Topological Analysis of Molecular Scaffolds: 2. Evaluation of Chemical Databases

Michael J. Wester[†], Sara Pollock[†], Evangelos A. Coutsias[†], Tharun Kumar Allu[‡], Sorel Muresan[♮], Tudor I. Oprea[‡]

† Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA; ‡ Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA; ♮ AstraZeneca, Sweden

Abstract

We have systematically enumerated graph representations of scaffold topologies for up to 8-ring molecules and 4-valence atoms, thus providing coverage of the lower portion of the chemical space of small molecules (Pollock et al.[1]). Here, we examine scaffold topology distributions for several databases: ChemNavigator and PubChem for commercially available chemicals, the Dictionary of Natural Products, a set of 2,742 launched drugs, and WOMBAT, a database of medicinal chemistry compounds. We also examined a virtual database of exhaustively enumerated small organic molecules, GDB,[2] and contrast the scaffold distribution from these collections to the complete coverage of up to 8-ring molecules. For reasons related, perhaps, to synthetic accessibility and complexity, scaffolds exhibiting 6 rings or more are poorly represented. Among all collections examined, PubChem has the greatest topological diversity, whereas GDB is the most limited topologically. More than 50% of all entries (13,000,000+ actual and 13,000,000+ virtual compounds) exhibit only 8 distinct topologies, one of which is the non-scaffold topology that represents all treelike structures. However, most of the topologies are represented by a single or very small number of examples. Within topologies, we found that 3-way scaffold connections (3-nodes) are much more frequent compared to 4-way (4-node) connections. Fused rings have a slightly higher frequency in biologically oriented databases. Scaffold topologies can be the first step toward an efficient classification scheme of the molecules found in chemical databases.

1 Introduction

Drugs are the cornerstone of allopathic medicine, and the vast majority have emerged from the private sector (pharmaceutical industry). Drug discovery is almost uniquely supported by the ability of the inventors to obtain patent rights regarding the usability and/or chemical structures of drugs. Pharmaceutical R&D, and more recently the National Institutes of Health (NIH) and other governmental agencies, have become more and more interested in tools and means to query the therapeutically relevant chemical space of small molecules (CSSM),[3–5] also known as ‘drug-like’

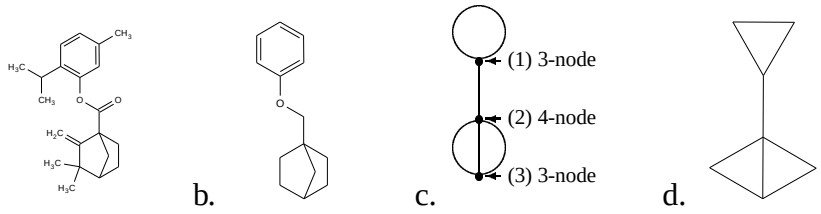

Figure 1 a. (5-methyl-2-propan-2-yl-phenyl) 3,3-dimethyl-2-methylidene-bicyclo[2.2.1]heptane-1-carboxylate [SMILES: CC(C)c1ccc(C)cc1OC(=O)C2(CCC3C2)C(=C)C3(C)C]. b. The scaffold corresponding to this molecule [C1CC2CCC1(C2)COc3ccccc3]. c. The topology corresponding to this scaffold (nodes are numbered as shown). d. A minimal representive of this topology [C1CC1C23CC2C3].

chemical space.[6] To this end, the question of how vast this chemical space is has been addressed in several ways—most of them related to *in silico* technologies, such as virtual chemical library enumeration starting from known lists of reagents. Such methods, however, explore only the limited space covered by (a) known chemical reactions and (b) available/known chemical reagents. The question of how large the drug-like chemical space is was recently extended with the launch of the NIH Roadmap molecular libraries initiative.[7] As the NIH is embarking in the selection and biological screening of 500,000 chemicals in search of biomolecular probes, the issue of which chemicals to acquire (from over 10,000,000 commercial structures) is not a trivial one.

2 Methods

The details of the mathematical methods we used are described in Pollock et al.[1] Here, we will summarize the definitions and algorithms that were needed for the analyses presented here.

2.1 Scaffold Topologies

A scaffold is the common portion of a series of related compounds from which it is possible to hang active groups or spacers to form more complex compounds (a well-known example of a scaffold is the peptide backbone). Here, we provide an operational definition:

Definition 1 We consider a scaffold to be a chemical graph composed solely of rings and optional linking linear structures. All branches of a scaffold terminate in a ring.

Traditionally, scaffolds can also admit atoms double-bonded to ring atoms,[8] but we do not include these special atoms in our description of topologies. Figure 1a,b shows a sample molecule and its corresponding scaffold.

To simplify matters, in the discussion that follows, we will disregard the distinction between single, double and triple bonds as well as between different atom types (e.g., C, N, O, etc.); note that by the nature of scaffolds, hydrogen atoms will be omitted from the molecular descriptions. We will use the graph theory terminology of nodes and edges to indicate atoms and bonds, respectively.

A k-node is defined to be a node of degree k, where the degree indicates the number of edge segments incident to the node (see Figure 1c). The valence of the atom represented by the node determines the maximum value of k, so, for example, carbon atoms in a dehydrogenated molecule exist as 1, 2, 3 or 4-nodes. An ℓ-edge consists of ℓ edges connecting two distinct nodes. A loop is an edge that connects a node to itself. In Figure 1c, node 1 has a loop, nodes 1 and 2 are connected by a 1-edge, and nodes 2 and 3 are connected by a 3-edge.

The topology of a molecule's scaffold is constructed from a molecule by recursively removing all of its 1-nodes (all branches that do not ultimately terminate in a ring on both ends), and by eliminating all of its 2-nodes (which simply divide an edge into two segments). The remaining nodes, which will be of degree three or greater, generate branching, initiating rings or ring connectors, and so establish the scaffold's topology. Topologies may contain multiple edges and loops, both features that are not found in molecular graphs. Nodes of degree five or more are rare in the databases that we examined (see Section 3), so we will only consider topologies consisting of 3-nodes and 4-nodes,[9] which correspond to carbon-based molecules.

Definition 2 A scaffold topology is constructed from a scaffold by
1. disregarding differences in atom type so nodes only differ by their connectivity,
2. treating multiple bonds as single edges, and
3. eliminating all 2-nodes from the resulting graph (except in the situation of a single ring in which case one 2-node is retained), 1-nodes having already been removed to produce the scaffold.

Let r and N_k count the number of independent rings and k-nodes, respectively, then for topologies,[1]

$$r = N_4 + \frac{N_3}{2} + 1 \,. \tag{1}$$

For a fixed value of r, N_3 and N_4 will thus take on the integer values

$$\begin{array}{rcc|c|c|c|c|c|c} N_3 & = & 2(r-1) & 2(r-2) & 2(r-3) & \cdots & 2(r-i-1) & \cdots & 0 \\ N_4 & = & 0 & 1 & 2 & \cdots & i & \cdots & r-1 \end{array} \,,$$

and hence, for a topology,

$$r-1 \le n \le 2(r-1) \qquad \text{and} \qquad 2(r-1) \le e \le 3(r-1) \,.$$

2.2 Comparing Topologies

Several schemes for uniquely characterizing molecular graphs have appeared (Trinajstíc et al.[10] describes a number of methods—see also 11–13). This has been a difficult task as complex graphs can have sophisticated symmetries that defy easy classification (see Berger et al.[14] for some remarkable counterexamples in ring perception).

We represent both molecular graphs and their topologies by adjacency matrices, A. Since we are only interested in the connectivity of atoms in molecules and scaffolds, and not whether a

$r =$	1	2	3	4	5	6	7	8
Total:	1	3	12	73	590	6454	88129	1452427

N_4 \ N_3	0	2	4	6	8	10	12	14
7	359							
6	97	13239						
5	28	2242	105188					
4	10	430	12905	326761				
3	4	88	1655	28301	483124			
2	2	22	228	2457	28649	365994		
1	1	5	30	193	1496	13343	136666	
0	1	2	5	17	71	388	2592	21096

Table 1 The total number of distinct scaffold topologies for 1 through 8 rings (top), and categorized by the number of 3-nodes, N_3, and 4-nodes, N_4 (bottom). The diagonal colors indicate the number of rings (r). Note that the (0, 0) topology is a loop with a 2-node.

r	1	2		3			4	
N_4	0	0	1	0	1	2	0	3
N_3	0	2	0	4	2	0	6	0
scaffolds	1	2–3	4	5–9	10–14	15–16	17–33	86–89

Table 2 Topology descriptors for the scaffolds in Figure 2.

bond is single, double or triple, all the molecular adjacency matrices will only have entries of zero or one. Topology adjacency matrices, however, can have nodes that are multiply connected with other nodes or with themselves (loops). From A, we compute the return-index R as discussed in the companion paper.[1] We note that the return-index is related to the characteristic polynomial and eigenvalues of a graph.[15–17]

We have exhaustively verified that after sorting with respect to the number of rings and the number of 3- or 4-nodes, the return-index is sufficient to distinguish topologies with up through 8 rings for molecules with atoms of valence up to 4.[1] For $r = 12$, we know of examples of topologies that have the same return-indices, yet are distinct.[1] The return-index is not sufficient to distinguish between graphs containing nodes of degree greater than four. Scaffolds with nodes of degree five or more are, however, rare as noted earlier.

Table 1 shows the results of enumerating all possible topologies up through 8 rings. In Figure 2a, the minimal scaffolds for all topologies with 1–3 rings are presented as well as those for the 3-node only and 4-node only 4-ring topologies. 52 mixed 3/4-node 4-ring topologies are not shown. See Table 2 for further identifications. Figure 2b exhibits examples of all the topologies shown in Figure 2a, except for number 17, which was not present in any of the databases examined.

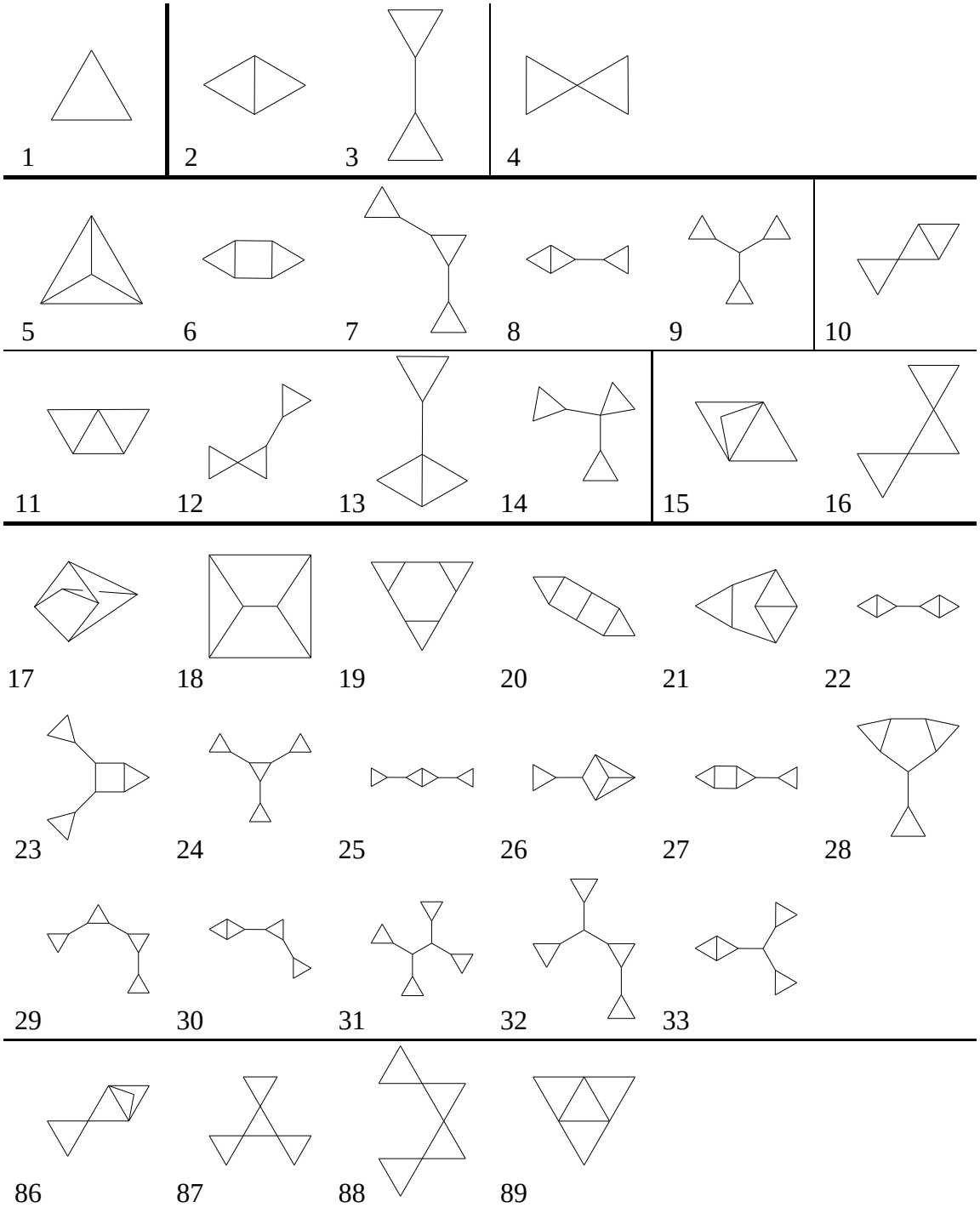

Figure 2 a. Minimal scaffolds for all 1–3-ring topologies and all 4-ring topologies possessing only 3-nodes or only 4-nodes. See Table 2 for further identification.

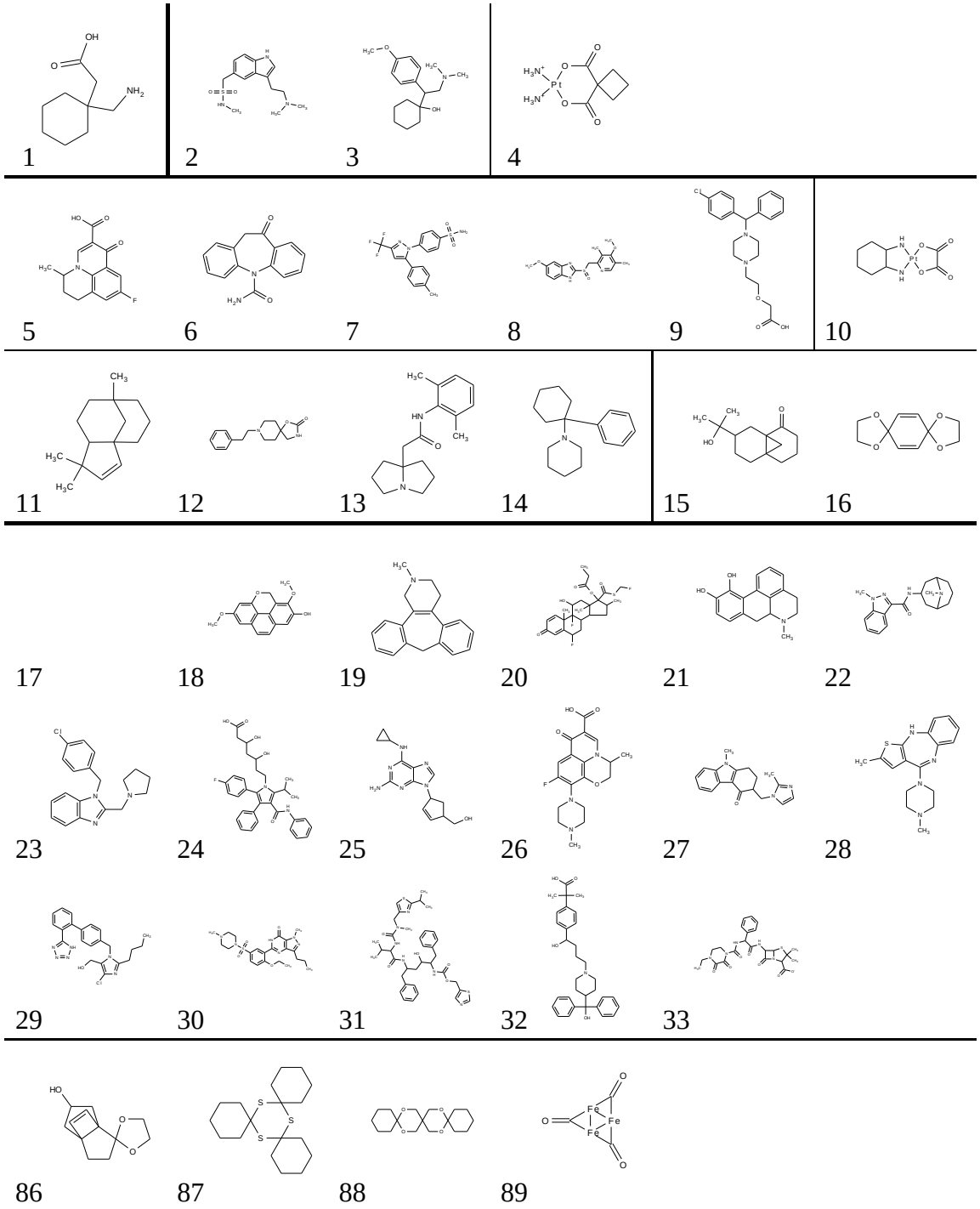

Figure 2 b. Examples[18] from the databases examined of molecules that exhibit each 1–3-ring topology and each 4-ring topology possessing only 3-nodes or 4-nodes, corresponding to the topologies in Figure 2. Note that none of the databases examined possessed an example of topology number 17. See Table 2 for further identification.

2.3 Spiro Atoms

A spiro atom is the unique common member of two or more otherwise disjoint ring systems.[19] As the topology fully describes the ring systems of a scaffold, the number of spiro atoms is an invariant for all scaffolds corresponding to a given topology. A scaffold's topology is in general a smaller graph than the scaffold itself, and so is a convenient tool for the analysis of spiro atoms. A spiro atom by its definition requires a node of degree at least four. We implement an exhaustive breadth-first search technique to determine if any node in the topology corresponds to a spiro atom. In a search of chemical libraries, we may encounter atoms of degrees greater than four (e.g., sulfur), and so we can apply the concept of spiro degree to count the number of otherwise disjoint ring-systems of which an atom is the unique common member. If the degree of a spiro is not specified, it is assumed to be two. In Figure 2a, the only topologies that have spiro atoms are 4, 10, 12 and 86 with one, 16 with two, and 87 and 88 with three.

2.4 Database Measures

Let N_{ik} count the number of k-nodes in the i^{th} molecule of a chemical database containing M molecules from which molecules lacking a scaffold (i.e., possessing no rings) have been excluded. Let $N_{ik}^{(s)}$ count the number of k-nodes in the scaffold corresponding to the i^{th} molecule. The average fraction of atoms per molecule that make up the scaffold is then

$$\frac{\sum_{i=1}^{M}\sum_{k\geq 2}N_{ik}^{(s)}}{\sum_{i=1}^{M}\sum_{k\geq 1}N_{ik}},$$

where the maximum value of k in the databases we examined was 6. The average fraction of branch points ($\geq$3-nodes) per scaffold is

$$\frac{\sum_{i=1,r\geq 2}^{M}\sum_{k\geq 3}N_{ik}^{(s)}}{\sum_{i=1,r\geq 2}^{M}\sum_{k\geq 2}N_{ik}^{(s)}},$$

which excludes single-ring ($r=1$) structures. The average scaffold connectivity (node degree) is

$$\frac{\sum_{i=1}^{M}\sum_{k\geq 2}kN_{ik}^{(s)}}{\sum_{i=1}^{M}\sum_{k\geq 2}N_{ik}^{(s)}}.$$

The average number of independent rings per scaffold is

$$\frac{\sum_{i=1}^{M}\left(\frac{1}{2}\left[\sum_{k\geq 3}(k-2)N_{ik}^{(s)}\right]+1\right)}{M}=1+\frac{\sum_{i=1}^{M}\sum_{k\geq 3}(k-2)N_{ik}^{(s)}}{2M}.$$

This last quantity is derived from a generalization of Equation 1.

Database	Version	Unique SMILES	Distinct topologies	% top. / merged top.	% top. / SMILES
ChemNavigator	October 2006	14,041,970	3,880	16.346	0.0242
DNP	April 2006	132,434	3,199	13.477	2.4155
Drugs	2006	2,742	155	0.653	5.6528
PubChem	November 7, 2006	11,595,690	22,612	95.261	0.1950
PC actives	November 7, 2006	38,881	1,052	4.432	2.7057
WOMBAT	December 2006	149,451	1,333	5.616	0.8919
merged		25,029,904	23,737	100.000	0.0948
GDB	2005	26,434,571	76	0.320	0.0003

Table 3 Databases examined, including a merged one constructed from all the others, their sizes, the number of distinct topologies discovered, the percentage this makes with respect to the total number of distinct topologies in the merged database, and the percentage ratio of topologies to SMILES (molecules). GDB, a generated database, is analyzed separately.

3 Analysis of Some Existing Databases

We computed scaffold topologies for the molecules found in several databases, as follows: ChemNavigator,[20] which collects commercially available chemicals; the Dictionary of Natural Products (DNP);[21] an in-house compilation of 2,742 launched drugs (Drugs); PubChem,[22] a public repository of small molecules which have been characterized for biological activity; PubChem "actives", which is the PubChem subset labeled as "active"; and WOMBAT,[23] a collection of small molecules with known biological activity from medicinal chemistry literature (see Table 3). For each database, we processed SMILES[24,25] for all the molecules, removed salts, hydration information and counter-ions, then eliminated non-unique entries. We converted each SMILES to an adjacency matrix using OEChem,[26] stripped each molecule down to its simplified scaffold (see Section 2), then extracted the distinct topologies and cataloged their frequencies. Furthermore, we carried out the same procedure on the non-redundant union of all databases,,[27] which was used to compare the topological coverage of the individual databases. We note that 10,153 (42.8%) of the distinct topologies found in the merged database had a single representative and 17,634 (74.3%) had 5 or less representatives. We also examined the Generated Database of Chemical Space of Small Molecules (GDB),[28] in which all organic molecules with 11 or less main atoms and molecular weight less than 160 Daltons have been algorithmically generated, then filtered down for synthetic feasibility and stability.[2]

In the last column of Table 3, the percentage ratio of topologies to SMILES was computed. This figure provides an indication of the databases' topological diversity. The smaller, more biologically oriented databases have the greatest ratios (especially Drugs, PC actives and DNP), while GDB, with only 76 unique topologies but over 26,000,000 SMILES, has a very low topology to SMILES ratio and therefore, low topological diversity.

Database	Fraction no rings	Maximum rings	$>$ 4-nodes population
ChemNavigator	0.002	62	95
DNP	0.086	32	61
Drugs	0.065	18	0
PubChem	0.025	165	6488
PC actives	0.039	23	198
WOMBAT	0.016	34	0
merged	0.012	165	6593
GDB	0.154	6	0

Table 4 For each database, the fraction of molecules that did not contain rings, the maximum number of rings found in a single compound, and the population of molecules that possessed at least one 5- or 6-node.

As can be seen in Table 4, nearly all the molecules contain rings and can be stripped down into scaffolds (these findings are similar to those of Lewell et al.[29] and Koch et al.[30]). Note, however, that 8.6% of the DNP structures, 6.5% of the Drugs and 3.9% of the PubChem actives, all biologically oriented, do not contain rings. 15.4% of the generated structures in GDB also lack rings. Note also that the larger databases of known chemicals contain, in general, larger structures. The most rings found in a single scaffold topology is from a PubChem compound with $r = 165$ ($N_6 = 8, N_5 = 88, N_3 = 32$). The next largest, also from PubChem, has 107 rings ($N_3 = 212$). In general, the largest examples in each database possess no 4-nodes, only 3-nodes and possibly 5- or 6-nodes.

Scaffold topologies containing a 5- or 6-node are rare; only 0.5% of the entries in the PubChem actives database (the most extreme case) contain nodes of such high degree. PubChem with 0.06% had the next greatest percentage of molecules possessing a scaffold with a 5- or 6-node, while Drugs, WOMBAT and GDB contain no such structures at all. We found no scaffolds that had nodes with degrees > 6. Therefore, we can safely ignore such higher degree nodes and concentrate on topologies that contain nodes of at most degree 4. A major reason why there are so few nodes of degree > 4 is that those atoms with high valence (e.g., P and S) are typically not ring members, so are commonly stripped off when scaffolds are created.

A variety of chemical, geometrical and topological criteria have been used to describe molecules and to map out chemical space. Here, we concentrate on measures based on topological properties to characterize the databases of interest, as illustrated in Table 5. One such measure is the average fraction of atoms per molecule that make up the scaffold (see the first data column). In the biologically oriented databases (DNP, Drugs, PubChem actives and WOMBAT), this fraction averages 0.610–0.714, while in the other known chemical databases, that average is higher, ranging 0.717–0.745. Thus, biologically oriented molecules tend to exhibit a higher fraction of the molecule that is represented by chemical substituents to the scaffold, rather than as part of it. This is likely to increase chemical and pharmacophore diversity at a scaffold, which is a traditional way of ex-

Database	Fraction scaffold	Fraction $\geq$ 3-nodes	Node degree	Number of rings
ChemNavigator	0.745	0.211	2.208	3.278
DNP	0.610	0.283	2.269	3.778
Drugs	0.636	0.236	2.202	2.854
PubChem	0.717	0.223	2.211	3.148
PC actives	0.714	0.249	2.232	3.311
WOMBAT	0.671	0.226	2.218	3.481
merged	0.733	0.217	2.210	3.235
GDB	0.605	0.307	2.049	1.653

Table 5 Basic database measures: average fraction of atoms per molecule that make up the scaffold, average fraction of branch points ($\geq$ 3-nodes) per scaffold, average scaffold connectivity (node degree), average number of independent rings per scaffold. See Methods for computational details.

ploring biological activity around a given scaffold. The lowest fraction of scaffold atoms (0.605) is in GDB, which indicates that these molecules contain a considerable fraction of non-scaffold structure. This is not surprising, since the goal of GDB is to exhaustively map chemical space and is, in a way, equivalent to the manner in which patents enumerate substituents for chemical completeness, a situation that rarely leads to synthesized compounds.

Others[29] have computed the scaffold molecular weight fraction, a related measure. The atoms that are stripped to produce the scaffold include all hydrogens; in general, the scaffold tends to retain a majority of the molecular mass. In a collection of approximately 10,000 preclinical and clinical phase candidates, including some marketed drugs, 56% of the molecular weight of the compounds[29] was present in the scaffolds (as we define them here).

Another topological measure is the fraction of scaffold atoms that are essential for defining the scaffold topology of multi-ring systems. This is the fraction of branching ($\geq$ 3)-nodes found within the scaffold. The second data column in the table lists the average fractions of scaffold atoms that define the scaffold topologies. These numbers tend to be around 0.217 for known chemicals, with somewhat higher values for the biologically oriented databases and GDB. GDB and DNP have by far the greatest branching structure within their scaffolds.

Bone and Villar[31] looked at the average connectivity (average node degree) of molecular structures as an indicator of diversity. The average node degree taken over all scaffolds is given in the third data column of Table 5. This measure is quite similar among databases of known chemicals, averaging around 2.21, with DNP having a marginally higher value. GDB scaffolds, averaging 2.049, are, on average, less connected, hence less diverse topologically.

Another such measure is the average number of independent rings per scaffold. Three-ring scaffolds are the most common in the version of DNP that Koch et al. examined, with the counts of two and four ringed-systems lying within one standard deviation.[30] Natural products have the highest average number of rings and marketed drugs the least, with natural product derivatives and combinatorially synthesized chemicals inbetween.[32] Our results show similar trends, but much

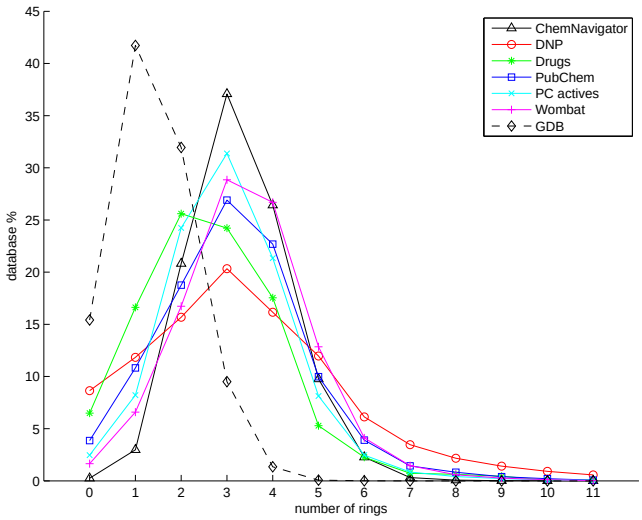

Figure 3 The population percentages in the indicated databases with respect to the total database population for the number of rings per scaffold.

less pronounced, since we examine larger collections (except for the drugs). GDB has a much lower average ring count than the other databases, which is merely indicative of the artificial limits imposed by enumeration (160 daltons, 11 atoms).

Figure 3 shows how the database population percentages correspond to the number of rings in more detail. All databases of known chemicals show fairly similar trends, peaking at three rings (except for Drugs which has 1.4% more two-ring than three-ring structures), with the majority of each database consisting of 2–4 ring molecules. DNP has the broadest peak, indicating that the number of rings in natural products are more evenly spread out than in other classes of chemicals. GDB has a different character than the other databases, peaking at one ring and then dropping sharply, nearly reaching zero at five rings. This is, of cours, consistent with the limitations imposed on the database by the upper bound of 11 heavy atoms.

In Figure 4, the populations of scaffold topologies in the ChemNavigator database are displayed as a function of N_3, N_4 and r. (All of the individual databases showed similar trends.) The populations drop sharply as the number of rings increases. In addition, in this three-dimensional representation, we can see that the currently explored portion of chemical space is strongly biased against 4-node scaffold topologies.

The above trends are again evident when the numbers of topologies in the various databases are compared with the theoretical maxima that we have computed in Table 1. In Table 6, the fractions of the topologies present versus the theoretical possibilities are tabulated as a function of the number of rings, while in Table 7, the fractions for $r = 1$–6, categorized by N_3 and N_4, are displayed. Note that a blank entry means no topologies of the indicated class were present in the specified database, while 0.000 means that there were some examples present, but the number is zero to three decimal places. The fractions for $r = 1$ and 2 were 1.0 for all databases and were generally 1.0 for $r = 3$, the exceptions being Drugs and WOMBAT, both smaller databases.

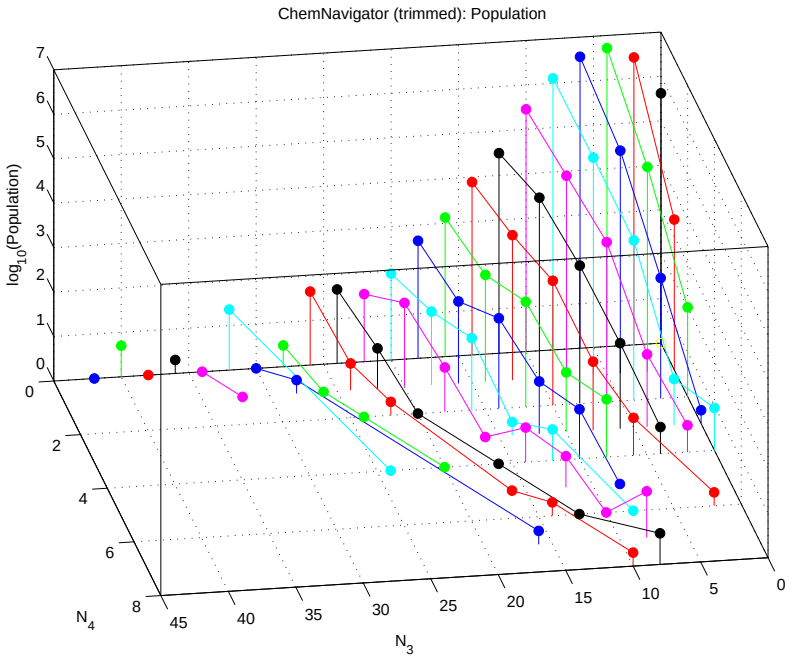

Figure 4 Populations of molecular scaffold topologies in the ChemNavigator database as a function of the number of 3- and 4-nodes, N_3 and N_4, and ordered, using connected stems of the same color, by the number of independent rings r. 5 outliers (topologies with $N_3 > 50$) have been excluded to make the main population trends of the graph easier to see.

For $r \geq 4$, the tendency towards structures with mostly 3-nodes starts to show up and becomes increasingly pronounced for higher values of r. This trend is especially notable in the Drugs collection.

Considering the 4-ring scaffolds in detail, in most of the databases examined, 16 out of the 17 possible topologies are present for the scaffolds consisting only of 3-nodes. The missing structure is the molecule labeled by 17 in Figure 2a which resembles a Möbius strip and is the only topology of the group that does not have a planar representation. Molecules with non-planar graphs are extremely rare; the first known example of a molecule with this topology was synthesized by Walba.[33] On the other extreme, most or all of the four 4-node only topologies are missing from the databases, except for PubChem which does have them all. For the mixed 3/4-node topologies, PubChem has examples of all and ChemNavigator nearly all, while the other databases are incomplete. The generated structures of GDB enumerate only 40–50% of the various 4-ring topologies. All of the minimal scaffolds of the 4-node only topologies and 13 out of 17 of the 3-node only topologies can be represented with 11 carbons or less, for example (see Figure 2a), so the filtering of chemically unstable and synthetically infeasible compounds (including non-planar graphs and all 3- and 4-member rings[2]) has removed a substantial fraction of topology types from this database.

The fraction of topologies compared to what is possible categorized by number of rings, or rings and 3- or 4-nodes, are indicators of the diversity of a database. Another is the population

$r =$	1	2	3	4	5	6	7	8
ChemNavigator	1.000	1.000	1.000	0.918	0.542	0.134	0.013	0.001
DNP	1.000	1.000	1.000	0.795	0.425	0.082	0.007	0.000
Drugs	1.000	1.000	0.750	0.411	0.078	0.005	0.000	0.000
PubChem	1.000	1.000	1.000	0.986	0.854	0.299	0.036	0.002
PC actives	1.000	1.000	1.000	0.712	0.280	0.039	0.002	0.000
WOMBAT	1.000	1.000	0.917	0.658	0.278	0.052	0.004	0.000
merged	1.000	1.000	1.000	0.986	0.859	0.310	0.039	0.002
GDB	1.000	1.000	1.000	0.425	0.041	0.001	0.000	0.000

Table 6 The fractions of scaffold topologies in the indicated databases with respect to the theoretical maxima per number of rings r.

r	N_4	N_3	Chem Navigator	DNP	Drugs	Pub Chem	PC actives	WOM BAT	merged	*GDB*
1	0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	0	2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3	0	4	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	1	2	1.000	1.000	0.800	1.000	1.000	1.000	1.000	1.000
	2	0	1.000	1.000		1.000	1.000	0.500	1.000	1.000
4	0	6	0.941	0.941	0.882	0.941	0.941	0.941	0.941	0.412
	1	4	1.000	0.900	0.467	1.000	0.900	0.933	1.000	0.433
	2	2	0.909	0.636	0.045	1.000	0.364	0.182	1.000	0.409
	3	0	0.250	0.250		1.000	0.250		1.000	0.500
5	0	8	0.930	0.887	0.479	0.944	0.831	0.831	0.944	0.127
	1	6	0.845	0.554	0.057	0.974	0.399	0.482	0.974	0.052
	2	4	0.364	0.303	0.004	0.868	0.127	0.053	0.873	0.022
	3	2	0.057	0.136		0.534			0.557	
	4	0	0.300			0.400			0.400	
6	0	10	0.642	0.451	0.054	0.851	0.345	0.482	0.851	0.008
	1	8	0.303	0.122	0.006	0.596	0.057	0.084	0.611	0.001
	2	6	0.059	0.053	0.000	0.228	0.011	0.009	0.241	
	3	4	0.009	0.022		0.071	0.002	0.001	0.080	
	4	2	0.007	0.007		0.060			0.060	
	5	0				0.214			0.214	

Table 7 The fractions of scaffold topologies in the indicated databases with respect to the theoretical maxima per numbers of 3- and 4-nodes, N_3 and N_4, for structures with $r = 1$–6 rings. Blank entries indicate that no representatives of that class of topologies were found in the specified database.

r	N_4	N_3	Chem Navigator	DNP	Drugs	Pub Chem	PC actives	WOM BAT	merged	*GDB*
0	0	0	0.245	8.633	6.492	2.466	3.837	1.641	1.225	15.414
1	0	0	2.979	11.831	16.630	8.212	10.771	6.588	5.248	41.721
2	0	2	20.808	15.390	25.492	24.094	18.384	16.680	21.981	29.521
	1	0	0.017	0.285	0.109	0.112	0.273	0.060	0.061	2.425
3	0	4	36.792	19.126	23.669	30.813	26.067	28.190	34.064	7.299
	1	2	0.287	1.172	0.547	0.523	0.664	0.659	0.399	2.090
	2	0	0.001	0.023		0.015	0.036	0.001	0.007	0.110
4	0	6	25.694	13.106	16.156	20.376	20.370	25.496	23.463	1.008
	1	4	0.729	2.829	1.349	0.931	2.132	1.184	0.838	0.300
	2	2	0.004	0.215	0.036	0.031	0.051	0.005	0.016	0.041
	3	0	0.000	0.006		0.002	0.013		0.001	0.001
5	0	8	9.178	8.721	4.413	7.382	8.652	11.800	8.492	0.057
	1	6	0.554	2.115	0.839	0.682	1.103	0.971	0.630	0.010
	2	4	0.028	1.097	0.036	0.064	0.180	0.073	0.047	0.002
	3	2	0.000	0.022		0.002			0.001	
	4	0	0.000			0.001			0.000	
6	0	10	2.004	3.517	1.714	2.044	3.001	3.524	2.071	0.001
	1	8	0.238	1.808	0.511	0.356	0.651	0.472	0.301	0.000
	2	6	0.028	0.657	0.036	0.063	0.219	0.106	0.046	
	3	4	0.000	0.137		0.006	0.013	0.010	0.003	
	4	2	0.000	0.005		0.001			0.000	
	5	0				0.000			0.000	

Table 8 The population percentages in the indicated databases with respect to the total database population for topologies with the given numbers of 3- and 4-nodes, N_3 and N_4, for structures with $r = 0$–6 rings. Blank entries indicate that no representatives of that class of topologies were found in the specified database. $r = 0$ values represent structures that contain no rings.

fraction of each distinct topology within the database. Table 8 displays the population percentages (with respect to the database's total population) of classes of topologies categorized by N_3 and N_4 for $r = 0$–6. Here, the bias against scaffolds containing 4-nodes is very strong. Moreover, while the distributions peak for 3-ring scaffolds containing only 3-nodes, there are significant percentages of structures containing 1–5 rings, and zero rings in some cases such as for DNP and Drugs.

Figure 5 displays for each database the population percentages of the scaffold topologies 1–33 shown in Figure 2a along with the situation when there are no rings present. Consider the six databases of known chemicals first. Several competing trends are evident. The fraction of topologies possessing even one 4-node (numbers 10–16) is very small. 3-node only topologies that contain a nonlinear cluster of three or more fused rings are also rare (i.e., topology numbers 5, 17–19, 21 and 26, as opposed to 6, 20, 27 and 28, which are well populated linear clusters).

Chem Navigator	DNP	Drugs	PubChem	PC actives	WOMBAT	*GDB*
2. 22.694	**4**. 11.831	**1**. 19.548	**1**. 20.740	**1**. 13.642	**3**. 13.200	**10**. 41.721
1. 19.646	**10**. 9.249	**4**. 16.630	**2**. 15.457	**3**. 11.101	**1**. 12.901	**14**. 24.765
3. 11.196	**18**. 9.226	**3**. 11.379	**3**. 11.509	**4**. 10.771	**2**. 10.160	**1**. 15.414
5. 6.474	**14**. 8.633	**14**. 6.492	**4**. 8.212	**2**. 7.652	**4**. 6.588	**4**. 4.755
6. 5.609	**3**. 6.643	**10**. 5.945	**5**. 4.033	**10**. 4.743	**5**. 5.101	**18**. 3.953
7. 3.590	**1**. 6.140	**26**. 5.872	**10**. 3.354	**18**. 4.681	**10**. 3.779	**46**. 2.765
4. 2.979	**26**. 5.356	**2**. 4.887	**6**. 2.824	**14**. 3.837	**6**. 3.510	**57**. 2.425
8. 2.505	**48**. 2.872	**18**. 3.939	**7**. 2.573	**11**. 3.130	**11**. 2.741	**58**. 0.977
9. 2.486	**2**. 2.437	**8**. 3.319	**14**. 2.466	**26**. 2.721	**18**. 2.399	**114**. 0.910
13. 2.094	**37**. 1.625	**23**. 2.553	**8**. 2.204	**7**. 2.220	**7**. 2.375	**122**. 0.610
79.273	64.012	80.564	73.372	64.498	62.754	98.295

Table 9 The percentages of the 10 most frequent topologies present in each of the databases examined. The numbers in boldface refer to the rank in the merged database; minimal scaffold topological representatives are displayed in Figure 6a. The numbers at the bottom are the sum of the 10 percentages above. At least half the population of each database lies above the horizontal line segment dividing the corresponding column.

Among the remaining topology types, those that consist of three or more rings emanating from a central vertex or vertices (i.e., 9, 31, 32 and 33) are the least common. In addition, it can be seen that the ChemNavigator and PubChem values show the same general qualitative trends compared to the other databases (ChemNavigator does, however, have fewer no-ring and single-ring structures than PubChem). Also, DNP topologies show a distinctive trend, having a higher proportion of linear fused ring assemblies than other databases (e.g., 6 and 20), but very few topologies involving multiple rings emanating from a central vertex or vertices. DNP (and Drugs) also have a considerable percentage of structures with no rings.

GDB also has a considerable percentage of structures with no rings. The other trends are also similar, except that unlike the other databases, topologies possessing a 4-node are not quite as rare. In addition, GDB favors the maximally fused 2- and 3-ring topologies, numbers 3 and 5, respectively, more than do the known chemicals.

Table 9 presents the population percentages of the 10 most frequent topologies in each of the databases. These topologies are identified by their rank in the merged database; their corresponding minimal scaffolds are displayed in Figure 6a and examples of actual molecules are provided in Figure 6b.

Only 18 distinct topologies are found in the collection of the 10 most common topologies from each of the six databases of known chemicals, making up from 62.8–80.6% of the total populations. None of these topologies possess 4-nodes. It can be seen that there is some tendency in DNP and Drugs to have scaffolds with more fused rings than in the other databases (see Table 10), although significant percentages of DNP and Drugs molecules do not contain any rings at all. In general, the

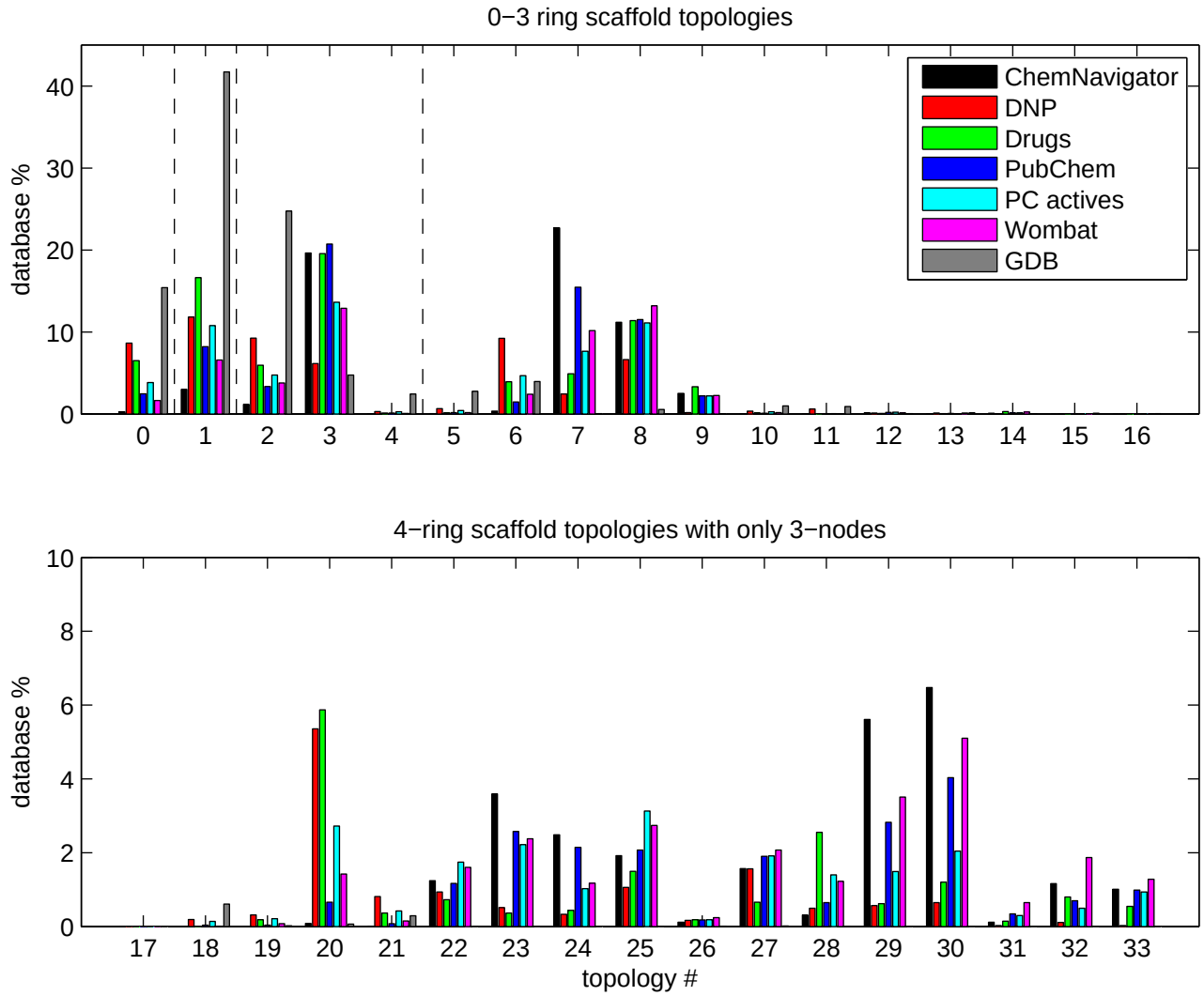

Figure 5 The percentage frequencies of the first 33 scaffold topologies of Figure 2 in the indicated databases. The entry labeled zero indicates the database percentages of structures that do not contain rings. The dashed lines in the top graph divide the results into sets of topologies possessing 0, 1, 2 or 3 rings, respectively. The bottom graph displays the frequencies for 4-ring topologies containing only 3-nodes. Note that the vertical scales in the two graphs are different.

Chem Navigator		DNP		Drugs		PubChem		PC actives		WOMBAT		merged		*GDB*	
3	1	1	1	2	1	2	1	2	1	3	2	2	1	2	2
2	1	2	2	1	1	3	1	3	2	2	1	3	1	0	0
3	2	3	3	3	2	3	2	1	1	3	1	3	2	2	1
4	2	0	0	0	0	1	1	3	1	1	1	1	1	1	1
4	1	3	2	2	2	4	2	2	2	4	2	4	2	3	3
4	2	2	1	4	4	2	2	3	3	2	2	4	1	3	3
1	1	4	4	3	1	4	1	0	0	4	1	4	2	2	1
3	1	5	5	3	3	4	2	4	2	4	2	3	1	3	2
4	1	3	1	3	1	0	0	4	4	3	3	4	1	3	3
5	2	5	4	4	3	3	1	4	2	4	2	2	2	4	4

Table 10 The number of rings (first number in each column pair) and the size of the largest fused ring system (second number) in each of the 10 most frequent topologies in the indicated databases.

biologically oriented databases had more top 10 topologies that exhibited fused rings than the more general databases (i.e., ChemNavigator and PubChem). For GDB, five additional topologies not included in the above 18 define its second five most frequent topologies (7.7% of the population; note that 90.6% of the population is included in the top five topologies). Three of these contain 4-nodes. There is also a tendency toward fused rings in this database.

4 Conclusions

We report the scaffold distribution and topological properties for six databases of existing chemicals: ChemNavigator, DNP, Drugs, PubChem, PubChem "actives" and WOMBAT, to which we include a comparison with GDB, a collection of virtual small organic molecules. The greatest topological diversity is observed in PubChem. This is not surprising, since this is a public repository where information providers routinely upload a large variety of chemical structures. For 6-ring scaffolds, PubChem molecules cover less than a third of the possible theoretical topological space (limited to $\leq$ 4-nodes), and this fraction declines rapidly for greater numbers of rings.

The least topologically diverse set is GDB, which is not surprising either. GDB has been developed using a "bottom-up" strategy for chemical space enumeration, where changes occur incrementally, one atom or one bond at a time algorithmically added to a list. By contrast, we regard this work on exhaustive enumeration as a "top-down" strategy, where the landscape of possibilities is mapped out to completeness. Our earlier, unpublished work, modifying one SMILES atom at a time, produced over 1.6 billion unique SMILES—all C.sp3 based, and all single bonds, up to 8 rings and 20 atoms.[35–37] We abandoned that strategy because this approach would quickly reach the asymptotic wall of combinatorial explosion: consider that, corresponding to the 1.6 billion alkanes, there are probably 1 billion mono-alkenes, mono-amines and mono-alcohols, to name a few possibilities approximating for symmetry-related redundancy. The GENSMI algorithm be-

came increasingly tedious to use at higher levels of complexity, as it had to compare every new SMILES to all others within a given class (e.g., 6 rings and 15 atoms). Using the "top-down" strategy, one can drill down and achieve completeness using a divide and conquer approach: completeness tests would be limited to only one topological subset, without having to compare all newly generated molecules to all others having the same number of rings and nodes. Thus, the GDB approach continues to be useful in exploring all possibilities of the low-molecular-weight chemical space, but topological landscaping brings a distinct perspective to the same problem.

We found a strong bias in all chemical collections of existing compounds toward 3-node topologies, i.e., vertices branching out in three different directions (see Tables 7 and 8). Other topological classes, such as those containing a nonlinear cluster of three or more fused rings (topology numbers 5, 17–19, 21 and 26 in Figure 2a) or three or more rings linked to a central vertex or vertices (topology numbers 9, 31–33), are relatively uncommon (the latter especially in the case of DNP) as was seen in Figure 5.

The average fraction of atoms that make up the scaffold tends to be lower for biologically active molecules, indicating that they have on average a higher number of chemical moieties substituted to the central scaffold, presumably to enhance pharmacophore diversity, thus contributing to biological activity.

We see a modest tendency toward more fused rings in the biologically oriented databases (especially linear fused ring assemblies in DNP), as well as a tendency toward fewer overall rings in DNP and Drugs, both of which also have significant fractions of molecules that do not contain any rings at all.

Looking at the 10 most frequent topologies for each database, we find that a small number of topologies characterize most of the molecules. Only 8 topologies (1–5, 10, 14 and 18 in Figure 6) are needed to characterize half the population of the each of the seven databases. 62.8–90.6% of the database populations are characterized by 18 topologies. On the other hand, most of the topologies encountered are represented by a single or very small number of examples.

We have developed a website[38] interfaced to a MySQL database, where one can enter a SMILES and get back a page displaying data relevant to the molecule's scaffold topology. Included are 2D diagrams of the original molecule and a minimal representative of the scaffold topology, some numerical details of the topology, the number of matches of this topology in the public database PubChem, and some further examples of this topology from this database. The SMILES of all molecules possessing this topology can also be extracted from the database.[39]

In addition, the user can access theoretical results from our enumeration of all possible scaffold topologies. Depictions of all minimal representatives of scaffold topologies up through 4 rings are available. We will continue to extend the capabilities of this site.

To compute a scaffold topology, we are in effect collapsing a molecule to its essential ring and connecting linear structure. In the paired paper of Pollock et al.,[1] scaffold topologies are systematically built up from the most basic topologies of one and two rings. Once a topology is available, a minimal or more complicated scaffold can be produced. The two papers, therefore, are looking at the problem of molecular classification from the opposing points of view of what is possible and what actually occurs.

Scaffold topologies are a first step toward an efficient classification scheme of the molecules found in chemical databases. For example, to extract the approximately 26 million topologies in the merged database on a 2.2 GHz Linux system with 32 Gb of memory required less than 4 hours of CPU time.

Acknowledgments

We wish to thank Cristian Bologa for his help and advice. This research was funded in part by Tobacco Settlement Funds and the University of New Mexico Initiative for Cross Campus Collaboration in the Biological and Life Sciences.

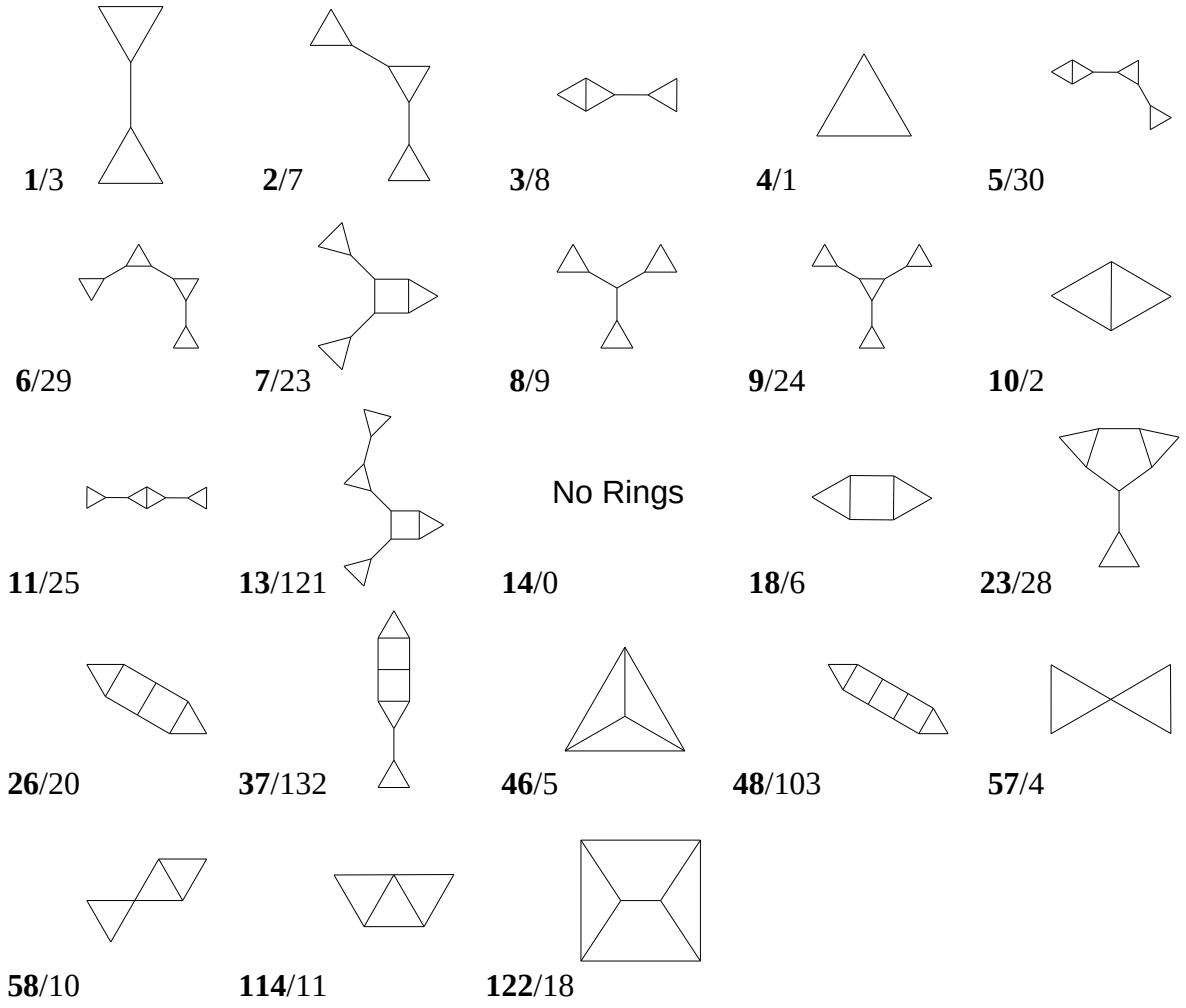

Figure 6 a. The most frequent topologies (represented by their minimal scaffolds) present in the databases examined, numbered (in boldface) by their rank in the merged database. The second value for each entry is the topology number, 1–33 and 86–89 of which are shown in Figure 2a.

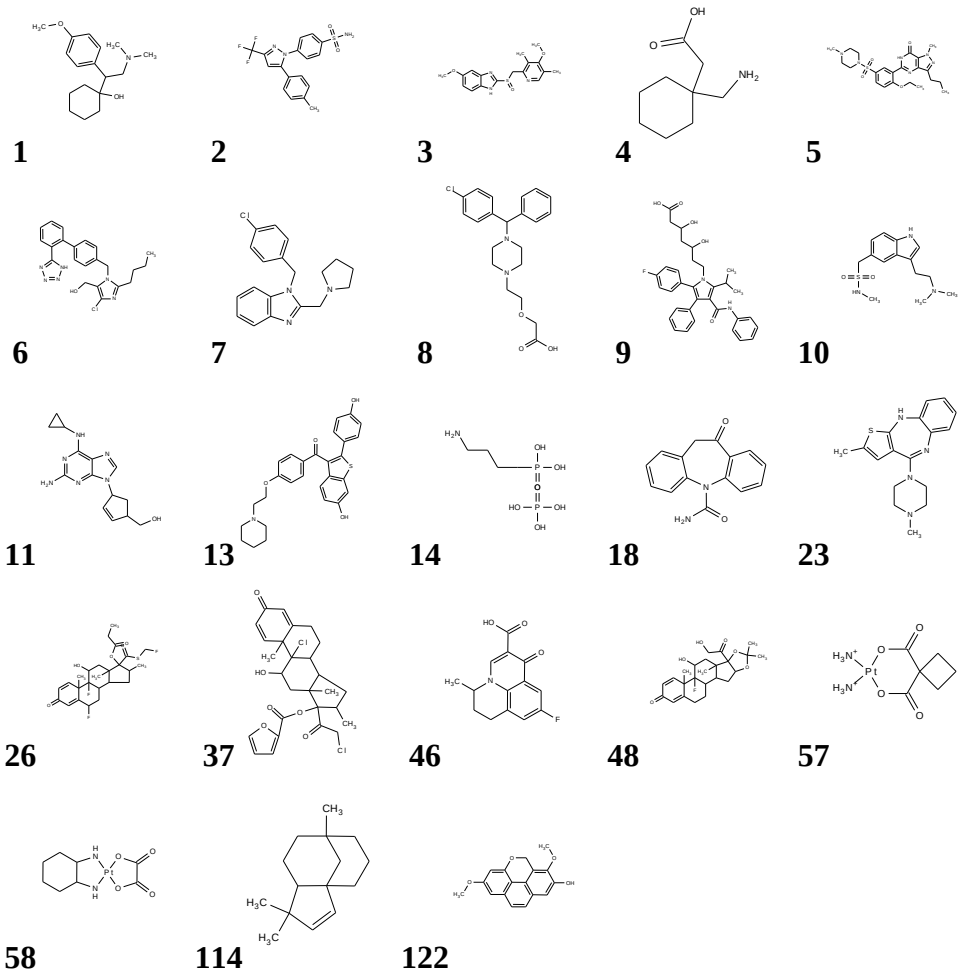

Figure 6 b. Examples[34] from the databases examined of the most frequent topologies present, numbered by their rank in the merged database (compare with Figure 6a).

References and Notes

1. Pollock, S.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. "Topological Analysis of Molecular Scaffolds: 1. Enumeration of Ring Topologies", *J. Chem. Info. Model., submitted (accompanying this paper)*.

2. Fink, T.; Bruggesser, H.; Reymond, J.-L. *Angewandte Chemie International Edition* **2005,** *44,* 1504–1508.

3. de Laet, A.; Hehenkamp, J. J. J.; Wife, R. L. *Journal of Heterocyclic Chemistry* **2000,** *37,* 669–674.

4. Hehenkamp, J. J. J.; de Laet, R. C.; Parlevliet, F. J.; Verheij, H. J.; Wife, R. L. Navigating the real and virtual chemical worlds. In *Proceedings of the 2000 Chemical Information Conference*; Collier, H., Ed.; Infonortics: Annecy, France, 2000.

5. Oprea, T. I.; Gottfries, J. *Journal of Combinatorial Chemistry* **2001,** *3,* 157–166.

6. Oprea, T. I. *Current Opinion in Chemical Biology* **2002,** *6,* 384–389.

7. http://nihroadmap.nih.gov/molecularlibraries/.

8. Wilkens, S. J.; Janes, J.; Su, A. I. *Journal of Medicinal Chemistry* **2005,** *48,* 3182–3193.

9. There is one exception to this statement for the situation when a scaffold consists of a single ring. Here, the topology will consist of a 2-node with a loop as otherwise there would be no node at all.

10. Trinajstić, N.; Nikolić, S.; Knop, J. V.; Müller, W. R.; Szymanski, K. *Computational Chemical Graph Theory: Characterization, Enumeration and Generation of Chemical Structures by Computer Methods;* Ellis Horwood: New York, 1991.

11. Filip, P. A.; Balaban, T.-S.; Balaban, A. T. *Journal of Mathematical Chemistry* **1987,** *1,* 61–83.

12. Mekenyan, O.; Bonchev, D.; Balaban, A. *Journal of Mathematical Chemistry* **1988,** *2,* 347–375.

13. Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. *Journal of Mathematical Chemistry* **1993,** *12,* 309–318.

14. Berger, F.; Flamm, C.; Gleiss, P. M.; Leydold, J.; Stadler, P. F. *Journal of Chemical Information & Computer Sciences* **2004,** *44,* 323–331.

15. Trinajstić, N. *Journal of Mathematical Chemistry* **1988,** *2,* 197–215.

16. Lee, S.-L.; Yeh, Y.-N. *Journal of Mathematical Chemistry* **1993,** *12,* 121–135.

17. West, D. B. *Introduction to Graph Theory;* Prentice Hall: Upper Saddle River, New Jersey, Second ed.; 2001.

18. 1. Neurontin, 2. Imitrex, 3. Effexor XR, 4. Paraplatin, 5. flumequine, 6. Trileptal, 7. Celebrex, 8. Nexium, 9. Zyrtec, 10. Eloxatin, 11. clovene, 12. fenspiride, 13. pilsicainide, 14. phencyclidine, 15. AIDS133821, 16. NSC263872, 18. Agrostophyllin, 19. setiptiline, 20. Flonase, 21. apomorphine, 22. Kytril, 23. clemizole, 24. Lipitor, 25. Trizivir, 26. Levaquin, 27. Zofran, 28. Zyprexa, 29. Cozaar, 30. Viagra, 31. Kaletra, 32. Allegra, 33. Zosyn, 86. NSC177445, 87. NSC160443, 88. CBDivE_010142, 89. tri-iron-dodecacarbonyl.

19. Moss, G. P. *Pure and Applied Chemistry* **1999,** *71,* 531–558.

20. ChemNavigator.com, Inc., "iResearch Library", 2006 http://www.chemnavigator.com/.

21. Chapman & Hall/CRC, London "Dictionary of Natural Products", 2006 Version 14.1.

22. National Center for Biotechnology Information, "PubChem", 2006 http://pubchem.ncbi.nlm.nih.gov/.

23. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004.

24. Weininger, D. *Journal of Chemical Information & Computer Sciences* **1988,** *28,* 31–36.

25. Daylight Chemical Information Systems, Inc., Aliso Viejo, California "Daylight Theory Manual", 2007 http://www.daylight.com/dayhtml/doc/theory/.

26. OpenEye Scientific Software, Inc., "OEChem — C++ Theory Manual Version 1.4", 2006 http://www.eyesopen.com/docs/.

27. Note that the merged database can have duplicate entries, even when duplicate SMILES are removed because there is no complete canonicalization algorithm for SMILES, but this will have no effect on the overall number of distinct topologies present.

28. Reymond, J.-L. "Reymond Group Cheminformatics Site", http://www.dcb.unibe.ch/groups/reymond/cheminf/index.html, 2007.

29. Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. *Journal of Medicinal Chemistry* **2003,** *46,* 3257–3274.

30. Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proceedings of the National Academy of Sciences of the USA* **2005,** *102,* 17272–17277.

31. Bone, R. G. A.; Villar, H. O. *Journal of Computational Chemistry* **1997,** *18,* 86–107.

32. Feher, M.; Schmidt, J. M. *Journal of Chemical Information and Computer Sciences* **2003,** *43,* 218–227.

33. Walba, D. M. *Tetrahedron* **1985,** *41,* 3161–3212.

34. 1. Effexor XR, 2. Celebrex, 3. Nexium, 4. Neurontin, 5. Viagra, 6. Cozaar, 7. clemizole, 8. Zyrtec, 9. Lipitor, 10. Imitrex, 11. Trizivir, 13. Evista, 14. Fosamax, 18. Trileptal, 23. Zyprexa, 26. Flonase, 37. Nasonex, 46. flumequine, 48. Nasacort AQ, 57. Paraplatin, 58. Eloxatin, 114. clovene, 122. Agrostophyllin.

35. Kappler, M. A.; Allu, T. K.; Oprea, T. I. "GENSMI: Generation of Genuine SMILES", http://www.daylight.com/meetings/mug04/Kappler/GenSmi.html, 2004 MUG'04: 18th Daylight User Group Meeting.

36. Kappler, M. A. "GENSMI: Exhaustive Enumeration of Simple Graphs", http://www.daylight.com/meetings/emug04/Kappler/GenSmi.html, 2004 EuroMUG 2004.

37. Kappler, M. A. "GENSMI: Exhaustive Enumeration of Simple Graphs", http://biocomp.health.unm.edu/events/Biocomputing@UNM2005/Presentations/Kappler/GenSmi.html, 2005 Biocomputing @ UNM 2005.

38. http://topology.health.unm.edu/.

39. Since the analyses were performed on chemical databases in which all salts were removed and then any non-unique entries created by this processing eliminated, there will often be more actual entries in the databases for a given topology than the number reported, but if only unique SMILES are considered, then these numbers will be identical.