# Scaffold Topologies. 2. Analysis of Chemical Databases

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

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131, Division of Biocomputing, Department of Biochemistry and Molecular Biology, University of New Mexico Health Sciences Center, Albuquerque, New Mexico 87131, and AstraZeneca R&D Mölndal, Sweden

Received September 13, 2007

We have systematically enumerated graph representations of scaffold topologies for up to eight-ring molecules and four-valence atoms, thus providing coverage of the lower portion of the chemical space of small molecules (Pollock et al. *J. Chem. Inf. Model.*, this issue). Here, we examine scaffold topology distributions for several databases: ChemNavigator and PubChem for commercially available chemicals, the Dictionary of Natural Products, a set of 2742 launched drugs, WOMBAT, a database of medicinal chemistry compounds, and two subsets of PubChem, "actives" and DSSTox comprising toxic substances. We also examined a virtual database of exhaustively enumerated small organic molecules, GDB (Fink et al. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508), and we contrast the scaffold topology distribution from these collections to the complete coverage of up to eight-ring molecules. For reasons related, perhaps, to synthetic accessibility and complexity, scaffolds exhibiting six rings or more are poorly represented. Among all collections examined, PubChem has the greatest scaffold topological diversity, whereas GDB is the most limited. More than 50% of all entries (13 000 000+ actual and 13 000 000+ virtual compounds) exhibit only eight distinct topologies, one of which is the nonscaffold 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 three-way scaffold connections (3-nodes) are much more frequent compared to four-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 coarse-grained 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 chemical structures of drugs. Pharmaceutical R&D, and more recently the National Institutes of Health (NIH) and other 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 "druglike" 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 is the chemical space received recent attention with the launch of the NIH Roadmap molecular libraries initiative.[7] As the NIH is embarking in the selection and biological screening of 300 000 chemicals in search of novel chemical probes, the issue of which chemicals to acquire (from over 10 000 000 commercial structures) is not a trivial one.

Previous enumerations of the CSSM include: Kappler,[8–11] who generated all single-bonded carbon-only structures up through $r = 8$ rings and $21 - r$ atoms; Kerber et al.,[12] who produced all valid nonionic molecular formulas composed of C, N, O, and H using standard valences up to a molecular weight of 150 Da and then generated all possible structures corresponding to each formula; and Fink et al.,[2,13] who completely enumerated all C, N, O, and F structures up to 11 atoms and 160 Da and then filtered them for simple valency, synthetic feasibility, and stability. Each of these studies created a fine-grained coverage of a lower portion of the CSSM in which potentially feasible organic molecules were produced.

Here, we compare the results of a coarser-grained classification, scaffold topologies, which themselves are not potential molecules but represent the elemental ring structures of organic molecules, against a variety of generic and biologically oriented chemical databases as well as the collection generated by Fink et al. This provides a high-level view of the fundamental topological character of these databases and a unique insight into a large class of known and possible new chemicals.

## 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.

* Corresponding author e-mail: toprea@salud.unm.edu.
† University of New Mexico.
‡ University of New Mexico Health Sciences Center.
§ AstraZeneca R&D Mölndal.

10.1021/ci700342h CCC: $40.75 
Published on Web 07/08/2008

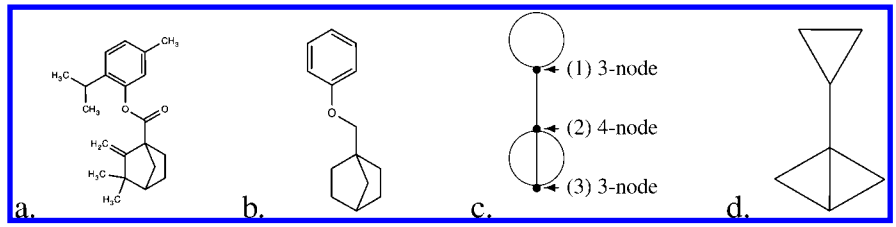



**Figure 1.** (a) (5-methyl-2-propan-2-yl-phenyl)-3,3-dimethyl-2-methylidene-bicyclo[2.2.1]heptane-1-carboxylate [SMILES: CC(C)c1ccc(C)cc1OC(=O)C2(CCC3C2)C(=C)C3(C)C]. (b) The scaffold corresponding to this molecule [C1CC2CCC1(C2)COc3ccccc3]. (c) The topology corresponding to this scaffold (nodes are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

**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.

Scaffolds can also admit atoms double-bonded to ring atoms,[14] but we do not include these special atoms in our description of scaffold 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. Scaffold 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 scaffold topologies consisting of 3-nodes and 4-nodes,[15] 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.

Since the recursive process of extracting a scaffold from a molecule involves, in the worst case, eliminating one atom (node) per step, where each step may require examining the entire adjacency matrix (i.e., $n_M^2$ entries, $n_M$ counting the number of atoms in the original molecule), the time complexity of this process cannot exceed $n_M^3$. Hereafter, for simplicity, we will often shorten the term *scaffold topology* to *topology*, but we will always mean a graph as constructed above unless indicated otherwise.

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

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $N_3 =$ | $2(r-1)$ | $2(r-2)$ | $2(r-3)$ | $\cdots$ | $2(r-i-1)$ | $\cdots$ | 0 |
| $N_4 =$ | 0 | 1 | 2 | $\cdots$ | $i$ | $\cdots$ | $r-1$ |

and hence, for a topology, the total number of nodes ($n$) and edges ($e$) satisfies

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

**2.2. Comparing Topologies.** Several schemes for uniquely characterizing molecular graphs have appeared (Trinajstíc et al.[16] describes a number of methods; see also refs 17–19). This has been a difficult task, as complex graphs can have sophisticated symmetries that defy easy classification (see Berger et al.[20] 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 bond is single, double, or triple, all of 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 ordered return index, an $n \times n$ matrix, as discussed in the companion paper.[1]

We have exhaustively verified that, after sorting with respect to the number of rings and the number of 3- or 4-nodes, the ordered return index is sufficient to distinguish topologies with up through eight rings for molecules with atoms of valence up to four.[1] Therefore, this set of values under the given conditions establishes a unique characterization of scaffold topologies. For $r = 11$, we know of examples of topologies that have the same ordered return indices yet are distinct.[1] The ordered 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.** The Total Number of Distinct Scaffold Topologies for One through Eight Rings (Top) and Categorized by the Number of 3-Nodes, $N_3$, and 4-Nodes, $N_4$ (Bottom)[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 |
| 1 | | | | 1 | 5 | 30 | 193 | 1496 |
| 0 | | | | 1 | 2 | 5 | 17 | 71 |

| $N_4$ \ $N_3$ | 10 | 12 | 14 |
|---|---|---|---|
| 7 | | | |
| 6 | | | |
| 5 | | | |
| 4 | | | |
| 3 | 365994 | | |
| 2 | 136666 | | |
| 1 | 13343 | 388 | 2592 | 21096 |
| 0 | | | |

[a] The diagonal colors indicate the number of rings ($r$). Note that the (0, 0) topology is a loop with a 2-node.

Moreover, we have found that the diagonal of the ordered return index is an excellent discriminator of topologies, which we use to speed database searches. We need only compare $n$ diagonal entries rather than perform full comparisons of $n \times n$ matrices in nearly all cases. Out of a total of 1 547 689 topologies containing eight rings or less, there are 2, 9, and 185 examples, respectively, in which groups of four, three, and two ordered return indices, respectively, share a common diagonal but the full matrices differ, resulting in a total of 405 ambiguous cases when the diagonal is used for discrimination. In such events, we fall back to full-matrix comparisons within the small groups of four, three, or two ordered return indices.

Table 1 shows the results of enumerating all possible topologies up through eight rings. In Figure 2a, all scaffold topologies with one to three rings are presented as well as the 3-node-only and 4-node-only four-ring topologies. A total of 52 mixed 3-/4-node four-ring topologies are not shown. See Table 2 for further identifications. The corresponding minimal scaffolds require 3, 4−6, 4−10, and 5−14 nodes, respectively, for $r = 1-4$. 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.

**2.3. Spiro Atoms.** A spiro atom is the unique common member of two or more otherwise disjoint ring systems.[22] 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 it 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 makes 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 (≥ 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

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

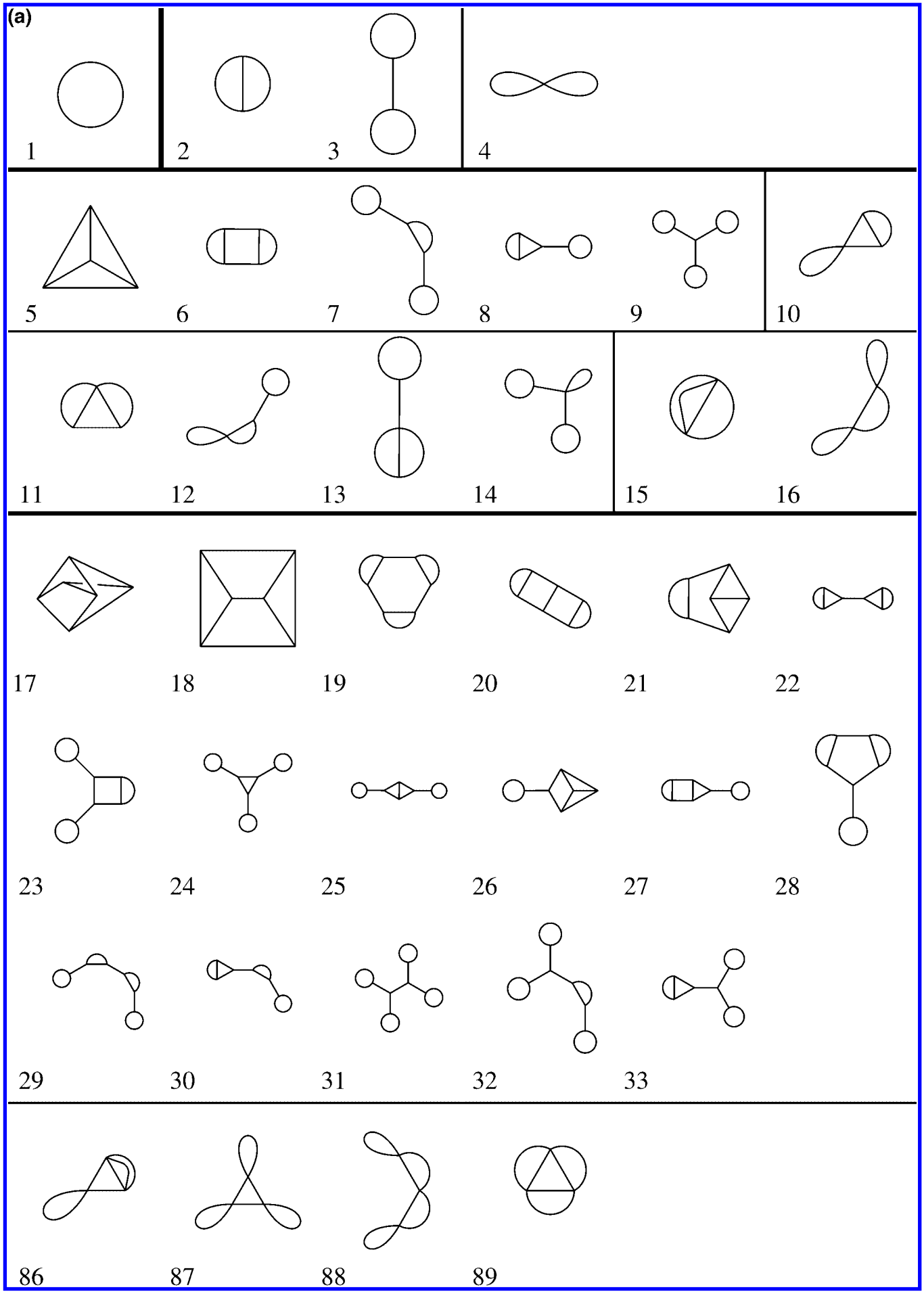
The average number of independent rings per scaffold is

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

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

## 3. ANALYSIS OF SOME EXISTING DATABASES

We computed scaffold topologies for the molecules found in several databases, as follows: ChemNavigator,[23] which collects commercially available chemicals; the Dictionary of Natural Products (DNP);[24] an in-house compilation of 2742 unique small molecules that are, or have been, launched drugs (Drugs); PubChem,[25] a public repository of small molecules which have been characterized for biological activity; PC "actives", which is the PubChem subset labeled as "active"; the Distributed Structure−Searchable Toxicity (DSSTox)[26] database, also a subset of PubChem; and WOMBAT,[27] a collection of small molecules with known biological activity from medicinal chemistry literature (see Table 3). For each database, we processed SMILES[28,29] for all of the molecules; removed salts, hydration information and counterions; and then eliminated nonunique entries. We converted each SMILES to an adjacency matrix using OEChem,[30] stripped each molecule down to its simplified scaffold (see section 2), and then extracted the distinct topologies and cataloged their frequencies. Furthermore, we carried out the same procedure on the nonredundant union of all databases,[31] 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 five or less representatives. We also examined the Generated Database of Chemical Space of Small Molecules (GDB),[32] in which all organic molecules with 11 or less main atoms and a molecular weight of less than 160 Da have been algorithmically generated and then filtered down for simple valency, synthetic feasibility, and stability.[2]


(a)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
86
87
88
89

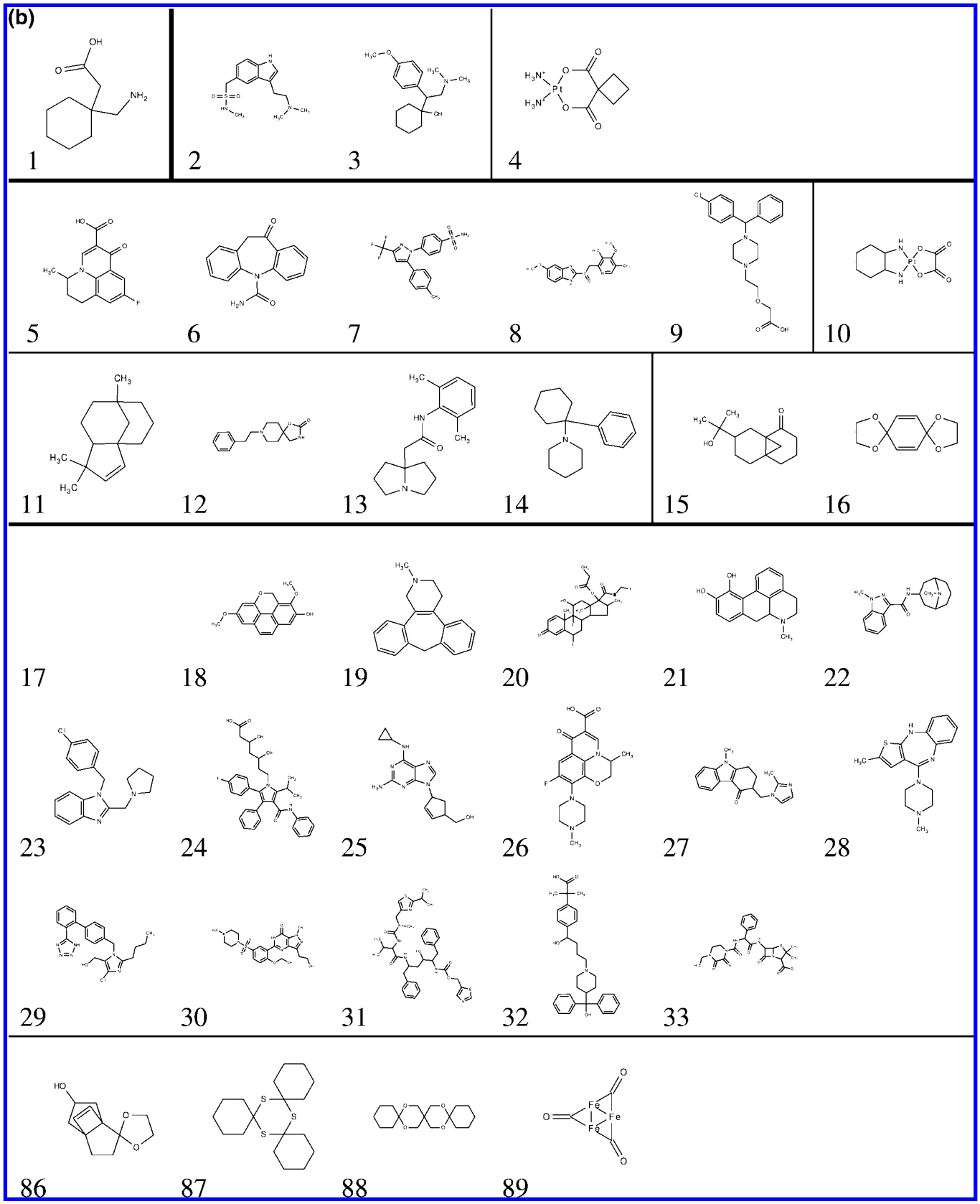



**Figure 2.** (a) All one- to three-ring scaffold topologies and all four-ring topologies possessing only 3-nodes or only 4-nodes. See Table 2 for further identification. (b) Examples[21] from the databases examined of molecules that exhibit each one- to three-ring topology and each four-ring topology possessing only 3-nodes or 4-nodes, corresponding to the topologies in part a. Note that none of the databases examined possessed an example of topology number 17. See Table 2 for further identification.

In Table 4, the scaffolds and topologies for each database are compared with the merged totals (columns 2 and 3), and then with the number of SMILES (molecules) in the database (columns 4 and 5). Relative to the merged database, of the two largest chemical databases, PubChem produced 5% fewer distinct scaffolds but nearly 6 times more topologies than

**Table 2.** Descriptors for the Scaffold Topologies in Figure 2a

| $r$ | 1 | 2 | | 3 | | | 4 | |
|---|---|---|---|---|---|---|---|---|
| $N_4$ | 0 | 0 | 1 | 0 | 1 | 2 | 0 | 3 |
| $N_3$ | 0 | 2 | 0 | 4 | 2 | 0 | 6 | 0 |
| topologies | 1 | 2−3 | 4 | 5−9 | 10−14 | 15−16 | 17−33 | 86−89 |

**Table 3.** Databases Examined, Including a Merged One Constructed from All the Others, Their Sizes, the Number of Distinct Scaffolds Produced, and the Number of Distinct Topologies Discovered[a]

| database | version | unique SMILES | distinct scaffolds | distinct topologies[b] |
|---|---|---|---|---|
| ChemNavigator | October 2006 | 14041970 | 1313911 | 3880 |
| DNP | April 2006 | 132434 | 31819 | 3199 |
| Drugs | 2006 | 2742 | 1312 | 155 |
| PubChem[c] | November 7, 2006 | 11595690 | 1210092 | 22612 |
| PC actives | November 7, 2006 | 38881 | 17200 | 1052 |
| DSSTox | November 7, 2006 | 3915 | 1067 | 115 |
| WOMBAT | December 2006 | 149451 | 44038 | 1333 |
| *merged* | | 25029900 | 2056025 | 23737 |
| *GDB* | 2005 | 26434571 | 1076051 | 76 |

[a] GDB, a generated database, was analyzed separately. [b] Since the ordered return index is not guaranteed to completely distinguish scaffold topologies for $r > 8$, the numbers presented in this table generally are lower bounds; however, we do believe them to be good estimates, as we employed additional strategies for >eight-ring structures to help provide further resolution, such as computing multiple ordered return indices using different values in the adjacency matrix to represent loops. In addition, the total numbers of topologies for each database with $r > 8$ were small: < 0.62%, except for DNP (3.68%) and PC actives (1.33%), both small databases. [c] PubChem substances were used, as at the time the analyses were performed, substances but not compounds could be identified as active.

**Table 4.** For Each Database Examined, the Percentage that the Number of Distinct Scaffolds (Topologies) Makes with Respect to the Total Number of Distinct Scaffolds (Topologies) in the Merged Database and the Percentage Ratio of Scaffolds and Topologies to Unique SMILES (Molecules) Present in the Database

| database | % scaf./ merged scaf. | % top./ merged top. | % scaf./ SMILES | % top./ SMILES |
|---|---|---|---|---|
| ChemNavigator | 63.905 | 16.346 | 9.357 | 0.0276 |
| DNP | 1.548 | 13.477 | 24.026 | 2.4155 |
| Drugs | 0.134 | 0.653 | 47.848 | 5.6528 |
| PubChem | 58.856 | 95.261 | 10.436 | 0.1950 |
| PC actives | 1.891 | 4.432 | 44.238 | 2.7057 |
| DSSTox | 0.190 | 0.484 | 27.254 | 2.9374 |
| WOMBAT | 7.269 | 5.616 | 29.467 | 0.8919 |
| *merged* | 100.000 | 100.000 | 8.214 | 0.0948 |
| *GDB* | | 0.320 | 4.071 | 0.0003 |

ChemNavigator. DNP made a small (1.5%) relative contribution of scaffolds, but a good-sized (13.5%) contribution of topologies. Nearly 99% of GDB's scaffolds did not overlap with the merged database; however, all of its topologies did.

The last two columns of Table 4 provide an indication of the databases' scaffold and scaffold topological diversities. The smaller, biologically oriented databases (especially Drugs) have the greatest diversities, while GDB, with only 76 unique topologies but over 26 000 000 SMILES, has a very low topology-to-SMILES ratio, although its scaffold-to-SMILES ratio is much more in line with the other, especially the two large, databases. Thus, collections of very small molecules (<160 Da) may have many scaffolds, but their underlying scaffold topologies remain quite limited. We note that the topology-to-SMILES ratio appears to be inversely correlated with the size of the databases (the larger the database, the smaller the ratio), and the scaffold-to-SMILES ratios are partially so, which suggests that a larger database typically contains more examples of a topology or a scaffold.

**Table 5.** For Each Database, the Percentage of Molecules That Do Not Contain Rings, the Maximum Number of Rings Found in a Single Compound, and the Population of Molecules That Possess at Least One 5- Or 6-Node

| database | % no rings | Maximum rings | >4-nodes population |
|---|---|---|---|
| ChemNavigator | 0.245 | 62 | 95 |
| DNP | 8.633 | 32 | 61 |
| Drugs | 6.492 | 18 | 0 |
| PubChem | 2.466 | 165 | 6488 |
| PC actives | 3.837 | 23 | 198 |
| DSSTox | 25.057 | 11 | 0 |
| WOMBAT | 1.641 | 34 | 0 |
| *merged* | 1.225 | 165 | 6593 |
| *GDB* | 15.414 | 6 | 0 |

Xue and Bajorath[33] found that the scaffold-to-compound percentage was 44.53% for the Optiverse screening library based on diversity design (117 976 chemicals) and 26.94% for the Maybridge collection of compounds and intermediates used in medicinal chemistry (58 239 chemicals). For the biologically oriented databases here, the numbers (and database sizes) are comparable, ranging between 47.85% for Drugs to 24.03% for DNP.

As can be seen in Table 5, nearly all of the molecules contain rings and can be stripped down into scaffolds (these findings are similar to those of Lewell et al.[34] and Koch et al.).[35] Note, however, that 8.6% of the DNP structures, 6.5% of the Drugs, and 3.9% of the PC actives, all biologically oriented, do not contain rings, as does 25.1% of DSSTox, by far the largest database percentage. A total of 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 a PubChem copper tetracarboranylphenylporphyrin with $r = 165$ ($N_6 = 8, N_5 = 88, N_3 = 32$). The next largest, a protein HIV inhibitor 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 PC 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, DSSTox, WOMBAT, and GDB contain no such structures at all. We found no scaffolds that had nodes with degrees > 6. Therefore, we ignored such higher-degree nodes and concentrated on topologies that contained 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 they 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,

**Table 6.** Basic Database Measures: Average Fraction of Atoms Per Molecule That Make up the Scaffold, Average Fraction of Branch Points (≥ 3-Nodes) Per Scaffold, Average Scaffold Connectivity (Node Degree), Average Number of Independent Rings Per Scaffold[a]

| database | fraction scaffold | fraction ≥ 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 |
| DSSTox | 0.649 | 0.239 | 2.133 | 2.225 |
| 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 |

[a] See Methods for computational details.

as illustrated in Table 6. One such measure is the average fraction of atoms per molecule that makes up the scaffold (see the first data column). In the biologically oriented databases (DNP, Drugs, PC actives, DSSTox, and WOMBAT), this fraction averages 0.61−0.71, while in the other known chemical databases, that average is higher, ranging 0.72−0.74. 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 exploring biological activity around a given scaffold. The lowest fraction of scaffold atoms (0.60) is in GDB, which indicates that these molecules contain a considerable fraction of nonscaffold 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 only occasionally leads to synthesized compounds.

Others[34] 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 was present in the scaffolds[34] (as we define them here).

Another topological measure is the fraction of scaffold atoms that are essential for defining the scaffold topology of multiring systems. This is the fraction of branching (≥ 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 0.22 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[36] 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 6. This measure is quite similar among databases of known chemicals, averaging around 2.21, with DNP having a marginally higher value and DSSTox a somewhat lower value. GDB scaffolds, averaging 2.05, are, on average, less connected.

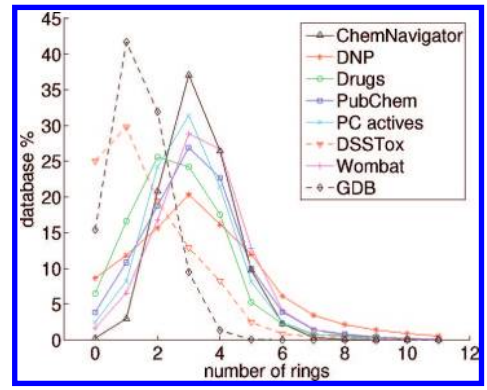



**Figure 3.** The population percentages in the indicated databases with respect to the total database population for the number of rings per scaffold.

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.[35] Natural products have the highest average number of rings and marketed drugs the least, with natural product derivatives and combinatorially synthesized chemicals in between.[37] Our results show generally similar trends, but much less pronounced, since we examine larger collections (except for the drugs). DSSTox is an exception, with a lower average number of rings than any of the other databases of known chemicals. GDB has a much lower average ring count than the other databases, which is merely indicative of the artificial limits imposed by enumeration (160 Da, 11 atoms).

Figure 3 shows how the database population percentages correspond to the number of rings in more detail. All databases of known chemicals except DSSTox 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 above databases, peaking at one ring and then dropping sharply, nearly reaching zero at five rings. This is, of course, consistent with the limitations imposed on the database by the upper bound of 11 heavy atoms. DSSTox also peaks at one ring; however, its tail drops gradually, more like the other known chemical databases. Nearly 3/4 of the scaffolds of toxic substances have two or less rings.

In Figure 4, the populations of scaffolds 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 scaffolds with 4-nodes and hence 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 7, the fractions of the topologies present versus the theoretical possibilities are tabulated as a function of the number of rings, while in Table 8, the fractions for $r = 1-6$, categorized by $N_3$ and $N_4$, are displayed. Note that a blank

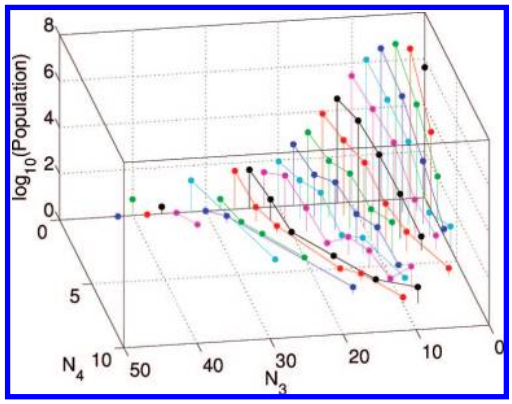



**Figure 4.** Populations of scaffolds in the ChemNavigator database as a function of the number of 3- and 4-nodes, $N_3$ and $N_4$, and ordered, using connected stems of the same color, by the number of independent rings $r$. Five outliers (scaffolds with $N_3 > 50$) have been excluded to make the main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

**Table 7.** The Fractions of Scaffold Topologies in the Indicated Databases with Respect to the Theoretical Maxima Per Number of Rings $r$

| $r =$ | 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 |
| DSSTox | 1.000 | 0.667 | 0.667 | 0.315 | 0.061 | 0.002 | 0.000 | 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 |

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 except DSSTox and were generally 1.0 for $r = 3$, the exceptions being Drugs, DSSTox, and WOMBAT, all smaller databases. For $r \geq 4$, the tendency toward 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 and DSSTox collections.

Considering the four-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 nonplanar graphs are extremely rare; the first known example of a molecule with this topology was synthesized by Walba.[38] 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 contain some fraction of the possibilities. The generated structures of GDB enumerate only 40−50% of the various four-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 nonplanar graphs and all three- and four-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, is an indicator of the diversity of a database. Another is the population fraction of each distinct topology within the database. Table 9 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 three-ring scaffolds containing only 3-nodes, there are significant percentages of structures containing one to five rings, and zero rings in some cases such as for DNP, Drugs, and DSSTox.

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 seven 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. The 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). Among the remaining topology types, those that consist of three or more rings emanating from a central vertex or vertices (i.e., 9 and 31−33) 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 has a considerable percentage of structures with no rings. DSSTox, as noted earlier, has a preponderance of no-ring and single-ring structures, and no examples at all of any 4-node-only topologies and very few with any 4-nodes at all.

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 two- and three-ring topologies, numbers 3 and 5, respectively, more than the other databases.

Table 10 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; they 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 seven databases of known chemicals, making up from 62.8 to 91.3% of the total populations. None of these topologies possess 4-nodes. There is some tendency for DNP to have more and DSSTox to have fewer scaffolds with linear assemblies of fused rings than the other databases (see Tables 10 and 11). In general, the biologically oriented databases, except DSSTox, have greater percentages within their top 10 topologies exhibiting linear fused-ring assemblies than

**Table 8.** 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[a]

| *r* | $N_4$ | $N_3$ | ChemNav. | DNP | Drugs | PubChem | PC actives | DSSTox | WOMBAT | *merged* | *GDB* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1.000 | 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 | 1.000 |
| | 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 | 1.000 |
| | 1 | 2 | 1.000 | 1.000 | 0.800 | 1.000 | 1.000 | 0.600 | 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 | 0.824 | 0.941 | 0.941 | 0.412 |
| | 1 | 4 | 1.000 | 0.900 | 0.467 | 1.000 | 0.900 | 0.300 | 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 | 0.394 | 0.831 | 0.944 | 0.127 |
| | 1 | 6 | 0.845 | 0.554 | 0.057 | 0.974 | 0.399 | 0.036 | 0.482 | 0.974 | 0.052 |
| | 2 | 4 | 0.364 | 0.303 | 0.004 | 0.868 | 0.127 | 0.004 | 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 | 0.031 | 0.482 | 0.851 | 0.008 |
| | 1 | 8 | 0.303 | 0.122 | 0.006 | 0.596 | 0.057 | 0.003 | 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 | |

[a] Blank entries indicate that no representatives of that class of topologies were found in the specified database.

**Table 9.** 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[a]

| *r* | $N_4$ | $N_3$ | ChemNav. | DNP | Drugs | PubChem | PC actives | DSSTox | WOMBAT | *merged* | *GDB* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0.245 | 8.633 | 6.492 | 2.466 | 3.837 | 25.057 | 1.641 | 1.225 | 15.414 |
| 1 | 0 | 0 | 2.979 | 11.831 | 16.630 | 8.212 | 10.771 | 29.808 | 6.588 | 5.248 | 41.721 |
| 2 | 0 | 2 | 20.808 | 15.390 | 25.492 | 24.094 | 18.384 | 19.515 | 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 | 12.746 | 28.190 | 34.064 | 7.299 |
| | 1 | 2 | 0.287 | 1.172 | 0.547 | 0.523 | 0.664 | 0.179 | 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 | 7.612 | 25.496 | 23.463 | 1.008 |
| | 1 | 4 | 0.729 | 2.829 | 1.349 | 0.931 | 2.132 | 0.664 | 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 | 2.095 | 11.800 | 8.492 | 0.057 |
| | 1 | 6 | 0.554 | 2.115 | 0.839 | 0.682 | 1.103 | 0.383 | 0.971 | 0.630 | 0.010 |
| | 2 | 4 | 0.028 | 1.097 | 0.036 | 0.064 | 0.180 | 0.026 | 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 | 0.741 | 3.524 | 2.071 | 0.001 |
| | 1 | 8 | 0.238 | 1.808 | 0.511 | 0.356 | 0.651 | 0.128 | 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 | |

[a] 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.

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, two of which are spiro. There is also a tendency toward linear assemblies of fused rings in this database (mostly due to the topology in Figure 6a ranked **10**); however, note that two of GDB's most frequent scaffold topologies (ranked **46** and **122** in Figure 6a) are nonlinear clusters of fused rings, which are rare in the other databases.

If the 32 most frequent scaffolds and the acyclic compounds found in Bemis and Murcko's analysis of the Comprehensive Medicinal Chemistry database[40] are converted to topologies, we find the following frequencies > 1%, where the boldfaced numbers indicate the rank in our merged database:

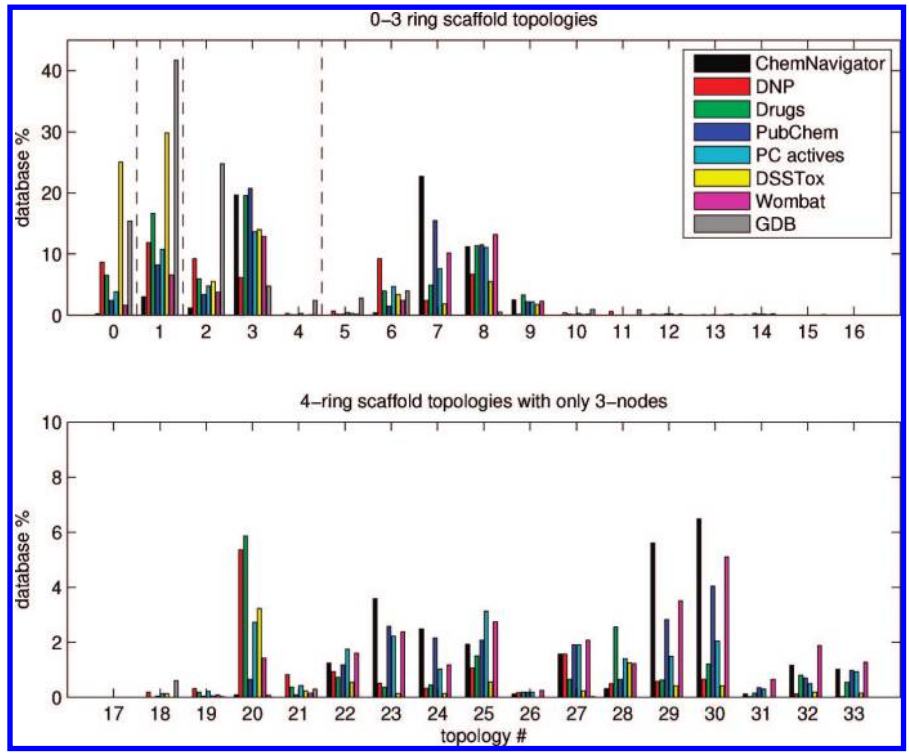



**Figure 5.** The percentage frequencies of the first 33 scaffold topologies of Figure 2 in the indicated databases. The entry labeled zero indicates the database percentages of structures that do not contain rings. The dashed lines in the top graph divide the results into sets of topologies possessing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

**Table 10.** The Percentages of the 10 Most Frequent Topologies Present in Each of the Databases Examined[a]

| Chem Navigator | | DNP | | Drugs | | PubChem | | PC actives | | DSSTox | | WOMBAT | | *GDB* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 22.694 | **4** | 11.831 | **1** | 19.548 | **1** | 20.740 | **1** | 13.642 | **4** | 29.808 | **3** | 13.200 | **10** | 41.721 |
| **1** | 19.646 | **10** | 9.249 | **4** | 16.630 | **2** | 15.457 | **3** | 11.101 | **14** | 25.057 | **1** | 12.901 | **14** | 24.765 |
| **3** | 11.196 | **18** | 9.226 | **3** | 11.379 | **3** | 11.509 | **4** | 10.771 | **1** | 13.997 | **2** | 10.160 | **1** | 15.414 |
| **5** | 6.474 | **14** | 8.633 | **14** | 6.492 | **4** | 8.212 | **2** | 7.652 | **10** | 5.517 | **4** | 6.588 | **4** | 4.755 |
| **6** | 5.609 | **3** | 6.643 | **10** | 5.945 | **5** | 4.033 | **10** | 4.743 | **3** | 5.492 | **5** | 5.101 | **18** | 3.953 |
| **7** | 3.590 | **1** | 6.140 | **26** | 5.872 | **10** | 3.354 | **18** | 4.681 | **18** | 3.372 | **10** | 3.779 | **46** | 2.765 |
| **4** | 2.979 | **26** | 5.356 | **2** | 4.887 | **6** | 2.824 | **14** | 3.837 | **26** | 3.218 | **6** | 3.510 | **57** | 2.425 |
| **8** | 2.505 | **48** | 2.872 | **18** | 3.939 | **7** | 2.573 | **11** | 3.130 | **2** | 1.865 | **11** | 2.741 | **58** | 0.977 |
| **9** | 2.486 | **2** | 2.437 | **8** | 3.319 | **14** | 2.466 | **26** | 2.721 | **8** | 1.737 | **18** | 2.399 | **114** | 0.910 |
| **13** | 2.094 | **37** | 1.625 | **23** | 2.553 | **8** | 2.204 | **7** | 2.220 | **23** | 1.252 | **7** | 2.375 | **122** | 0.610 |
| | 79.273 | | 64.012 | | 80.564 | | 73.372 | | 64.498 | | 91.315 | | 62.754 | | 98.295 |

[a] The numbers in boldface refer to the rank in the merged database; the corresponding scaffold topologies 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.

**1**. 16.582, **4**. 14.355, **14**. 5.977, **10**. 5.527, **26**. 4.824, **3**. 4.336, **18**. 2.812

These values are remarkably similar to the results for Drugs in Table 10. Note that a substantial fraction (44.26%) of Bemis and Murcko's data (of less-frequent scaffolds) was not published. Only topology **3** has a significantly different placement in the two orderings.

The total number of scaffold topologies containing eight rings or less is 1 547 689 (see Table 1). Of these, 850 878 (54.98%) contain spiro nodes, and 164 375 (10.62%) are nonplanar as determined by nauty.[41] There are 9474 topologies in the merged database with eight or less rings, so 99.39% of the possible scaffold topologies are not found in any of the databases examined. Of those missing, 51.58% are planar and have spiro nodes, 3.60% are nonplanar with spiro nodes, and 7.09% are nonplanar and lack spiro nodes. Only 12 nonplanar and 2099 spiro node topologies (all of which are planar) are present in the merged database. Nine of the nonplanar topologies are found only in PubChem, and the total number of molecules represented by such topologies in the merged database is a mere 44, agreeing with Walba's assessment[38] concerning the rarity of chemicals with nonplanar graphs. Of the databases that have topologies unique to them for $r \leq 8$, the only biologically oriented ones are DNP and WOMBAT, with just a few examples (372 and 49 molecules, respectively, representing about half as many topologies), while 55.48% of PubChem's $r \leq 8$ topologies (4959/8939) are present only there.

We computed the scaffold-to-SMILES ratios of the various known chemical databases for the 17 scaffold topologies that

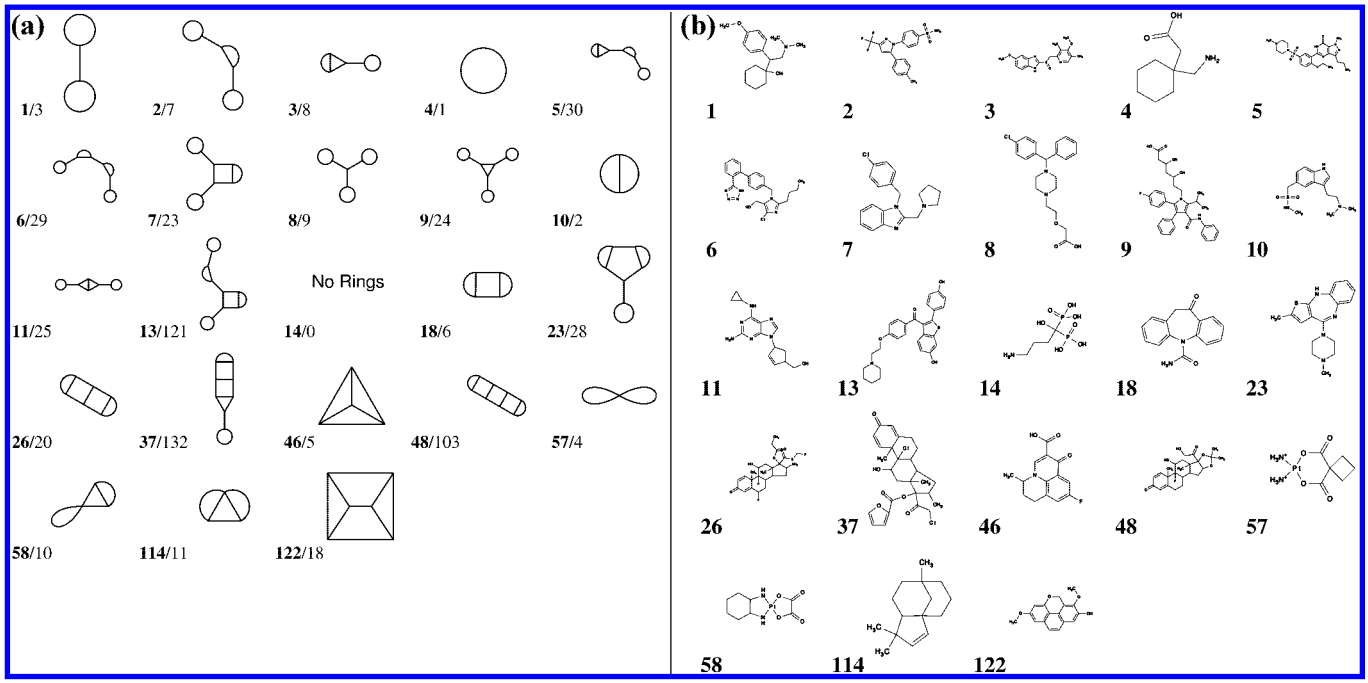



**Figure 6.** (a) The most frequent topologies present in the databases examined, numbered (in boldface) by their rank in the merged database. The second value for each entry is the topology number, 1−33 and 86−89 of which are shown in Figure 2a. (b) Examples[39] from the databases examined of the most frequent topologies present, numbered by their rank in the merged database (compare with Figure 6a).

**Table 11.** 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[a]

| Chem Nav. | | DNP | | Drugs | Pub Chem | PC actives | DSSTox | WOMBAT | *merged* | *GDB* |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | 1 1 | 2 1 | 2 1 | 2 1 | 2 1 | 1 1 | 3 2 | 2 1 | 2 2 |
| 2 1 | | 2 2 | 1 1 | 3 1 | 3 2 | 0 0 | 3 1 | 3 1 | 0 0 | |
| 3 2 | | 3 3 | 3 2 | 3 2 | 1 1 | 2 1 | 3 1 | 3 2 | 2 1 | |
| 4 2 | | 0 0 | 0 0 | 1 1 | 3 1 | 2 2 | 1 1 | 1 1 | 1 1 | |
| 4 1 | | 3 2 | 2 2 | 4 2 | 2 2 | 3 2 | 4 2 | 4 2 | 3 3 | |
| 4 2 | | 2 1 | 4 4 | 2 2 | 3 3 | 3 3 | 2 2 | 4 1 | 3 3* | |
| 1 1 | | 4 4 | 3 1 | 4 1 | 0 0 | 4 4 | 4 1 | 4 2 | 2 1 | |
| 3 1 | | 5 5 | 3 3 | 4 2 | 4 2 | 3 1 | 4 2 | 3 1 | 3 2 | |
| 4 1 | | 3 1 | 3 1 | 0 0 | 4 4 | 3 1 | 3 3 | 4 1 | 3 3 | |
| 5 2 | | 5 4 | 4 3 | 3 1 | 4 2 | 4 3 | 4 2 | 2 2 | 4 4* | |

[a] Nonlinear assemblies of fused rings are marked by an asterisk.

are common to the corresponding 10 most frequent topology collections in Table 10 (topologies ranked **1**−**11**, **13**, **18**, **23**, **26**, **37**, and **48** in Figure 6a), comprising at least 55% of the population of each of the databases. The average numerical rank (1−8) of the ratios taken from highest to lowest

| Drugs | DSSTox | PC actives | DNP | WOMBAT |
|---|---|---|---|---|
| 1.412 | 1.765 | 2.882 | 4.412 | 4.706 |

| PubChem | ChemNavigator | merged |
|---|---|---|
| 6.706 | 6.824 | 7.294 |

follow exactly the order of the database sizes from smallest to largest, reinforcing the observation for Table 4 that the size of the database has a significant influence on the observed ratio.

For the same set of databases and scaffold topologies, the average number of atoms per scaffold that make up each topology class is graphed in Figure 7. The two general databases, ChemNavigator and PubChem, have been omitted as they follow very similar trends to the merged database. The black bars indicate the number of atoms necessary to produce minimal scaffolds (a minimal loop is defined by three atoms), and the ratio of the merged averages to these is nearly constant, approximately 2.33, due in large part to the wealth of six-membered rings throughout chemistry (note topology **4**). (We note that the minimal scaffold is achieved in the merged database for eight of the topologies, typically the smaller ones.) The anomalous jump at **9** for Drugs is derived from only 12 examples, one of which is the 128-atom scaffold of nesiritide. Omitting this outlier brings the mean down to 31.82. Topologies ranked **6**−**9** and **23** exhibit the most variability (three- and four-ring structures with one to three dangling rings). Generally, DNP scaffolds have the most and DSSTox scaffolds the fewest atoms per topology class, although there are some exceptions.

## 4. CONCLUSIONS

We report the scaffold distribution and topological properties for seven databases of existing chemicals: ChemNavigator, DNP, Drugs, PubChem, PubChem "actives", DSSTox, 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. The databases analyzed in this paper are already dated, but updating the values will not change the qualitative aspect of our results. We will provide semiannual updates for some of these tables on our UNM Biocomputing Web site. For six-ring scaffolds, PubChem molecules cover less than a third of the possible theoretical topological space (limited to ≤ 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

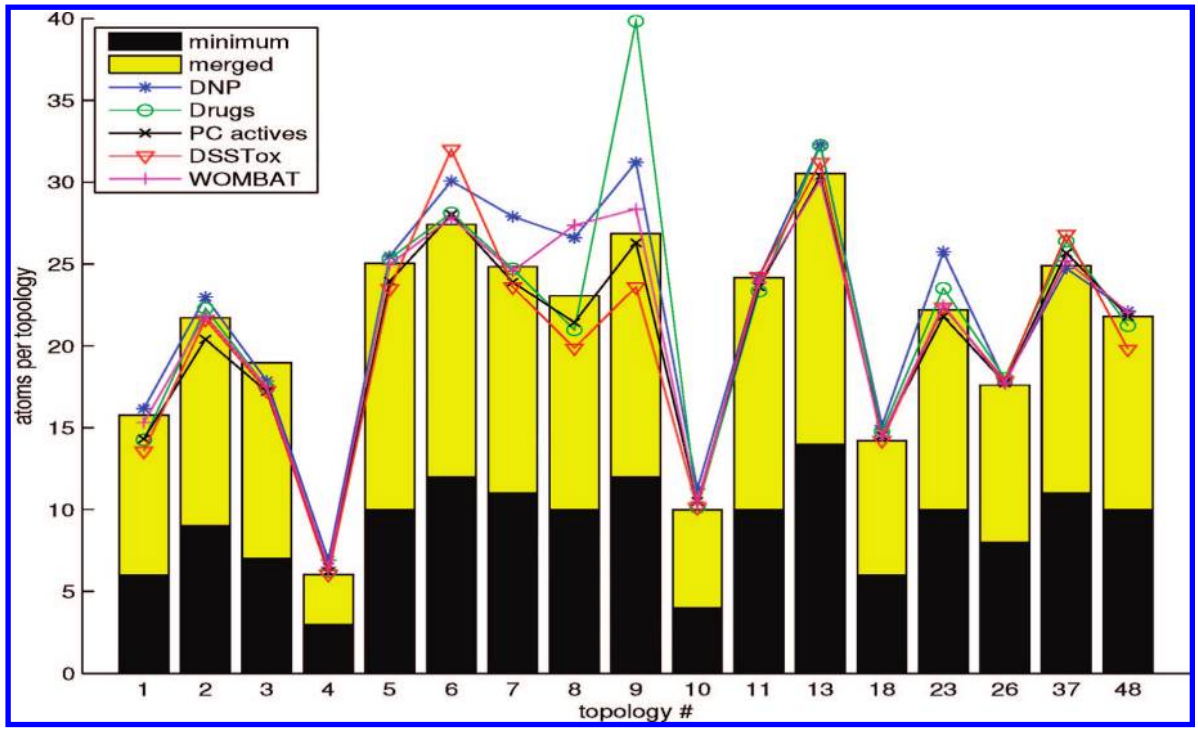



**Figure 7.** The average number of atoms comprising the scaffolds in the indicated databases that are members of the given ranked topologies (see Figure 6a). Minimum refers to the number of nodes needed to produce a minimal representative of the topology (see Figure 1d). The values for the merged database are the total bar heights.

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 1.45 billion unique SMILES—all C.sp3-based, and all single bonds, up to eight rings and 20 atoms.[8–11] We abandoned that strategy because this approach would quickly reach the asymptotic wall of combinatorial explosion: consider that, corresponding to the 1.45 billion alkanes, there are probably 1 billion monoalkenes, monoamines, and monoalcohols, to name a few possibilities while approximating for symmetry-related redundancy. The GENSMI algorithm became increasingly tedious to use at higher levels of complexity. 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.

Fine-grained enumerations of the CSSM do provide potential organic molecules from which a variety of chemical, geometrical, and topological properties can be extracted, as well as possible drug leads and so forth. Coarse-grained approaches like ours sacrifice details such as atom and bond types in the interest of restraining the inevitable combinatorial explosion, allowing for a much broader but shallower perspective, which restricts itself to topological properties. Even coarser-grained explorations can be performed, such as the one by Lipkus,[42] which classified the CSSM with a trio of topological descriptors. This work was performed before complete enumerations were available, so comparisons with the theoretical possibilities were limited.

The granularity of scaffold topological enumeration has an important feature when applied to real chemical databases. Lightly populated regions of structures rich in complexity, where the combinatorics make it infeasible to perform fine-grained enumeration, are well broken apart by our classification. Alternatively, heavily populated regions of simple topologies, where the combinatorics are much easier, are well-suited for complete fine-grained subclassifications, and so the two levels of granularity are actually complementary. Scaffold topologies can be viewed as a low-resolution atlas of the major topological classes of organic ring systems ($r \leq 8$), while fine-grained enumerations act as detailed roadmaps of particular regions.

In our analyses, we found a strong bias in all collections of existing chemical compounds (especially DSSTox, which is nearly devoid of 4-nodes) toward 3-node topologies, that is, vertices branching out in three different directions (see Tables 8 and 9). 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. Indeed, we see a modest tendency toward more linear fused-ring assemblies in the biologically oriented databases (especially DNP), except for DSSTox, which is underrepresented by these structures. There is also a tendency toward fewer overall rings in DNP, Drugs, and especially DSSTox, all of which also have significant fractions of molecules that do not contain any rings at all. Finally, we note that compounds possessing nonplanar graphs are quite rare.

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. The scaffolds of natural products generally have more atoms than average, however.

Looking at the 10 most frequent topologies for each database, we find that a small number of topologies characterize most of the molecules. Only eight topologies (**1**−**5**, **10**, **14**, and **18** in Figure 6) are needed to characterize half the population of the each of the eight databases. A total of 62.8−91.3% 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. This is consistent with the findings of other researchers in the context of scaffolds.[33,40] Only 0.61% of the possible scaffold topologies containing eight rings or less have actual chemical representatives. As has also been seen by others,[10,12,13] the CSSM is vast and almost completely unexplored. The various databases examined, especially the biologically oriented ones, occupy very restricted regions.

We have developed a Web site[43] 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. The output includes 2D diagrams of the original molecule and a minimal representative of the scaffold topology, some numerical details related to the topology, the number of matches of this topology in the public database PubChem, and some examples of this topology from PubChem. The SMILES of all molecules possessing this topology can also be extracted from the database.[44] 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 four rings are available. We will continue to extend the capabilities of this site and provide updates of scaffold topology distributions for a number of databases.

To generate a scaffold topology, we effectively collapse a molecular structure to its essential ring and connecting linear structure. In the companion paper of Pollock et al.,[1] scaffold topologies are systematically built up from the most basic topologies of one and two rings, and then they are uniquely characterized. Once a topology is available, a minimal or more complicated scaffold can be produced. The two papers, therefore, look at the problem of CSSM exploration from the opposing points of view of what is possible and what actually occurs.

The unique characterization of scaffold topologies makes it possible to create an efficient, searchable database that allows for rapid coarse-grained classification of organic molecules. For example, to analyze the scaffold topologies for the approximately 25 million unique SMILES in the merged database required less than 4 CPU-hours on a 2.2 GHz Linux system with 32 GB of RAM. Such population-based topological analyses can easily be performed using this categorization technique, so this methodology complements existing techniques for CSSM mapping.

## ACKNOWLEDGMENT

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

## REFERENCES AND NOTES

(1) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. Scaffold Topologies. 1. Exhaustive Enumeration up to Eight Rings. *J. Chem. Info. Model.* **2008**, *48*, 1304–1310.

(2) 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.

(3) de Laet, A.; Hehenkamp, J. J. J.; Wife, R. L. Finding Drug Candidates in Virtual and Lost/Emerging Chemistry. *J. Heterocycl. Chem.* **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. Chemography: The Art of Chemical Space Navigation. *J. Comb. Chem.* **2001**, *3*, 157–166.

(6) Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **2002**, *6*, 384–389.

(7) Molecular Libraries and Imaging. http://nihroadmap.nih.gov/molecularlibraries/ (accessed May 2008).

(8) Kappler, M. A.; Allu, T. K.; Oprea, T. I. GENSMI: Generation of Genuine SMILES, Presented at MUG'04: 18th Daylight User Group Meeting, 2004. http://www.daylight.com/meetings/mug04/Kappler/GenSmi.html (accessed Dec 7, 2007).

(9) Kappler, M. A. GENSMI: Exhaustive Enumeration of Simple Graphs, Presented at EuroMUG 2004. http://www.daylight.com/meetings/emug04/Kappler/GenSmi.html (accessed Dec 7, 2007).

(10) Kappler, M. A. GENSMI: Exhaustive Enumeration of Simple Graphs, Presented at Biocomputing @ UNM 2005. http://biocomp.health.unm.edu/events/Biocomputing@UNM2005/Presentations/Kappler/GenSmi.html (accessed Dec 7, 2007).

(11) Oprea, T. I.; Kappler, M. A.; Allu, T. K.; Mracec, M.; Olah, M. M.; Rad, R.; Ostopovici, L.; Hadaruga, N.; Baroni, M.; Zamora, I.; Berellini, G.; Aristei, Y.; Cruciani, G.; Bologa, C. G.; Edwards, B. S.; Sklar, L. A.; Balakin, K. V.; Savchuk, N.; Brown, D.; Larson, R. S. *QSAR and Molecular Modelling in Rational Design of Bioactive Molecules*; Computer Aided Drug Design & Development Society in Turkey: Istanbul, Turkey, 2006.

(12) Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. Molecules in silico: potential versus known organic compounds. *MATCH* **2005**, *54*, 301–312.

(13) Fink, T.; Reymond, J.-L. 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. *J. Chem. Inf. Model.* **2007**, *47*, 342–353.

(14) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.

(15) 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.

(16) 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.

(17) Filip, P. A.; Balaban, T.-S.; Balaban, A. T. A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlation ability. *J. Math. Chem.* **1987**, *1*, 61–83.

(18) Mekenyan, O.; Bonchev, D.; Balaban, A. Topological indices for molecular fragments and new graph invariants. *J. Math. Chem.* **1988**, *2*, 347–375.

(19) Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *J. Math. Chem.* **1993**, *12*, 309–318.

(20) Berger, F.; Flamm, C.; Gleiss, P. M.; Leydold, J.; Stadler, P. F. Counterexamples in Chemical Ring Perception. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 323–331.

(21) 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.

(22) Moss, G. P. Extension and revision of the nomenclature for spiro compounds (IUPAC Recommendations 1999). *Pure Appl. Chem.* **1999**, *71*, 531–558.

(23) iResearch Library, ChemNavigator.com, Inc., 2006. http://www.chemnavigator.com/ (accessed Dec 7, 2007).
(24) Dictionary of Natural Products, version 14.1; Chapman & Hall/CRC: London, 2006.
(25) PubChem, National Center for Biotechnology Information, 2006. http://pubchem.ncbi.nlm.nih.gov/ (accessed Dec 7, 2007).
(26) U.S. Environmental Protection Agency, Distributed Structure-Searchable Toxicity (DSSTox), 2007. http://epa.gov/ncct/dsstox/ (accessed Dec 7, 2007).
(27) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Schreiber, S. L., Kapoor, T. M., Wess, G. Eds.; Wiley-VCH: New York, 2007.
(28) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(29) Daylight Theory Manual; Daylight Chemical Information Systems, Inc.: Aliso Viejo, California, 2007. http://www.daylight.com/dayhtml/doc/theory/ (accessed Dec 7, 2007).
(30) OEChem — C++ Theory Manual, Version 1.4, OpenEye Scientific Software, Inc., 2006. http://www.eyesopen.com/docs/ (accessed Dec 7, 2007).
(31) 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.
(32) Reymond, J.-L. Reymond Group Cheminformatics Site, 2007. http://www.dcb.unibe.ch/groups/reymond/cheminf/index.html (accessed Dec 7, 2007).
(33) Xue, L.; Bajorath, J. Distribution of Molecular Scaffolds and R-Groups Isolated from Large Compound Databases. *J. Mol. Model.* **1999**, *5*, 97–102.
(34) Lewell, X. Q.; Jones, A. C.; Bruce, C. L.; Harper, G.; Jones, M. M.; Mclay, I. M.; Bradshaw, J. Drug Rings Database with Web Interface. A Tool for Identifying Alternative Chemical Rings in Lead Discovery Programs. *J. Med. Chem.* **2003**, *46*, 3257–3274.
(35) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17272–17277.
(36) Bone, R. G. A.; Villar, H. O. Exhaustive Enumeration of Molecular Substructures. *J. Comput. Chem.* **1997**, *18*, 86–107.
(37) 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.
(38) Walba, D. M. Topological Stereochemistry. *Tetrahedron* **1985**, *41*, 3161–3212.
(39) 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.
(40) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(41) McKay, B. D. *nauty User's Guide*, version 2.4; Australian National University: Canberra, Australia, 2007.
(42) Lipkus, A. H. Exploring Chemical Rings in a Simple Topological-Descriptor Space. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 430–438.
(43) Molecular Scaffold Topology Server. http://topology.health.unm.edu/ (accessed May 2008).
(44) These analyses were performed on unique parent compound entries extracted from chemical databases, in which all salts were removed and then nonunique entries eliminated. There will actually be more entries in these databases for any given topology, generally, than the numbers reported here, but if only unique SMILES are considered, then these numbers should be identical.

CI700342H