Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery

Tobias Fink and Jean-Louis Reymond*

Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3, CH-3012 Berne, Switzerland

Received September 27, 2006

All molecules of up to 11 atoms of C, N, O, and F possible under consideration of simple valency, chemical stability, and synthetic feasibility rules were generated and collected in a database (**GDB**). **GDB** contains 26.4 million molecules (110.9 million stereoisomers), including three- and four-membered rings and triple bonds. By comparison, only 63 857 compounds of up to 11 atoms were found in public databases (a combination of PubChem, ChemACX, ChemSCX, NCI open database, and the Merck Index). A total of 538 of the 1208 ring systems in **GDB** are currently unknown in the CAS Registry and Beilstein databases in any carbon/heteroatom/multiple-bond combination or as a substructure. Over 70% of **GDB** molecules are chiral. Because of their small size, all compounds obey Lipinski's bioavailability rule. A total of 13.2 million compounds also follow Congreve's "Rule of 3" for lead-likeness. A Kohonen map trained with autocorrelation descriptors organizes **GDB** according to compound classes and shows that leadlike compounds are most abundant in chiral regions of fused carbocycles and fused heterocycles. The projection of known compounds into this map indicates large uncharted areas of chemical space. The potential of **GDB** for drug discovery is illustrated by virtual screening for kinase inhibitors, G-protein coupled receptor ligands, and ion-channel modulators. The database is available from the author's Web page.

INTRODUCTION

Drug discovery today critically depends on the high-throughput screening of compound libraries in silico and in vitro.[1,2] New structural types, so-called chemotypes, are of particular interest since these might display unforeseen properties and expand the scope of chemistry. Most approaches to create new compounds rely on combining known building blocks using known reactions and are not well-suited to uncover such new chemotypes, although the concept of diversity-oriented organic synthesis addresses this issue.[3−7] On the other hand, new structures with interesting properties are regularly found in natural products, suggesting that a large and literally unimaginable structural diversity is possible in organic molecules.[8−10]

Irrespective of what nature has already synthesized, it would be highly interesting to contemplate the ensemble of all possible organic molecules, which collectively form the so-called chemical universe or chemical space.[11,12] The number of molecules in chemical space might be very large but is finite if one limits oneself to a maximal molecular size of interest, for example, 300−500 Daltons as an upper limit for drug-type compounds,[13] which would provide 10^{20}−10^{200} molecules.[11,14] Compound enumeration was approached in the 1960s from a fundamental science point of view and with the practical aim of aiding spectroscopic structure determination.[15−20] However, the efforts were never pushed to an actual enumeration of all possible molecules or the analysis of their properties, most likely because of the limited computational power available at that time.

Herein, we report the generation and analysis of a database (**GDB**) of 26.4 million compounds and 110.9 million stereoisomers, enumerating all possible organic molecules of up to 11 atoms containing C, N, O, and F, under consideration of simple valency, chemical stability, and synthetic feasibility rules.[21] We detail the database generation process starting from mathematical graphs and the analysis in terms of new ring systems, stereochemistry, physicochemical properties, and compound classes. We also illustrate the possible use of **GDB** for drug discovery.

RESULTS AND DISCUSSION

Database Construction. The database was computed using an in-house-developed program written in Java (Figure 1). An exhaustive enumeration strategy was used, starting with mathematical graphs corresponding to simple hydrocarbons in which unsaturations and heteroatoms were introduced combinatorially. The resulting set of theoretical molecules was then reduced to a subset of chemically relevant compounds by applying "filters" for chemical stability and synthetic feasibility of functional groups, and the remaining molecules were analyzed for tautomeric forms and finally expanded to all possible stereoisomers, as detailed below. The database was computed up to 11 atoms of the second-row elements C, N, O, and F.

I. Graph Selection. Saturated hydrocarbons can be considered as the chemical equivalents of graphs. Carbon

* Corresponding author fax: + 41 31 631 80 57; e-mail: jean-louis.reymond@ioc.unibe.ch.

10.1021/ci600423u CCC: $37.00
Published on Web 01/30/2007

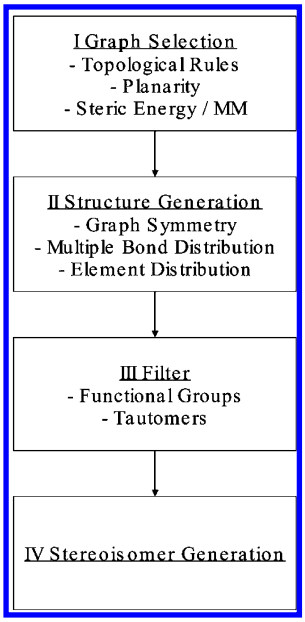

Figure 1. Assembly principle for the chemical universe database **GDB**.

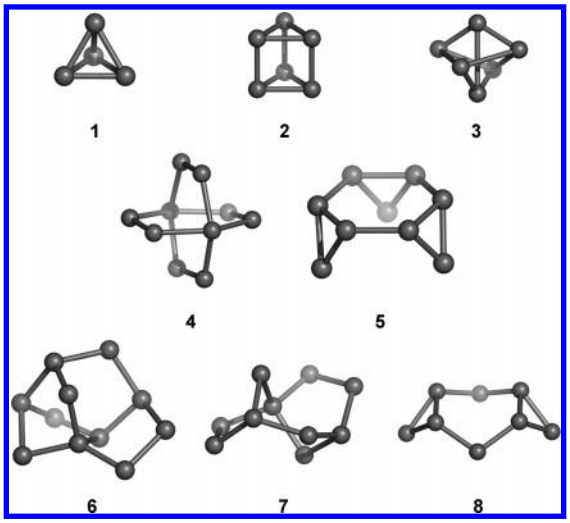

Figure 2. Examples or graphs represented as their hydrocarbons: **1**, tetrahedrane; **2**, prismane; **3**, Claus-benzol; **4**, tricyclo[2.2.2.2]-decane; **5**, tetracyclo[6.1.0.0.^{2,4}0^{5,7}]nonane (only one stereoisomer shown); **6**, highest accepted graph; **7**, lowest rejected graph; and **8**, tricyclo[5.1.0.0^{3,5}]octane (only one stereoisomer shown).

atoms correspond to graph nodes, and the carbon−carbon single bonds correspond to graph edges.[22] Hydrogen atoms are not considered in this representation and are automatically added to complement valency. The starting collection of 843 335 connected graphs up to 11 nodes with a maximum node connectivity of 4 was obtained using the program GENG, which lists all possible graphs for a given number of nodes and a given maximum node connectivity.[23] This graph collection contained a vast majority of graphs (98.1%) corresponding to highly strained polycyclic hydrocarbons. Some strained molecules such as tetrahedrane (**1**), prismane (**2**), or fused cyclopropanes are sometimes synthetically accessible but have a high steric energy content which renders them generally unstable (Figure 2).[24] Strain is particularly problematic when concentrated in single carbon atoms showing large distortions from the optimal tetrahedral geometry, for example, in tricyclo[2.2.2.2]decane (**4**) containing two pyramidal carbon atoms. On the other hand, tetracyclo[6.1.0.0.^{2,4}0^{5,7}]nonane (**5**) shows a very high total steric energy, but the steric strain is in this case equally distributed over six carbon atoms, and hence, the molecule is not synthetically problematic.[25]

Table 1. Graph Selection Table[a]

nodes	graphs[b]	passed Topo I[c]	passed Topo II[d]	planar graphs[e]	unstrained graphs[f]	with unsaturations
1	1	1	1	1	1	1
2	1	1	1	1	1	3
3	2	2	2	2	2	4
4	6	4	4	4	4	13
5	21	8	8	8	8	33
6	78	20	20	20	20	123
7	353	57	57	57	57	445
8	1929	199	194	194	194	1956
9	12 207	780	712	708	705	8863
10	89 402	3600	2893	2845	2822	43 443
11	739 335	19 215	12 575	12 169	11 912	221 336
total	**843 335**	**23 887**	**16 467**	**16 009**	**15 726**	**276 220**

[a] Each graph corresponds to a saturated hydrocarbon. [b] Generated by the program GENG[23] using a maximum connectivity of 4. [c] The Topo I filter eliminates any graph with a node present in two different three- or four-membered rings. [d] The Topo II filter eliminates graphs with a tetravalent bridgehead in a small ring. [e] Nonplanar graphs cannot be drawn in a plane without crossing edges, e.g., **3**. [f] Graphs containing highly distorted centers as determined by a energy-minimization in a MM2 force field were removed, see text.

Graphs were selected for reduced ring strain of the corresponding hydrocarbon in four steps (Table 1). The first three steps involved topological filters eliminating (a) graphs containing one or more nodes in two small (three- or four-membered) rings (Topo I), (b) graphs containing a ring system with a tetravalent bridgehead in a small ring (Topo II), and (c) all nonplanar graphs. A ring system is a graph without acyclic bonds. Nonplanar graphs cannot be drawn in a plane without crossing edges and correspond to chemically impossible hydrocarbons such as Claus-benzol (**3**, Figure 2), which is the chemical equivalent of the complete bipartite graph $K_{3,3}$.[26] These topological filters reduced the graph collection by 98.0% from 843 335 to 16 009, eliminating a plethora of fused small rings with highly distorted centers. Although famous ring systems such as cubane[27] or prismane[28] were lost in the process, their loss was negligible considering the very large number of graphs still available.

The fourth step involved the elimination of graphs containing highly distorted centers not identified by topology. To this end, the 16 009 remaining graphs were reduced to their parent ring systems. All stereoisomers and, if necessary, multiple conformations for each ring system were generated and subsequently energy-minimized using an adapted MM2 force field (details on the used force field are in the Methods section).[29] The contribution of each carbon center to the total steric energy was calculated from its deviation from the optimal tetrahedral geometry. A cutoff value of +17 kcal/mol was selected, which separated **6** as the highest accepted ring system from **7** as the first rejected ring system (Figure S1 of the Supporting Information and Figure 2). This procedure was necessary to identify distorted graphs such as the above-mentioned tricyclo[2.2.2.2]decane **4**. This molecular-mechanics-based procedure eliminated 125 (9.3%) of the 1349 ring systems tested, corresponding to 283 graphs (1.8%) of the 16 009 graphs retained after the topology filters, leaving a final set of 15 726 graphs for structure generation.

Table 2. Overview of the Structure Generation Process

nodes	graphs[a]	generated[b]	accepted[c]	unique tautomers (**GDB**)[d]	all tautomers	stereoisomers[e]
1	1	4	4	4	4	4
2	1	10	9	9	9	9
3	2	52	20	20	21	20
4	4	332	80	80	88	87
5	8	2294	357	352	397	469
6	20	18 066	1906	1850	2135	2911
7	57	154 542	10 953	10 568	12 438	19 904
8	194	1 445 073	69 563	66 706	79 899	153 601
9	705	14 213 741	464 402	444 313	540 002	1 258 963
10	2822	146 004 340	3 259 036	3 114 041	3 827 907	10 898 065
11	11 912	1 558 491 448	23 875 101	22 796 628	28 240 425	98 645 474
total	**15 726**	**1 720 329 902**	**27 681 431**	**26 434 571**	**32 703 325**	**110 979 507**

[a] Number of graphs selected for molecule generation, see Table 1. [b] Number of molecules generated by introducing atom types at each node using the following rules: nodes with connectivity 4, C only; connectivity 3, C and N; connectivity 2, C, N, and O; connectivity 1, C, N, O, and F. [c] Number of molecules accepted as chemically stable and feasible following functional group selection rules listed in Table S1 (Supporting Information). [d] Number of accepted molecules remaining after elimination of equivalent tautomers. [e] Number of stereoisomers obtained from unique tautomers (**GDB**). The generation of **GDB** for 12 atoms exceeded the computing power available to us.

II. Structure Generation. The graph symmetry of the selected 15 726 graphs was determined to allow the introduction of unsaturations and element types without generating duplicates. Graph symmetry was derived from the graph automorphism group perception algorithm described by Bohanec and Perdin,[30] using the priority rules defined by Weininger et al.,[31] adding the frequencies at which an atom occurs in rings of size 3−7 as supplementary atomic invariants. This procedure derived the correct graph symmetry in all 15 726 graphs, including problem cases such as tricyclo[5.1.0.0^{3,5}]octane **8** (Figure 2). Double and triple bonds were then introduced combinatorially, resulting in 276 220 molecular graphs, excluding bridgehead double bonds, triple bonds in rings smaller than nine, and allenes,[32] all considered potentially problematic for synthesis.

All mathematically possible molecular structures were generated from the molecular graphs by introducing the four second-row elements C, N, O, and F under consideration of valency rules, allowing fluorine (as a representative halogen) at methyl (CH_3) positions only; oxygen at methyl and methylene (CH_2) group; and nitrogen at methyl, methylene, and methyne (CH) positions. The procedure was carried out under consideration of the reduced graph symmetry of each of the 276 220 molecular graphs taking unsaturations into account, avoiding the generation of duplicates. The combinatorial distribution resulted in over 1.7 billion unique structures, which we call the "dark matter universe" (**DMU**). Sulfur was not considered in order to avoid a much larger combinatorial explosion. Sulfur-containing molecules, which make up 25.5% of known drugs in the Ashgate drug database,[33] can be derived from oxygen analogs using simple substitution rules (O → S, C═O → S═O or SO_2).

III. Filters for Chemical Stability, Tautomers, and Aromaticity. The 1.7 billion molecules of **DMU** are not useful as an entry into the chemical universe or as possible molecules. Indeed, almost all molecules in this ensemble contain one or more unstable functional groups, in particular, heteroatom−heteroatom bonds (e.g., N−F and N−O−O), gem-diols, aminals, enols, orthoacids, acyl fluorides, and similarly labile functional groups. These were removed by applying "filters" identifying these substructures (Table S1, Supporting Information), leaving 27.7 million structures (1.6%) as possible molecules. These molecules were analyzed for redundant tautomeric forms by exhaustively generating all tautomers of each compound and eliminating doubles, resulting in a final collection of 26.4 million structures. The tautomer analysis included the correct identification of aromatic rings, whereby aromaticity was determined according to the qualitative molecular orbital model of Frost and Musulin.[34,35] The database was formulated either as a set of unique tautomers retaining the most probable tautomer in each case, referred to here as **GDB**, or as a database listing all possible tautomers of each compound except for nonaromatic enols and enamines.

IV. Stereoisomer Generation. Stereoisomers were enumerated following an elegant algorithm described by Djerassi et al.,[36] which correctly identifies all asymmetric centers and *Z*/*E* double bonds in a molecule. The algorithm was improved by including a determination of the correct relative stereochemistry for bridgehead pairs, such as in bicyclo[2.2.2]-octane. *Z*/*E* isomerism was blocked for double bonds inside rings smaller than 10. Atropisomerism due to axial chirality was not considered in this analysis. A total of 110.9 million stereoisomers were generated from the 26.4 million structures in **GDB**, corresponding to an average of 4.2 stereoisomers per molecule (Table 2).

Overview of the Structure Generation Process. The number of molecules produced by the structure generation process increased exponentially with the square of the number of atoms, such that more than 90% of **GDB** was molecules of 11 atoms (Table 2). The number of molecules generated was also strongly dependent on the graph type and varied between 4 and 97 274 molecules per graph. Monocyclic and bicyclic graphs gave rise to the largest numbers of molecules (Figure 3). The number of molecules generated also varied strongly according to the elemental composition, with the largest diversity arising from molecules containing C, N, and O (Table 3).

Database Analysis. GDB was analyzed for (1) new ring systems, (2) stereochemistry, (3) physicochemical properties, (4) compound classes, and (5) drug and lead discovery. The database was compared to a reference database of known compounds obtained by extracting all organic molecules[37] of up to 11 atoms from the PubChem,[38] ChemACX,[39]

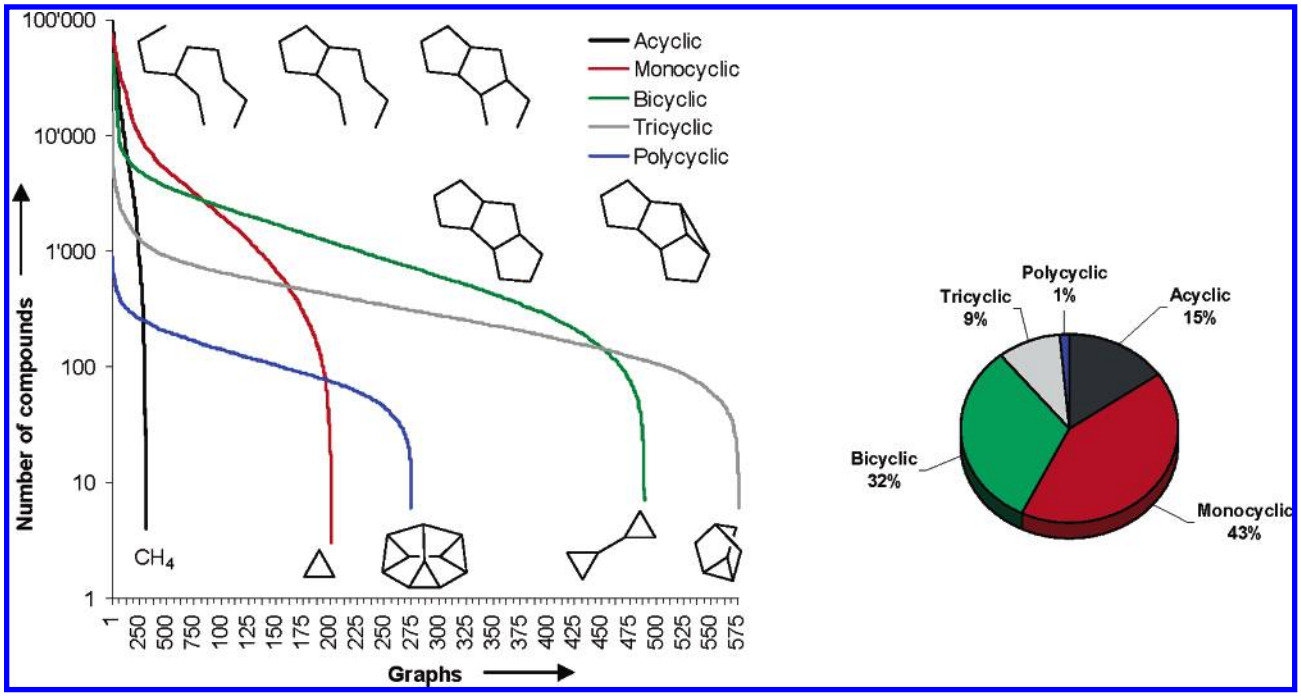

Figure 3. Number of molecules generated per graph, in descending order. Hydrocarbons corresponding to the graphs giving the largest (upper structures) and the smallest (lower structures) number of compounds are shown for each category.

Table 3. Number of Compounds in **GDB** as a Function of Elemental Composition (Rows) and Number of Heavy Atoms (Columns)

	number of heavy atoms											
elemental composition	1	2	3	4	5	6	7	8	9	10	11	**total**
C	1	3	4	12	29	102	347	1468	6413	30 582	152 117	**191 078**
N	1	0	0	0	0	0	0	0	0	0	0	**1**
O	1	1	0	0	0	0	0	0	0	0	0	**2**
F	1	1	0	0	0	0	0	0	0	0	0	**2**
CN	0	1	6	22	91	443	2255	12 832	76 063	472 756	3 049 435	**3 613 904**
CO	0	2	5	19	74	338	1671	9302	54 733	337 024	2 164 860	**2 568 028**
CF	0	1	3	12	39	151	622	2954	14 598	77 225	429 664	**525 269**
CNO	0	0	2	11	82	526	3578	24 858	176 888	1 299 010	9 819 032	**11 323 987**
CNF	0	0	0	2	17	122	818	5594	39 052	279 803	2 059 976	**2 385 384**
COF	0	0	0	2	18	126	809	5437	37 148	260 489	1 883 487	**2 187 516**
CNOF	0	0	0	0	2	42	468	4401	39 910	358 528	3 236 049	**3 639 400**
total	**4**	**9**	**20**	**80**	**352**	**1850**	**10 568**	**66 846**	**444 805**	**3 115 417**	**22 794 620**	**26 434 571**

ChemSCX,[40] NCI open databases[41] and the Merck Index,[42] referred to as **RDB**. This reference database contained 63 857 unique compounds after neutralization, the removal of counter ions, the replacement of isotopes, and the substitution of fluorine for all halogen atoms. A total of 37 393 (58.6%) compounds from **RDB** were found in **GDB**. The 26 464 **RDB** compounds not found in **GDB** contained features not considered during database generation, including graphs violating the topological rules used for graph selection (Topo I or Topo II: 218 compounds) or the restrictions applied on multiple bonds (allenes or bridgehead olefins: 629); molecules with sulfur (14 048), phosphorus (1661), silicon (1481), or combinations of these three elements (645); molecules not considered for chemical stability criteria, in particular hydrolytic stability (e.g., acyl halides, hemiacetals: 7774); and radicals (8). We also extracted 25 000 000 compounds at random from the 1.7 billion theoretical structures (**DMU**) obtained during database generation prior to the application of filters.

1. New Ring Systems. The 15 726 saturated hydrocarbon graphs selected for molecule generation were analyzed for new structural types. The graphs were fragmented into acyclic graphs (graphs without cyclic bonds) and ring systems (graphs without acyclic bonds). Both purely acyclic and purely cyclic graphs were analyzed for occurrence in the CAS Registry (SciFinder) and the Beilstein database either as pure hydrocarbons or in any heteroatom and/or multiple bond, including as substructures. All of the 309 acyclic graphs found in **GDB** were found in molecules from these two databases. By contrast, only 670 (55.5%) of the 1208 ring systems present in **GDB** could be identified in molecules from the CAS Registry or Beilstein database (Table 4). Interestingly, 367 (68.2%) of the 538 unknown ring systems are chiral, but only 310 (46.3%) of the 670 known ring systems are chiral, suggesting chiral ring systems might be more difficult to make or to imagine. Compounds **9**−**16** in Figure 4 exemplify such yet unknown ring systems (Figure 4). Note the particularly elegant chiral, *C*2-symmetrical, and yet unknown ring system **16** containing three bicyclo[2.2.1]heptane systems. The related chiral *C*3-symmetrical ring system **17** containing three symmetry-related bicyclo[2.2.1]heptane systems was reported in 1976 in its trioxa version, which was obtained from the acid-catalyzed cyclization of the *C*3-symmetrical tris-epoxide of bicyclo[2.2.2]octa-triene.[43]

Table 4. Classification of Ring Systems According to Number of Three- (Rows) and Four-Membered (Columns) Rings[a]

number of three-membered rings	number of four-membered rings			
	0	1	2	**total**
0	124 [3]	189 [60]	103 [67]	**416 [130]**
1	225 [50]	238 [177]	20 [19]	**483 [246]**
2	201 [88]	55 [48]		**256 [136]**
3	53 [26]			**53 [26]**
total	**603 [167]**	**482 [285]**	**123 [86]**	**1208 [538]**

[a] Main number: number of ring systems. Number in brackets: unknown ring systems that were not found in **RDB**, the MDL Beilstein database, or the CAS registry database, in any heteroatom and multiple combination, including substructures. The three unknown ring systems without any three- or four-membered rings are shown as **14**−**16** in Figure 4. Stereochemical composition: chiral ring systems, 677 (367 unknown); achiral ring systems, 455 (150 unknown); ring systems with both chiral and achiral stereoisomers, 76 (21 unknown).

2. Stereochemistry. Stereochemistry in organic molecules may arise from the graphs themselves. For instance, there are at least two stereoisomers for 166 (53.7%) of the 309 acyclic graphs, 849 (70.3%) of the 1208 ring systems, and 13 859 (88.1%) of the 15 726 graphs. At least one chiral stereoisomer (corresponding to two enantiomers) is found in 166 (53.7%) of the acyclic graphs, 753 (62.3%) of the ring systems, and 12 771 (81.2%) of the graphs. Additional stereochemistry is created through the introduction of unsaturations and heteroatoms at the nodes of both chiral and achiral graphs to form molecules.

Stereochemistry was first analyzed as a function of molecular size, which showed that very small molecules below five heavy atoms were mostly achiral, with some exceptions such as propylene oxide, the smallest chiral molecule with only four atoms. On the other hand, over two-thirds of all molecules with 10 and 11 atoms were chiral, with a clear trend suggesting that almost all possible structures are chiral for molecules of 12 or more atoms (Figure 5). Only very few structures (0.5% of **GDB**) had both chiral and achiral stereoisomers, for example, 2,3-butane diol with a chiral pair and one meso form.

The prevalence of chirality in larger molecules simply followed from the larger number of stereocenters. Molecules with 11 atoms thus contained the largest stereochemical diversity, with an average of 4.3 stereoisomers per molecule, compared to 3.4 stereoisomers per molecule for those with 10 atoms or less (Table 2). A comparison of the number of stereoisomers versus the number of stereocenters in a molecule showed that the most abundant stereochemical classes of molecules (4 023 245 structures) were those with one stereocenter and two stereoisomers, corresponding to simple enantiomeric pairs generated by a single asymmetric carbon center (Table 5). There were 66 molecules with 10 stereocenters but only one possible stereoisomer (because of polycyclic constraints) and 1395 molecules with 64 stereoisomers (generated from five or six stereocenters). A similar comparison in terms of the number of chiral versus achiral stereoisomers per molecule showed that almost one-third of the database (7 984 499 structures) was purely chiral molecules with four different stereoisomers (Table 6). On the other hand, there was a comparable number of purely achiral molecules (7 052 511 structures), with more than half of these having only a single stereoisomer (3 970 798 structures). Examples of stereoisomeric molecules are given in Figure 6.

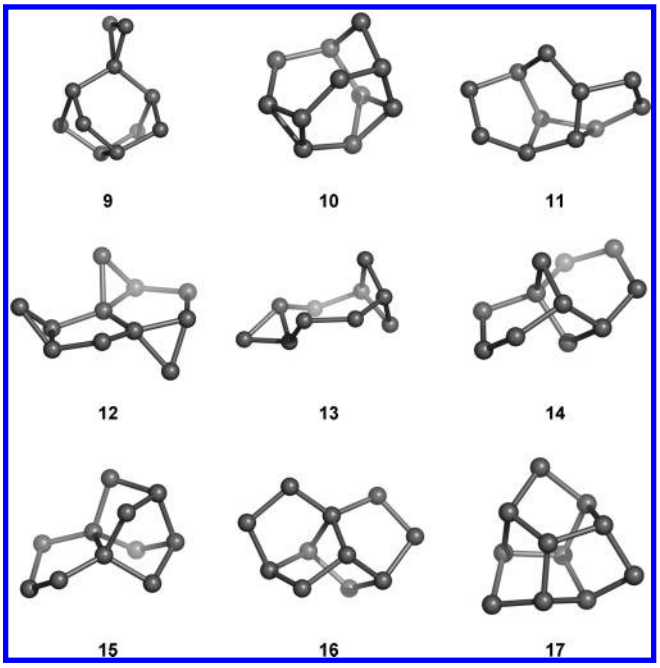

Figure 4. Examples of ring systems. **9**−**16** are unknown either as pure hydrocarbons or in any heteroatom and multiple-bond combination including as substructures. The C2-symmetrical chiral **16** contains three bicyclo[2.2.1]heptane ring systems. The chiral C3-symmetrical ring system **17** contains three symmetry-related bicyclo[2.2.1]heptane ring systems and is known as the trioxa derivative (oxygen atoms at the three divalent nodes).[43]

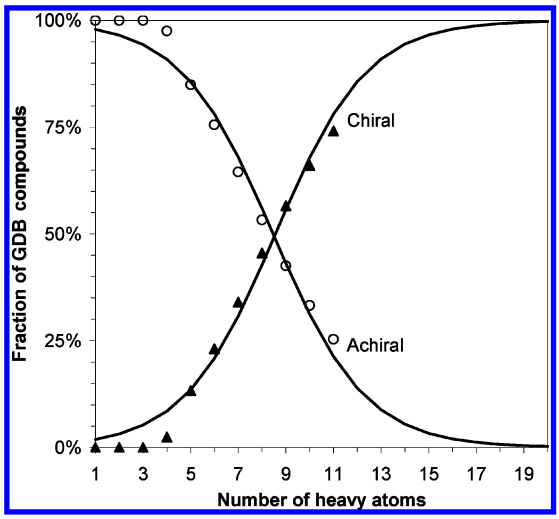

Figure 5. Stereoisomer distribution in **GDB**.

3. Physicochemical Properties. The predicted physicochemical properties of **GDB** compounds, which are particularly important with respect to drug properties, were computed and compared with those of compounds in the reference database **RDB** and in the "dark matter universe" **DMU**. Descriptor values were determined for the molecular weight (MW) and the number of rotatable bonds in the molecule (RBC), which describe molecular size and flexibility, as well as for the octanol−water partition coefficient (logP),[44] the topological polar surface area (TPSA),[45] the number of hydrogen-bond donor (HBDC) and acceptor (HBAC) sites, which all describe molecular polarity and are primarily a function of heteroatom density in a molecule

Table 5. Number of Stereocenters vs Number of Stereoisomers for GDB Compounds

number of stereoisomers[b]	number of stereocenters[a]											
	0	1	2	3	4	5	6	7	8	9	10	**total**
1	3 805 751[c]	0	127 959	0	25 340	186	10 245	68	1125	58	66	**3 970 798**
2	2 193 503[d]	4 023 245	1 160 206	189 290	317 283	33 036	88 602	7945	7488	228	158	**8 020 984**
3	2481[d]	0	13 182	0	2493	0	115	0	31	0	0	**18 302**
4	572 362[d]	1 664 209	3 524 711	2 137 679	380 906	320 836	53 880	43 573	3240	889	5	**8 702 290**
5	0	0	0	0	61	0	14	0	0	0	0	**75**
6	229[d]	38	1860	1253	422	87	256	14	3	0	0	**4162**
7	0	0	0	0	6	0	0	0	0	0	0	**6**
8	49 980[d]	188 701	733 641	1 869 820	1 265 196	362 856	93 685	19 829	3 794	130	0	**4 587 632**
10	62[d]	0	300	0	1537	52	521	12	26	0	0	**2510**
12	0	0	142	164	510	44	4	0	0	0	0	**864**
16	991[d]	3822	29 787	165 346	443 339	286 119	108 252	8965	1096	89	0	**1 047 806**
20	4[d]	0	0	0	6	0	18	0	1	0	0	**29**
24	0	0	0	0	14	107	0	0	0	0	0	**121**
32	19[d]	0	118	2208	11 041	41 237	17 192	5605	134	1	0	**77 555**
36	0	0	0	0	0	0	42	0	0	0	0	**42**
64	0	0	0	0	0	186	1209	0	0	0	0	**1395**
total	6 625 382	5 880 015	5 591 906	4 365 760	2 448 154	1 044 746	374 035	86 011	16 938	1395	229	**26 434 571**

[a] A stereocenter is defined as an atom at which the interchange of two groups produces a stereoisomer, with the exception of *E*/*Z* isomers. [b] Total number of stereoisomers resulting from a compound. [c] Molecules containing no stereocenters and no *E*/*Z* double bonds. [d] Molecules resulting in *E*/*Z* isomers only.

Table 6. Number of Chiral and Achiral Stereoisomers for GDB Compounds

chiral stereoisomers[b]	achiral stereoisomers[a]												
	0	1	2	3	4	5	6	8	10	16	20	32	**total**
0	0	3 970 798	2 380 175	3046	636 839	75	257	59 670	74	1554	4	19	**7 052 511**
2	5 640 809	15 256	80 952	0	0	0	0	0	0	0	0	0	**5 737 017**
4	7 984 499	0	1967	6	29 712	0	0	0	0	0	0	0	**8 016 184**
6	1938	0	0	0	84	0	0	0	0	0	0	0	**2022**
8	4 498 250	0	2268	0	131	0	0	2730	0	0	0	0	**4 503 379**
10	84	0	22	0	0	0	0	0	0	0	0	0	**106**
12	711	0	0	0	1214	0	0	0	0	0	0	0	**1925**
16	1 042 308	0	0	0	6	0	0	0	0	0	0	0	**1 042 314**
20	19	0	0	0	0	0	0	0	0	0	0	0	**19**
24	121	0	0	0	0	0	0	0	0	0	0	0	**121**
32	77 536	0	0	0	42	0	0	0	0	0	0	0	**77 578**
64	1395	0	0	0	0	0	0	0	0	0	0	0	**1395**
total	**19 247 670**	**3 986 054**	**2 465 384**	**3052**	**668 028**	**75**	**257**	**62 400**	**74**	**1554**	**4**	**19**	**26 434 571**

[a] Stereoisomers containing a mirror plane or a center of inversion are considered achiral. [b] Stereoisomers containing neither a mirror plane nor a center of inversion are considered chiral.

(Figure S2, Supporting Information). **GDB** showed a very sharp distribution of MW around 153 ± 7 Da, assessing the combinatorial enumeration of molecules with 11 atoms, which form 92% of the database. **DMU** showed an even sharper MW distribution (158 ± 7 Da), while the reference set **RDB** was shallower with a maximum at a lower MW (141 ± 24 Da), assessing the only very partial coverage of combinatorial possibilities in known molecules. The distribution of molecules in the property space described by the six selected molecular properties was visualized by principal component analysis (PCA; Figure 7), showing that **GDB** contained compounds in high-polarity regions (high PC1) not covered by **RDB**, which can be attributed to the presence of multiple polar functional group combinations. The compounds from **DMU**, which contain a very high proportion of heteroatoms arising from forbidden heteroatom combinations, expanded to even higher PC1 values. The pure combinatorial nature of **DMU** is apparent in the smoothness of the occupancy of PC space (Figure 7, 3D plot at right). The uneven distribution of compounds in this space by **GDB** reflects the effect of the filters used to eliminate chemically impossible or unstable structural elements.

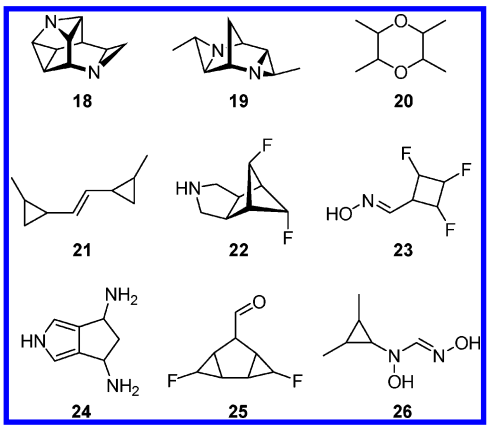

Figure 6. Examples of compounds. **18**: 10 stereocenters and four chiral stereoisomers. **19**: eight stereocenters and 20 chiral stereoisomers. **20**: four stereocenters and four chiral and three achiral stereoisomers. **21**: four stereocenters and 16 chiral and four achiral stereoisomers. **22**: six stereocenters and six chiral and four achiral stereoisomers. **23**: four stereocenters and eight chiral and eight achiral stereoisomers. **24**: two stereocenters and two chiral and one achiral stereoisomers. **25**: seven stereocenters and 12 chiral and four achiral stereoisomers. **26**: three stereocenters and four chiral and four achiral stereoisomers.

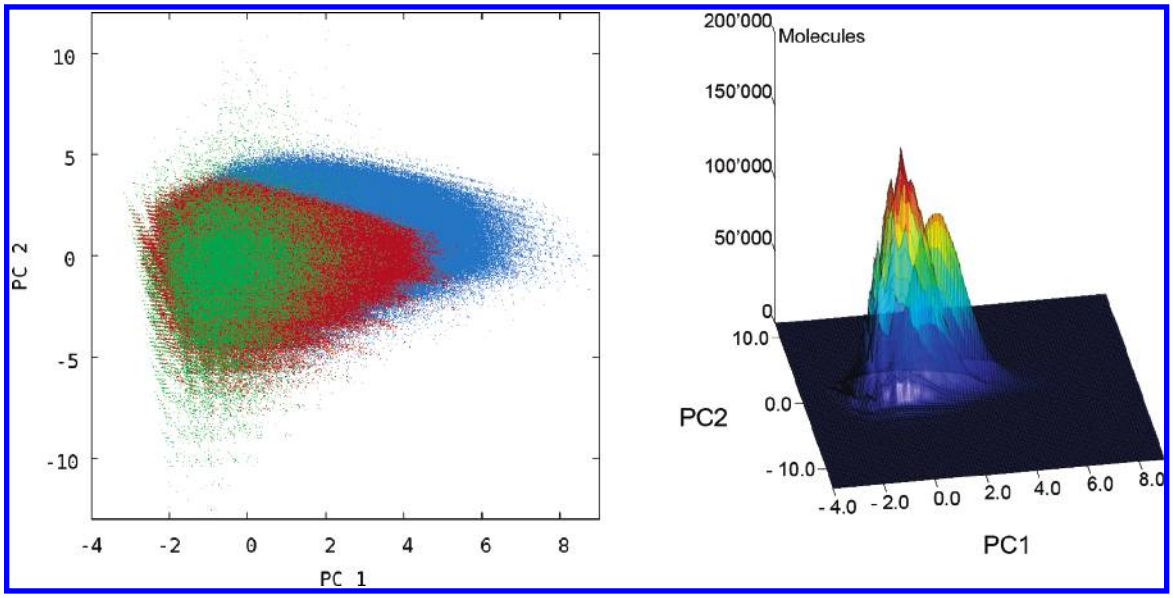

Figure 7. PCA plot including **GDB** (left, red; right, spiky mountain), **RDB** (left, green; right, not shown), and **DMU** (left, blue; right, smooth mountain). Principal component analysis of **GDB** resulted in the following loadings: PC1 (50.16% of variance): MW 0.014, logP −0.302, HBDC 0.282, HBAC 0.259, TPSA 0.311, RBC −0.092. PC2 (19.63% of variance): MW 0.609, logP 0.147, HBDC −0.050, HBAC −0.016, TPSA −0.015, RBC 0.647. **RDB** covers, albeit sparsely, a broader range of PC2 values than **GDB** and **DMU** because of the presence of heavy heteroatoms (P, S, and Si) not included in **GDB** and **DMU** (high PC2) and the presence of reactive small molecules rejected by our filters, such as isonitriles, isocyanates, or phosgene (low PC2 values).

Because of their small size, all of the **GDB** database molecules obeyed Lipinski's "Rule of 5" for predicting bioavailability, which stipulates that no more than two descriptor values should fall outside the limits of MW < 500, logP < 5, HBDC < 5, and HBAC < 10.[46] Furthermore, half of the **GDB** compounds (13.2 million) followed the more restrictive "Rule of 3" (RO3) used to select leadlike molecules, under which all parameter values must fall below the limits of MW < 300, RBC < 3, logP < 3, HBDC < 3, HBAC < 3, and TPSA < 60 Å^2.[47] A large proportion of the leadlike compounds (61.5%) contained three- or four-membered rings or triple bonds, while these were much less frequent in the nonleadlike series (41.4%). A total of 8.6 million of the nonleadlike compounds in **GDB** broke RO3 for being too polar (HBDC, HBAC, TPSA), 2.6 million because they were too flexible (RBC), another 1.8 million because of the combination of both, and only 90 840 because of low polarity (high logP).

4. Compound Classes. Organic chemistry classifies compounds by classes such as heteroaromatics (e.g., pyridine, 11.3% of **GDB**), aromatics (e.g., benzene, 0.1% of **GDB**), fused heterocycles (e.g., cyclohexene-oxide, 29.5% of **GDB**), fused carbocycles (e.g., decaline, 4.8% of **GDB**), heterocycles (e.g., tetrahydrofurane, 27.3% of **GDB**), carbocycles (e.g., cyclopentanol, 11.5% of **GDB**), and acyclic compounds with at least one nonterminal heteroatom (here, "heteroacyclic", e.g., ethyl acetate, 11.2% of **GDB**) or a continuous carbon chain (here, "carboacyclic", e.g., ethylene glycol, 4.2% of **GDB**). To examine how these compound classes were distributed in **GDB**, we generated a Kohonen map of the database using autocorrelation descriptors. Autocorrelation descriptors represent pairwise topological relationships between atoms to which properties have been assigned.[48−50] They were determined as described by Moreau and Broto[48] using partial σ and π charges,[51] atomic polarizability[52] and the topological steric effect index,[53] the atomic number, and the identity function (= 1 for each atom) as atomic properties over a distance of 0−7 bonds, resulting in a 48-dimensional vector for each structure in **GDB** (eq 1)

$$\mathrm{AC}_d = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta(p_i p_j)_d \qquad (1)$$

where d = the considered topological distance, N = the number of atoms, p = the atomic property, and

$$\delta_{ij} = \begin{cases} 1 & \text{if } d_{ij} = d \\ 0 & \text{if } d_{ij} \neq d \end{cases}$$

(Kronecker Delta)

Kohonen maps,[54,55] also called "self-organizing maps", are a specialized subtype of artificial neural networks and consist of a grid of neurons on a toroidal surface conveniently represented as its two-dimensional projection in a plane. Each neuron is initially assigned a random weight vector, here consisting of 48 numerical values according to the dimensionality of the input autocorrelation vectors. Learning is achieved by repeatedly presenting series of input vectors to the map, determining the most similar neuron to each input vector in respect to the Euclidean distance, and gradually adjusting the weight vector of this neuron and its topological neighbors to become more similar to the presented input vector. The learning process results in an organized pattern grouping similar input vectors in topologically close neurons. For the present analysis, a Kohonen map consisting of 200 × 200 neurons was trained using autocorrelation vectors of 1 000 000 randomly chosen **GDB** molecules for 250 000 epochs, with 100 molecules presented per epoch. After the training was completed, all **GDB** and **RDB** compounds were presented to the map in the form of their autocorrelation vectors and assigned to the neuron with the most similar weight vector.

The different compound classes appeared in separated regions of the Kohonen map, revealing a well-defined

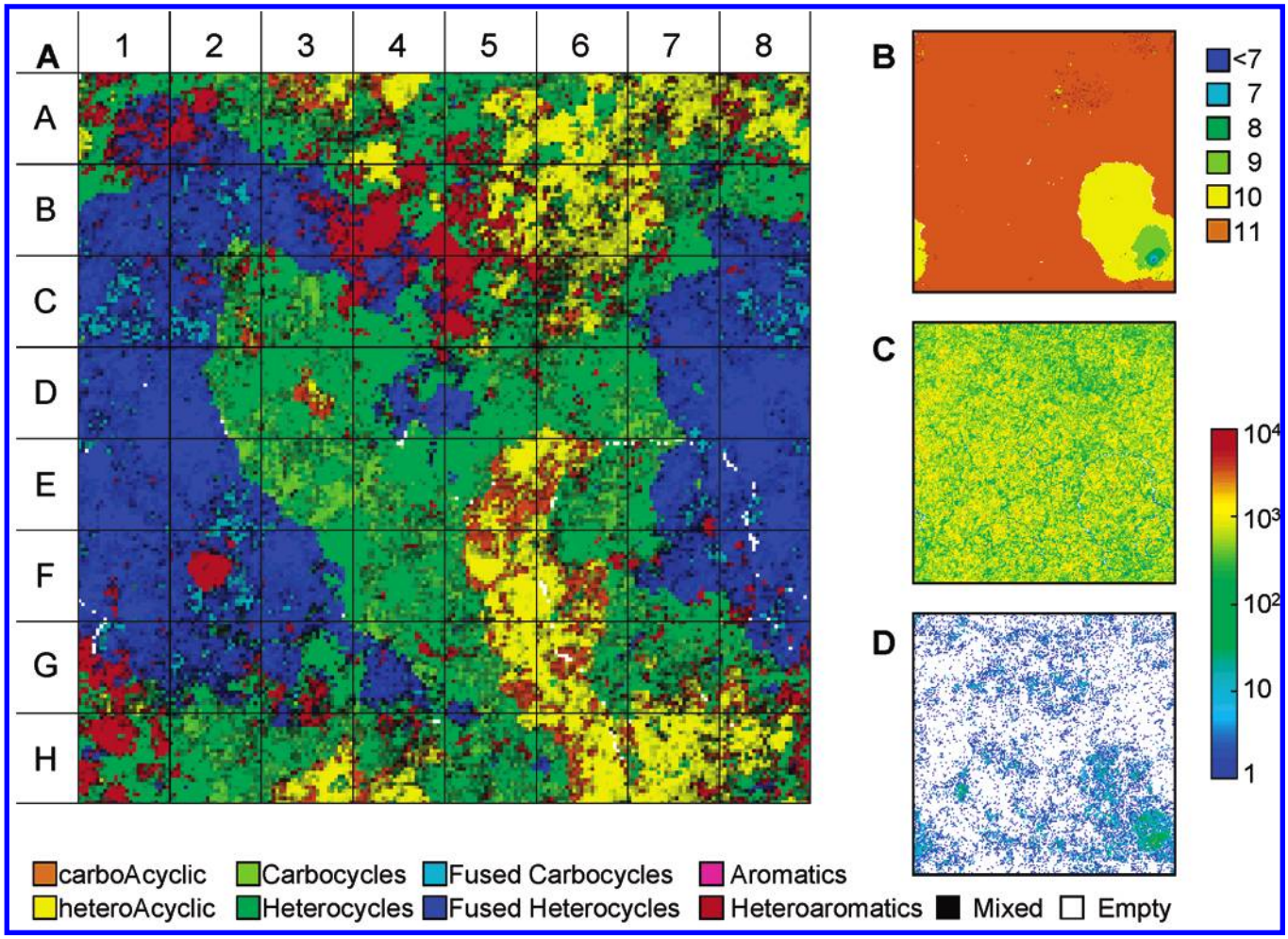

Figure 8. Structural types landscape of **GDB**. (A) Compound type distribution. Neurons are colored according to the most abundant compound class in the neuron according to the color code given below. The color is weighed to grayscale in proportion to all other combined classes also present in the neuron. Each molecule is assigned to a single category in the following priority order: heteroaromatic > aromatic > fused heterocycle > fused carbocycle > heterocycle > carbocycle > heteroacyclic > carboacyclic. (B) Color-coding according to the most frequent number of atoms per molecule in each neuron following the color code given at right. (C and D) Color-coding according to neuron occupancy by **GDB** (C) and **RDB** (D) following the color scale at right.

landscape for **GDB** (Figure 8). Heteroacyclic (yellow) and carboacyclic compounds (orange) formed arid islands surrounded by forested regions of carbocycles (lime green), heterocycles (green), and a few concentrated regions with wildfires of aromatics (purple) and heteroaromatic (red) compounds. These cyclic compounds formed a coastal system touching an ocean of fused carbocycles (cyan) and fused heterocycles (blue). An analysis of neuron occupancy by **GDB** compounds (Figure 8C) showed an even distribution across the entire map, with between 1 and 3627 molecules per neuron. There were only 95 empty neurons in the occupancy map of **GDB**, which encircled all compounds of 10 atoms and less forming an island at the lower right portion of the map. By contrast, **RDB** compounds were found in only 32% of the neurons (Figure 8D), leaving large portions of the map unoccupied. The only area well-covered by **RDB** was the inside portion of the small molecule island mentioned above, which contained all compounds of nine atoms and smaller clustered around three neurons containing molecules with less than seven atoms (Figure 8A). The dense occupation of **RDB** in this area suggests that the chemical space of these small molecules is already fairly well sampled in known molecules.

The organization of **GDB** by the Kohonen map also segregated compounds on the basis of their chirality and their leadlike properties. Thus, color-coding by the percentage of chiral molecules in each neuron showed that chiral molecules were mostly present among fused carbocycles and fused heterocycles (Figure 9A). Color-coding for lead-likeness (following the "Rule of 3" discussed above) showed that these chiral regions also concentrated on the most leadlike molecules, suggesting that chirality is an interesting property to consider for drug design (Figure 9B). The more leadlike nature of fused ring systems is probably related to their structural rigidity and a lower number of terminal positions constraining heteroatoms to mostly secondary and tertiary positions, which should lower polarity by allowing fewer hydrogen-bond donor sites. The leadlike regions also contained most compounds identified by virtual screening for drug classes as discussed below (Figure 9C). A selection of compounds corresponding to different regions of the Kohonen map is shown in Figure 10.

5. Drug and Lead Discovery. As mentioned in the Introduction, the major possible application of **GDB** could be the identification of new drugs or lead compounds. The

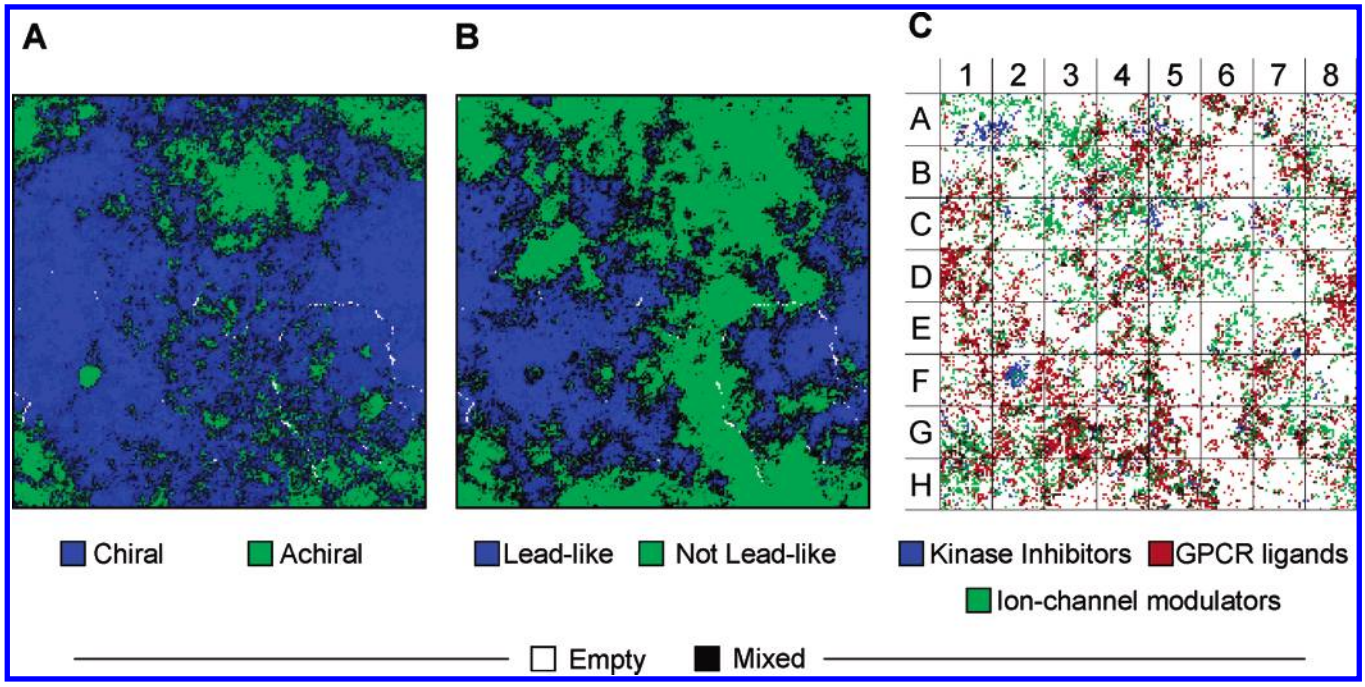

Figure 9. Kohonen maps of **GDB**. (A) Color-coding according to the presence of chiral vs achiral compounds. Mixed neurons are in grayscale. (B) Color-coding according to leadlike vs nonleadlike compounds as defined by the "Rule of 3" (see main text). (C) Color-coding according to the presence of virtual hits as obtained by virtual screening using the Molinspiration toolkit for three different drug classes (see main text).

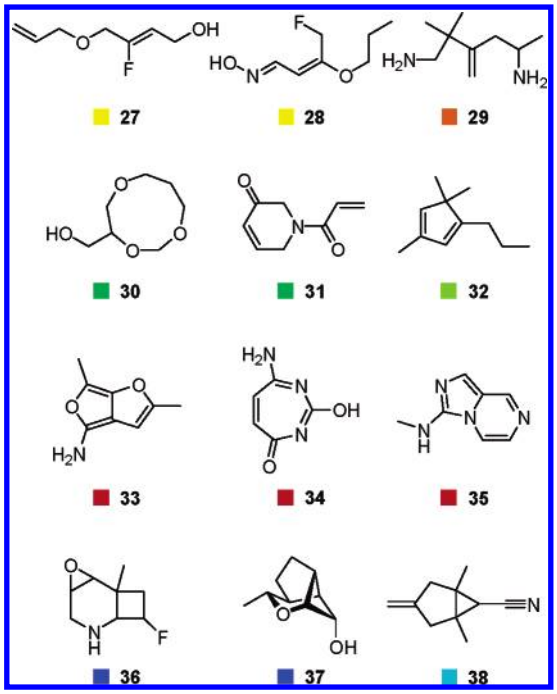

Figure 10. Examples of **GDB** compounds. Their location in the Kohonen map of Figure 7 is given in parentheses. **27**: heteroacyclic (center of H7). **28**: heteroacyclic (border between G5 and G6). **29**: carboacyclic (border between E5 and E6). **30**: heterocycle (upper-left corner of E5). **31**: heterocycle (center of B8). **32**: carbocycle (border between E3 and E4). **33**: heteroaromatic (center of F2). **34**: heteroaromatic (top of H1). **35**: heteroaromatic (border between A1 and A2). **36**: fused heterocyle (lower-left corner of D2). **37**: fused heterocycle (center of D8). **38**: fused carbocycle (top of C1).

facts that 99.8% of **GDB** compounds are not found in the reference database **RDB** of known compounds and that 68% of the Kohonen map is not populated at all by **RDB** indeed suggest that there should be many new structures in **GDB**. The presence of many new ring systems in **GDB** as discussed above also suggests the possibility of new chemotypes.

With an average MW of 153 ± 7 Da, **GDB** compounds are rather small in comparison to the average drug molecule (MW 340 Da, 25 atoms) and should be generally best suited as lead molecules that can be optimized by adding substituents, as shown by the large proportion of leadlike compounds in **GDB** (see above). Nevertheless, a few drugs and bioactive natural products fall within the MW range covered by **GDB** and are found in the database (Figure 11). Examples include the drugs paracetamol (**39**, 10 atoms), pregabalin (**40**, 11 atoms), amantadine (**41**, 11 atoms), and fluorouracil (**42**, 9 atoms); the anesthetic desflurane (**43**, 10 atoms); neurotransmitters such as GABA (**44**, 7 atoms), glutamate (**45**, 10 atoms), and dopamine (**46**, 11 atoms); the vitamin niacin (**47**, 9 atoms); natural products such as the pheromones frontalin (**48**, 10 atoms) and brevicomin (**49**, 11 atoms); monoterpenes such as carvone (**50**, 11 atoms) or menthol (**51**, 11 atoms); and other fragrances such as benzaldehyde (**52**, 8 atoms) or cinnamaldehyde (**53**, 10 atoms).

A small-scale virtual screening experiment was carried out to test the potential of **GDB** for finding new drugs and leads. In virtual screening, one attempts to identify promising compounds of a certain bioactivity on the basis of computational methods such as similarity searching,[56] quantitative structure−activity relationships,[57,58] or machine-learning methods such as artificial neural networks.[59] We used a commercial package (Molinspiration miscreen toolkit) that analyzes compounds on the basis of Bayesian statistics.[60] This method assigns a bioactivity probability score to any structure on the basis of its substructural fragments and their occurrence in bioactive versus non-bioactive molecules, as derived from a reference set of active and inactive compounds. Virtual hits from this virtual screening comprise compounds that contain a high density of fragments occurring

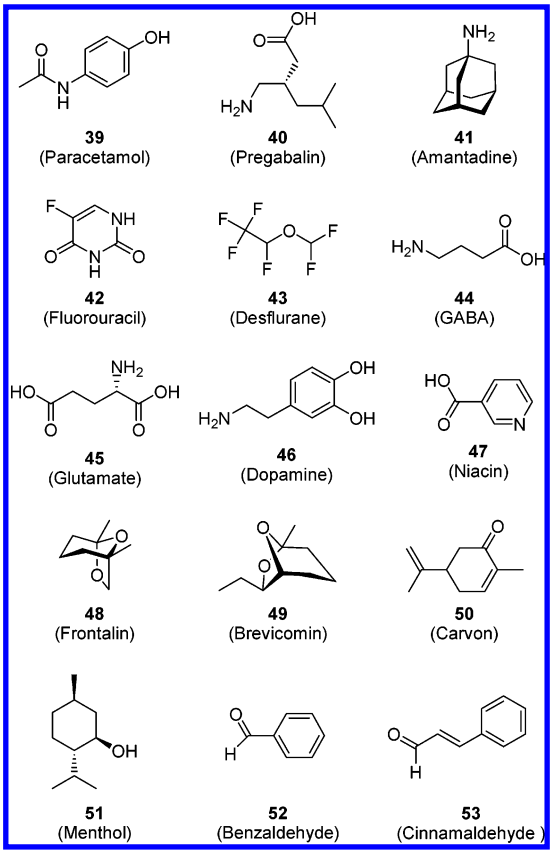

Figure 11. Examples of known drugs and bioactives present in **GDB**.

preferentially in known bioactives of the targeted class and might correspond to either drugs or lead compounds.

The virtual screening experiment was carried out for three important drug targets using the included predefined models by Molinspiration, and compounds exceeding the suggested threshold of 0.5 were considered as possibly active:[61] G-protein coupled receptors (GPCRs), kinases, and ion channels. Screening the entire database **GDB** (**RDB**) returned 3043 compounds (83 in **RDB**) with features of kinase inhibitors, 24 489 (85) with features of G-protein coupled receptors, and 19 696 (79) with features of ion-channel modulators, corresponding to a total of 42 804 virtual hits in **GDB** (239 in **RDB**). An analysis of the position of these virtual hits on a Kohonen map (Figure 9C) showed that the kinase inhibitor virtual hits were clustered in three well-defined areas corresponding to achiral heteroaromatic compounds and featured a large number of adenine analogs. G-protein coupled receptor inhibitor and ion-channel modulator virtual hits were mostly fused heterocycles occurring in leadlike regions of the map, whereby the structural diversity was consistent with the very heterogeneous structural types found in these types of compounds. Most importantly, 59.8% of the identified virtual hits occupied neurons of the Kohonen map not populated by **RDB** and represent previously unknown structures. Although a full presentation of these virtual hits lies beyond the scope of the present paper, a few examples are shown in Figure 12.

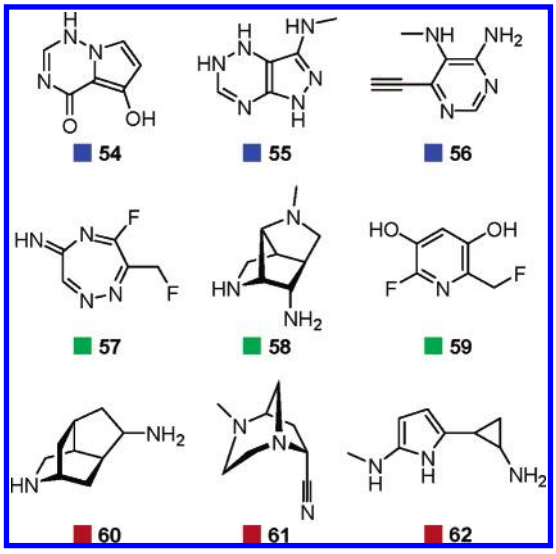

Figure 12. Examples of virtual hits from **GDB** found in neurons not occupied by **RDB** (the location in the Kohonen map of Figure 9C is indicated in parentheses). **54**: kinase inhibitor (bottom-center side of A1). **55**: kinase inhibitor (center-left side of C5). **56**: kinase inhibitor (lower-left corner of H2). **57**: ion-channel modulator (center-left side of C3). **58**: ion-channel modulator (lower-right corner of F8). **59**: ion-channel modulator (upper-right corner of D6). **60**: GPCR ligand (center-right side of D8). **61**: GPCR ligand (center-left side of F3). **62**: GPCR ligand (border between A4 and B4).

CONCLUSION

In summary, we have shown the construction and properties of **GDB**, a database of all small molecules up to 11 atoms of C, N, O, and F possible under consideration of simple valency, chemical stability, and synthetic feasibility rules. Molecules were generated from mathematical graphs, via the selection of graphs with moderate ring strain, the introduction of unsaturations and element types, the selection of functional groups by chemical stability criteria, and the determination of tautomeric forms and absolute and relative stereochemistry. All compounds in **GDB** follow Lipinski's "Rule of 5" for bioavailability because of their small size. Furthermore, half of **GDB** compounds also have leadlike properties according to Congreve's "Rule of 3". A Kohonen map based on autocorrelation vectors organized **GDB** by compound classes and revealed that leadlike molecules are most abundant in regions of chiral fused carbocycles and fused heterocycles.

One of the most interesting possible uses of **GDB** concerns its use for drug and lead discovery. **GDB** contains mostly small molecules (MW ~157.3 Da) in the range of typical lead compounds yet contains some known drugs and bioactives. The potential of **GDB** for drug and lead discovery was tested by virtual screening for kinase inhibitors, GPCR ligands, and ion-channel modulators using a virtual screening based on Bayesian statistics, leading to 42 804 possible bioactives. Many other virtual screening approaches are of course conceivable to mine **GDB** for new drugs, some of which will be the subject of future reports.

The exhaustive enumeration approach chosen here to explore chemical space provides a rich source of information, but the approach might not be utilizable for molecules larger than 11 atoms. Indeed, the size of **GDB** up to 11 atoms is already comparable to the contents of the current entire Chemical Abstracts index. The use of more computer power and memory should allow reaching 12 or perhaps 13 atoms, extending the database by 2−4 orders of magnitude. Since

the number of molecules in **GDB** increases exponentially with the square of the number of atoms, one can extrapolate that there should be approximately 10^{27} different molecules at 25 atoms, the average size of druglike molecules, a number comparable to an independent estimate based on a combination of known fragments.[14] A database of this size is currently out of reach for an exhaustive enumeration. Other approaches such as the molecular breeding of **GDB** molecules by a genetic algorithm offer a more realistic even if not an exhaustive option to explore the chemical universe for these larger molecules and will be reported in due course.

METHODS

General Remarks. All applications implemented in this work were entirely written in Java (J2SE v5.0) using the JChem v3.1 and Marvin v4.0 API provided by ChemAxon.[62,63]

Structure Generation. Structure generation including tautomer and stereoisomer generation was achieved using a collection of in-house-developed software applications. The generation process was parallelized 80-fold on two AMD Opteron 252 2.6 GHz CPUs, yielding the obtained results in approximately 20 h.

MM2 Force Field. All force-field calculations were carried out with our own implementation of the MM2 force field. The following functional form of the force field was used:

$$\begin{aligned} E_{\text{Steric}} = & \sum_{\text{bonds}} k_b(l_i - l_{0,i})^2[1 + k_b'(l_i - l_{0,i}) + k_b''(l_i - l_{0,i})^2] + \\ & \sum_{\text{angles}} k_\theta(\theta_i - \theta_{0,i})^2[1 + k_\theta'(\theta_i - \theta_{0,i})^4] + \\ & \sum_{\text{angles}} k_{b,\theta}(\theta_i - \theta_{0,i})^2[(l_a - l_{0,a}) + (l_b - l_{0,b})] + \\ & \sum_{\text{torsions}} \frac{V_1}{2}(1 + \cos\omega) + \frac{V_2}{2}(1 - \cos 2\omega) + \\ & \frac{V_3}{2}(1 + \cos 3\omega) + \\ & \sum_{i=1}^{N}\sum_{j=i+1}^{N} \epsilon_{ij}\left[A \exp\left(\frac{-Br_{ij}}{\sum r_{ij}^*}\right) - C\left(\frac{r_{ij}}{\sum r_{ij}^*}\right)^6\right] \end{aligned} \quad (2)$$

This functional form is identical to the original MM2 force field except for the quartic term of the bond stretching that was added later by Allinger in the MM3 force field.[29,64] This term is included to avoid a lengthening of the bonds caused by the fact that the cubic term has a maximum far from the equilibrium state. The parameter set was taken from the MM2 force field as published.[29]

Statistical Analysis. The analysis of univariate measures and principle component analysis were performed using SAS v8.

Autocorrelation vectors were calculated using the in-house-implemented application ACD, and Kohonen maps were produced using the in-house-developed application SOM v1.0, following the well-known algorithm of Kohonen for training.[54]

ACKNOWLEDGMENT

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, ChemAxon for the use of the JChem and Marvin API, and Molinspiration for providing access to their virtual screening toolkit miscreen.

Supporting Information Available: Distribution of the highest atomic contribution of ring systems involved in the database construction process (Figure S1); list of filters applied during the structure generation process (Table S1); and descriptor value distribution profiles for GDB, RDB, and DMU (Figure S2). This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bleicher, K. H.; Böhm, H.-J.; Müller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.
(2) Olah, M. M.; Bologa, C. G.; Oprea, T. I. Strategies for Compound Selection. *Curr. Drug Discovery Technol.* **2004**, *1*, 211−220.
(3) Koch, M. A.; Waldmann, H. Protein Structure Similarity Clustering and Natural Product Structure as Guiding Principles in Drug Discovery. *Drug Discovery Today* **2005**, *10*, 471−83.
(4) Fergus, S.; Bender, A.; Spring D. R. Assessment of Structural Diversity in Combinatorial Synthesis. *Curr. Opin. Chem. Biol.* **2005**, *9*, 304−309.
(5) Reayi, A.; Arya, P. Natural Product-Like Chemical Space: Search for Chemical Dissectors of Macromolecular Interactions. *Curr. Opin. Chem. Biol.* **2005**, *9*, 240−247.
(6) Tan, D. S. Diversity-Oriented Synthesis: Exploring the Intersections between Chemistry and Biology. *Nat. Chem. Biol.* **2005**, *1*, 74−84.
(7) Noren-Müller, A.; Reis-Corrêa, I., Jr.; Prinz, H.; Rosenbaum, C.; Saxena, K.; Schwalbe, H. J.; Vestweber, D.; Cagna, G.; Schunk, S.; Schwarz, O.; Schiewe, H.; Waldmann, H. From the Cover: Discovery of Protein Phosphatase Inhibitor Classes by Biology-Oriented Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10606−10611.
(8) Chin, Y. W.; Balunas, M. J.; Chai, H. B.; Kinghorn, A. D. Drug Discovery from Natural Sources. *AAPS J.* **2006**, *8*, E239−E253.
(9) Baker, D. D.; Alvi, K. A. Small-Molecule Natural Products: New Structures, New Activities. *Curr. Opin. Biotechnol.* **2004**, *15*, 576−583.
(10) Haefner, B. Drugs from the Deep: Marine Natural Products as Drug Candidates. *Drug Discovery Today* **2003**, *8*, 536−544.
(11) Bohacek, R. S.; Martin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modelling Approach. *Med. Res. Rev.* **1996**, *16*, 3−50.
(12) Gorse, A. D. Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **2006**, *6*, 3−18.
(13) Feher, M.; Schmidt J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(14) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-Like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.
(15) Lederberg, W. Topological Mapping of Organic Molecules. *Proc. Natl. Acad. Sci. U.S.A.* **1965**, *53*, 134−139.
(16) Buchanan, B. G.; Feigenbaum, E. A. DENDRAL and Meta-DENDRAL − Their Applications Dimension. *Artif. Intell.* **1978**, *11*, 5−24.
(17) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inferenc. 17. Approach to Computer-Assisted Elucidation of Molecular-Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755−5762.
(18) Sasaki, S.; Kudo, Y.; Ochiai, S.; Abe, H. Automated Chemical Structure Analysis of Organic Compounds - Attempt to Structure Determination by Use of NMR. *Mikrochim. Acta* **1971**, 726−742.
(19) Shelley, C. A.; Munk, M. E. CASE, a Computer-Model of the Structure Elucidation Process. *Anal. Chim. Acta Comp. Tech. Opt.* **1981**, *5*, 507−516.
(20) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. MOLGEN(+), a Generator of Connectivity Isomers and Stereoisomers for Molecularstructure Elucidation. *Anal. Chim. Acta* **1995**, *314*, 141−147.
(21) A preliminary report of this work has been published. Fink, T.; Bruggesser, H.; Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem., Int.*

Ed. **2005**, *44*, 1504−1508. The database is available via the Internet at http://www.dcb.unibe.ch/groups/reymond/ (accessed Dec 12, 2006).
(22) Balaban, A. T. *Chemical Application of Graph Theory*, 1st ed.; Academic Press: London, 1976.
(23) McKay, B. D. Practical Graph Isomorphism. *Congr. Numerantium* **1981**, *30*, 45−87.
(24) Moman, E.; Nicoletti, D.; Mourino, A. Strained Polycycles by H^5C^{5x}-C-5 Free-Radical Cascades. *Org. Lett.* **2006**, *8*, 1249−1251.
(25) Rücker, C.; Prinzbach, H. cis-tris-σ Homobenzenes from cis-Benzenetrioxide. *Tetrahedron Lett.* **1983**, *24*, 4099−4102.
(26) Kuratowski, K. Sur le Problème des Courbes Gauches en Topologie. *Fund. Math.* **1930**, *15*, 271−283.
(27) Klunder, A. J. H.; Zwanenburg, B. Chemistry of Strained Polycyclic Compounds. III. Synthesis of Cubane and Homocubane Alcohols. *Tetrahedron* **1972**, *28*, 4131−4138.
(28) Katz, T. J.; Acton, N. Synthesis of Prismane. *J. Am. Chem. Soc.* **1973**, *95*, 2738−2739.
(29) Allinger, N. L. Conformational Analysis. 130. MM2. A Hydrocarbon Force Field Utilizing V_1 and V_2 Torsional Terms. *J. Am. Chem. Soc.* **1977**, *99*, 8127−8134.
(30) Bohanec, S.; Perdih, M. Symmetry of Chemical Structures: A Novel Method of Graph Automorphism Group Determination. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 119−726.
(31) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(32) Ma, S. Some Typical Advances in the Synthetic Applications of Allenes. *Chem. Rev.* **2005**, *105*, 2829−2872.
(33) *Ashgate Drugs*, version 2.1; CambridgeSoft Corporation: Cambridge, MA, 2005.
(34) Frost, A. A.; Musulin, B. A Mnemonic Device for Molecular Orbital Energies. *J. Chem. Phys.* **1953**, *21*, 572−573.
(35) This method avoids counting highly reduced heteroaromatic rings such as 3,4,5,6-tetraza-1-cyclohexene as an aromatic compound as it is done in various commercial toolkits (e.g. Daylight). However, examples where the molecule does not fulfill the planarity criterion of Hueckel still pass as aromatic, like the well-known nonaromatic cyclodeca-1,3,5,7,9-pentaene.
(36) Nourse, J. G.; Carhart, R. E.; Smith, D. H.; Djerassi, C. Exhaustive Generation of Stereoisomers for Structure Elucidation. *J. Am. Chem. Soc.* **1979**, *101*, 1216−1223.
(37) An organic molecule is considered as a molecule containing at least one carbon atom and only elements of the organic subset C, N, O, F, Cl, Br, I, S, P, and Si.
(38) National Center for Biotechnology Information. The PubChem Project. http://pubchem.ncbi.nlm.nih.gov/ (accessed Aug 9, 2006).
(39) *ChemACX Ultra*, version 8.0; CambridgeSoft Corporation: Cambridge, MA, 2005.
(40) *ChemSCX*, version 8.0; CambridgeSoft Corporation: Cambridge, MA, 2005.
(41) National Cancer Institute. NCI Open Database. http://cactus.nci.nih.gov/ncidb2/download.html. (accessed Aug 9, 2006).
(42) *The Merck Index*, version 13.4; CambridgeSoft Corporation: Cambridge, MA, 2005.
(43) Weitemeyer, C.; De Meijere, A. 3,7,10-Trioxapentacyclo[3.3.3.$0^{2,4}$.$0^{6,8}$.-$0^{9,11}$)undecane (3,7,10-Trioxatris(homobarrelene)). *Angew. Chem., Int. Ed. Engl.* **1976**, *15*, 686−687.
(44) Crippen, G. M.; Wildman, S. A. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(45) 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.
(46) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(47) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A Rule of Three for Fragment-Based Lead Discovery? *Drug Discovery Today* **2003**, *8*, 876−877.
(48) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures. Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.
(49) Zakarya, D.; Tiyal, F.; Chastrette. M. Use of the Multifunctional Autocorrelation Method to Estimate Molar Volumes of Alkanes and Oxygenated Compounds - Comparison between Components of Autocorrelation Vectors and Topological Indexes. *J. Phys. Org. Chem.* **1993**, *6*, 574−582.
(50) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205−1213.
(51) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(52) Miller, K. J.; Savchik, J. A. New Empirical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(53) *Topology Analysis Calculator Plugin*, version 4.0; ChemAxon: Budapest, Hungary, 2005.
(54) Kohonen, T. *Self-Organizing Maps*, 3rd ext. ed.; Springer: Berlin, 2001.
(55) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999.
(56) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(57) Liu, S.-S.; Yin, C.-S.; Li, Z.-L.; Cai, S.-X. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321−329.
(58) Stiefl, N.; Baumann, K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure−Activity Relationship Technique. *J. Med. Chem.* **2003**, *46*, 1390−1407.
(59) Gasteiger, J.; Teckentrup, A.; Terfloth, L.; Spycher, S. Neural Networks as Data Mining Tools in Drug Design. *J. Phys. Org. Chem.* **2003**, *16*, 232−245.
(60) *miscreen - Molinspiration Cheminformatics Virtual Screening Toolkit*, version 2005.03; Molinspiration Cheminformatics: Slovensky Grob, Slovak Republic, 2005.
(61) The training and test sets that were used to build these models are the property of Molinspiration and were not accessible for the authors.
(62) *JChem*, version 3.1; ChemAxon: Budapest, Hungary, 2005.
(63) *Marvin*, version 4.0; ChemAxon: Budapest, Hungary, 2005.
(64) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. Molecular Mechanics − The MM3 Force-Field for Hydrocarbons. *J. Am. Chem. Soc.* **1989**, *111*, 8551−8556.

CI600423U