphs with most compounds are 2-methylpentane (249 cpds, acyclic), 1,2,4-trimethylcyclohexane (325 cpds, monocyclic), 4,6-dimethyloctahydroindene (231 cpds, bicyclic), and 1-methyladamantane (15 cpds, polycyclic). There are more graph types for Rdb since 3- and 4-membered rings are also present (5.4% of Rdb, see text).

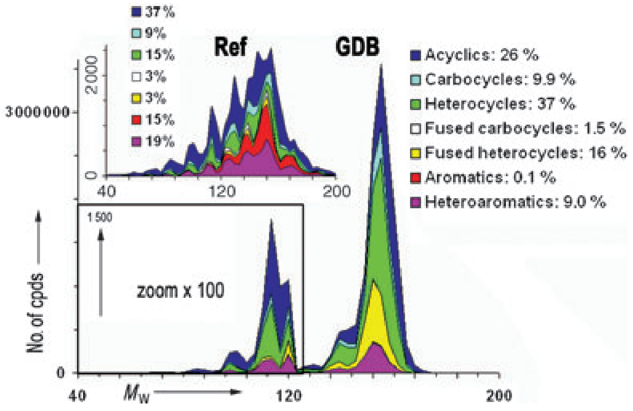


Figure 2. Database composition by structural categories and M_W. Main plot: 13.9 million GDB compounds containing up to 11 main atoms, average $M_W = 153.2$, $\sigma = 7.5$ Da, 87% of GDB-compounds have $M_W < 160$. The lower-left portion of GDB for $M_W < 120$ has been expanded 100× for visibility. Inset: 36227 Rdb compounds containing up to 11 main atoms. The M_W distribution in Rdb is broader owing to the heavier elements (P, S, Si). The maximum number of compounds occurs in the interval 155.2 ± 1.6 Da for both GDB and Rdb. For details of element composition per structural category see also Table 3.

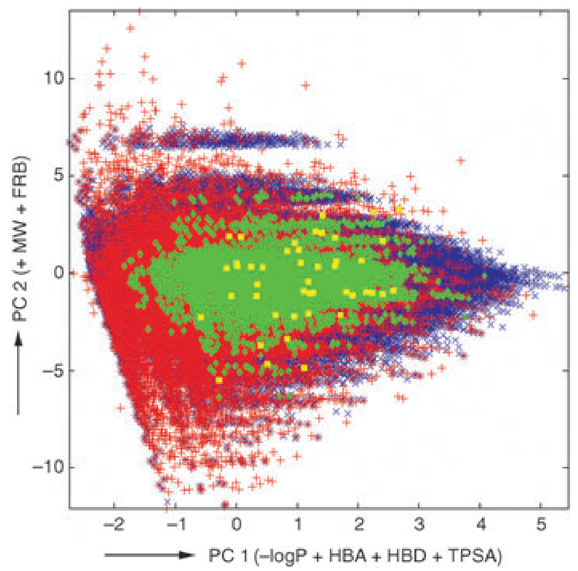


Figure 3. Coverage of property space by database compounds. × GDB, 13.9 millions cpds; + Rdb, 36227 cpds, overlayed on GDB; ◆ virtual hits from GDB, 25676 cpds, overlayed on previous series; ■ virtual hits from Rdb, 160 cpds, only the 47 hits not present in GDB are shown, overlayed on previous series. Principal component analysis (PCA) of GDB + Rdb gave: PC1: M_W 0.031, FRB 0.030, logP −0.930, HBA 0.915, HBD 0.827, TPSA 0.943; PC2: M_W 0.756, FRB 0.775, logP 0.080, HBA 0.039, HBD 0.085, TPSA 0.098. Virtual screening was performed using the Molinspiration software.[24] The hit rates were 0.2% for GDB and 0.4% for Rdb.

number of hydrogen-bond donors (HBD) and acceptors (HBA), fraction of rotatable bonds (FRB), and topological polar surface area (TPSA).[20] The chemical space defined by these properties was represented in a 2-dimensional projection along the first two principal components (covering 75% of the diversity), with PC1 reflecting the hydrophobic/hydrophilic balance and PC2 depending on molecular weight and conformational flexibility (Figure 3). GDB covered this chemical space much more densely and exhaustively than Rdb, with coverage extending in particular into high-polarity regions where no Rdb compounds were found.

The relevance of GDB for drug discovery was tested by virtual screening for bioactivity. Virtual screening uses quantitative structure–activity relationship (QSAR) methods, such as similarity searching,[21] statistical methods (principal component regression, partial least squares),[22] or neural networks.[23] We used a commercial package based on Bayesian statistics (Molinspiration miscreen toolkit)[24] trained for three important drug targets: G-protein coupled receptors (GPCR), kinases, and ion channels. The virtual screening returned a large number of high-scoring compounds in each case (GPCR ligands: 17106; Ion-channel modulators: 7527; Kinase inhibitors: 2071). While 90% of these virtual hits fell into regions of chemical space well covered by both databases, 10% of these hits were found in regions of space covered only by GDB but not by Rdb (Figure 3 and Figure 4).

The small-molecule chemical universe appears as a large yet tractable entity. Large portions of the chemical universe remain hidden as invisible "dark matter" in GDB, such as the overwhelming multitude of 3- and 4-membered ring compounds. Nevertheless, the current study reveals a wealth of

Table 3: Composition of GDB and Rdb databases by structural categories (columns) and element composition (rows).[a]

Elements				Structure category				
	Heteroaromatics	Aromatics	Fused Heterocycles	Fused Carbocycles	Heterocycles	Carbocycles	Acyclics	**Total**
GDB								
C	0	396	0	19 683	0	30 183	9364	**59 626**
C,F	0	1049	0	34 648	0	117 334	72 111	**225 142**
C,F,N	124 285	2437	174 875	19 638	453 588	170 793	377 756	**1 323 372**
C,F,N,O	232 910	2253	213 437	8120	866 091	177 613	746 025	**2 246 449**
C,F,O	22 037	2151	130 109	30 170	305 971	228 877	333 692	**1 053 007**
C,N	204 782	2761	435 108	29 359	723 949	143 720	352 117	**1 891 796**
C,N,O	648 694	5025	1 041 291	28 544	2 512 535	324 149	1 495 785	**6 056 023**
C,O	23 309	2273	246 184	43 477	335 434	179 018	207 323	**1 037 018**
No C	0	0	0	0	0	0	3	**3**
Total	**1 256 017**	**18 345**	**2 241 004**	**213 639**	**5 197 568**	**1 371 687**	**3 594 176**	**13 892 436**
Rdb								
C	0	269	0	352	0	457	594	**1672**
C,F	0	231	0	66	0	126	388	**811**
C,F,N	630	291	23	6	74	37	281	**1342**
C,F,N,O	510	275	10	9	241	36	508	**1589**
C,F,O	33	492	34	39	181	149	971	**1899**
C,N	1404	574	231	68	611	251	979	**4118**
C,N,O	2509	868	305	50	2046	427	3190	**9395**
C,O	159	779	301	391	850	978	2704	**6162**
No C	1	0	1	2	4	1	226	**235**
Others[b]	1665	1512	168	117	1229	782	3531	**9004**
Total	**6911**	**5291**	**1073**	**1100**	**5236**	**3244**	**13 372**	**36 227**

[a] Compounds are assigned to one category only with the following priorities: heteroaromatics > aromatics > fused heterocycles (including spiro compounds) > fused carbocycles (including spiro compounds) > heterocycles > carbocycles > acyclics. For example furyl-benzene is classified as heteroaromatic only. [b] "Others" are C containing compounds also containing elements other than C, N, O, or halogen (e.g. S, Si, or P). For details and M_W distribution by categories, see also Figure 2. For clarity, totals are highlighted in bold.

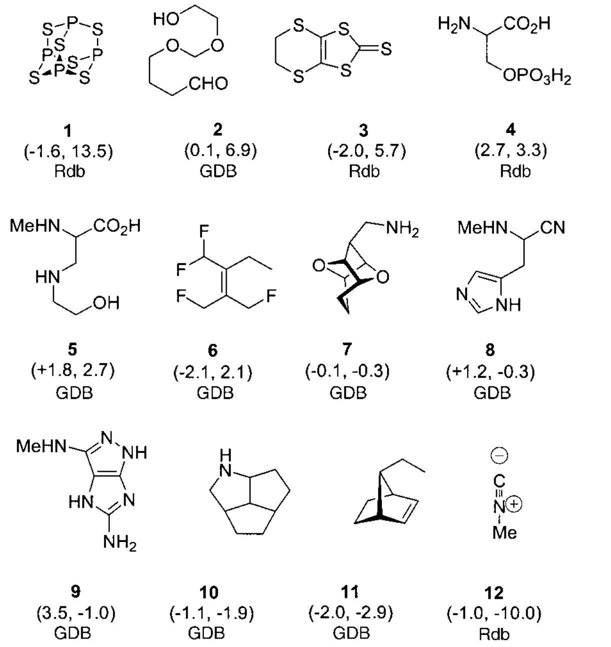


Figure 4. Examples of GDB and Rdb compounds in property space. PC-coordinates are given in parenthesis as (PC1, PC2) and the compounds are ordered by decreasing PC2. See also Figure 2 and 3. The Rdb compounds (**1**, **3**, **4**, **12**) are from areas of chemical space covered only by Rdb. Compounds **5**–**9** are not registered in the Chemical Abstracts System (CAS) and are therefore considered as unknown. Compounds **2** and **10** are also unknown although derivates are registered in CAS. Virtual screening[24] gives high scores for compounds **4** (GPCR ligand), **5** (ion-channel modulator), **8** (GPCR ligand), and **9** (kinase inhibitor).

organic structures below 160 Da covering property space broadly and extensively, with many possibly bioactive compounds. The database construction strategy chosen also ensures that the majority of GDB, although presently unknown, should be synthetically accessible.

Received: October 28, 2004
Published online: January 26, 2005

Keywords: chemoinformatics · combinatorial chemistry · computer chemistry · drug design · molecular diversity

[1] K. H. Bleicher, H.-J. Böhm, K. Müller, A. I. Alanine, *Nat. Rev. Drug Discovery* **2003**, *2*, 369–378.

[2] M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

[3] D. A. Erlanson, R. S. McDowell, T. O'Brien, *J. Med. Chem.* **2004**, *47*, 3463–3482.

[4] a) S. Petit-Zeman, Charting chemical space: finding new tools to explore biology. *4th Horizon Symposium*, Palazzo Arzaga, Italy, October 23–25, **2003**; b) R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.

[5] P. Ertl, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

[6] Note that there are many more substituents than molecules because any one molecule gives rise to several substituents because the attachement point behaves as a virtual atom. For example toluene (methylbenzene) gives four possible substituents: *alpha*-tolyl, *ortho*-tolyl, *meta*-tolyl, and *para*-tolyl.

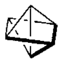
[7] a) R. E. Carhart, D. H. Smith, H. Brown, C. Djerassi, *J. Am. Chem. Soc.* **1975**, *97*, 5755–5762; b) R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Application of Artificial Intelligence for Chemistry: The DENDRAL Project.*

New York, McGraw-Hill, **1980**; c) R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse, C. Djerassi, *J. Org. Chem.* **1981**, *46*, 1708–1718; d) C. Benecke, R. Grund, R. Hohberger, *Anal. Chim. Acta* **1995**, *314*, 141–147; e) A. Kerber, R. Laue, T. Gruner, *Commun. Math. Co.* **1998**, *37*, 205–208; f) T. Gruner, A. Kerber, R. Laue, *Commun. Math. Co.* **1999**, *39*, 135–137.

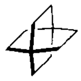
[8] a) B. D. McKay, *Congressus Numerantium* **1981**, *30*, 45–87.

[9] Many small-ring combinations are highly strained and unstable, such as tetrahedrane or prismane. Compounds containing multiple cyclopropanes are known, but their number is insignificant compared to the combinatorial possibilities. For a striking example of molecules with multiple cyclopropanes, see A. de Meijere, M. von Seebach, S. Zöllner, S. I. Kozhushkov, V. N. Belov, R. Boese, T. Haumann, J. Benet-Buchholz, D. S. Yufit, J. A. K. Howard, *Chem. Eur. J.* **2001**, *7*, 4021–4034.

[10] Nonplanar graphs cannot be drawn in a plane without crossing edges (bonds) between nodes (atoms), and contain the $K_{3,3}$ graph as a subgraph. Tricyclic bridgeheads occur in tricyclo[2.2.2.2]decane and related compounds, and are highly distorted.



$K_{3,3}$ graph



tricyclo[2.2.2.2]decane

[11] S. Bohanec, M. Perdih, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 719–726.

[12] Unstable combinations include: bridgehead olefins, bonds between heteroatoms (except in hydrazones, oximes, nitro, and in certain aromatic heterocycles), acyl halide, enamines, acyclic imines, enols, hemiacetals, orthoesters, and similar hydrolytically labile functions. Triple bonds were not used except for nitriles, and allenes were not used.

[13] a) D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36; b) D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

[14] a) J. Gasteiger, C. Rudolph, J. Sadowski, *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547; b) J. Sadowski, J. Gasteiger, *Chem. Rev.* **1993**, *93*, 2567–2581; c) http://www.mol-net.de/index.html.

[15] Tricyclo[3.3.3.0]undecane is present in Rdb as 1-Aza-tricyclo[3.3.3.0]undecane and 1,5-diaza-tricyclo[3.3.3.0]undecane.

[16] http://dtp.nci.nih.gov/index.html

[17] http://www.camsoft.com.

[18] A ring type is a graph not containing any node of connectivity 1. The Chemical Abstracts Registry or Beilstein databases are suitable for comparison. A simple total count comparison would be of little value since many entries in these databases correspond to isotopic combinations and salts of the same compounds, and sometimes to theoretical molecules that have never been synthesized.

[19] logP was calculated according to: A. K. Ghose, G. M. Crippen, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21–35.

[20] Topological polar surface area was calculated according to: P. Ertl, B. Rohde, P. Selzer, *J. Med. Chem.* **2000**, *43*, 3714–3717.

[21] V. J. Gillet, P. Willett, J. Bradshaw, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.

[22] a) S.-S Liu, C.-S. Yin, Z.-L. Li, S.-X. Cai, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321–329; b) N. Stiefl, K. Baumann, *J. Med. Chem.* **2003**, *46*, 1390–1407.

[23] J. Gasteiger, A. Teckentrup, L. Terfloth, S. Spycher, *J. Phys. Org. Chem.* **2003**, *16*, 232–245.

[24] Sets of active and inactive compounds for a specific drug target serve as an input for this application. After fragmentation of those compounds a pharmacophore model is created which is able to give an activity score for unknown compounds. http://www.molinspiration.com