JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

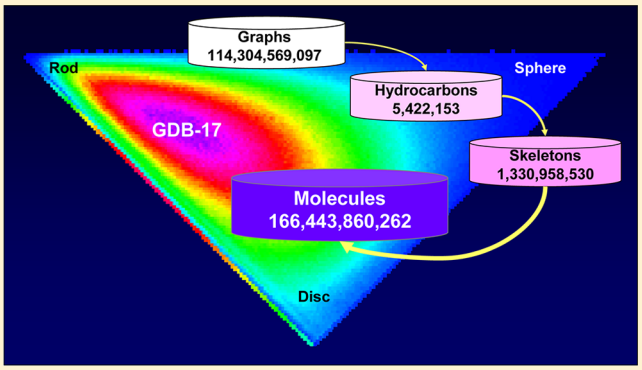
# Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17

Lars Ruddigkeit,[†] Ruud van Deursen,[‡] Lorenz C. Blum,[†] and Jean-Louis Reymond*[,†]

[†]Department of Chemistry and Biochemistry, NCCR TransCure, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland

[‡]Biomolecular Screening Facility, NCCR Chemical Biology, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

**ABSTRACT:** Drug molecules consist of a few tens of atoms connected by covalent bonds. How many such molecules are possible in total and what is their structure? This question is of pressing interest in medicinal chemistry to help solve the problems of drug potency, selectivity, and toxicity and reduce attrition rates by pointing to new molecular series. To better define the unknown chemical space, we have enumerated 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens forming the chemical universe database GDB-17, covering a size range containing many drugs and typical for lead compounds. GDB-17 contains millions of isomers of known drugs, including analogs with high shape similarity to the parent drug. Compared to known molecules in PubChem, GDB-17 molecules are much richer in nonaromatic heterocycles, quaternary centers, and stereoisomers, densely populate the third dimension in shape space, and represent many more scaffold types.






## INTRODUCTION

The cumulated efforts of synthetic chemistry over the last century has produced over 60 million compounds as collected by Chemical Abstracts Service.[1,2] Since the implementation of combinatorial and parallel synthesis by academic and industrial drug discovery, the number of druglike small molecules (organic compounds of intermediate polarity with MW ≤ 500 Da) has increased even further.[3−5] The combined corporate, academic, and commercial collections worldwide probably total over 100 million different small molecules.[6] Despite these impressive numbers, it has become increasingly difficult to develop new small molecule drugs, largely due to lack of efficacy, side effects, and toxicity issues.[7,8] *De novo* drug design[9−12] may help to address this problem by investigating even much larger numbers of yet unknown molecules by virtual screening[13−15] in search of innovative structures that might exhibit improved selectivity and ADMET profiles.

The majority of *de novo* drug design methods generate molecules within genetic algorithms that optimize a desired property such as a docking score by evolving a molecule population through breeding and mutation cycles. In most cases these algorithms generate new molecules by recombining known building blocks with known reactions, which severely limits their innovative potential. To circumvent this limitation, we recently approached the direct enumeration of chemical space by extending an approach to *de novo* design pioneered by Cayley, the inventor of graph theory, to count acyclic hydrocarbons[16] and later used in computer assisted structure elucidation.[17−19] The idea is to enumerate molecules from first principles starting from mathematical graphs irrespective of pre-existing building blocks to avoid a historical bias in structure selection. Geometrical strain and functional group stability criteria are used to ensure that the molecules produced are chemically meaningful. By this method we obtained the chemical universe database GDB-11 enumerating 26.4 million different molecules up to 11 atoms of C, N, O, and F (110.9 million molecules when including stereoisomers).[20,21] The number increased to almost 1 billion (not counting stereoisomers) for GDB-13 listing all molecules up to 13 atoms of C, N, O, Cl, and S.[22,23] Both databases were later shown to be useful sources of molecular diversity to discover new receptor ligands by virtual screening, synthesis, and testing.[24−30]

While GDB-11 and GDB-13 uncovered impressive numbers of possible molecules, the databases only addressed very small organic molecules (MW < 200 Da), which are of interest as relatively small fragments[31] but rarely correspond to actual drugs. Herein we report the enumeration of organic molecules up to 17 atoms of C, N, O, S, and halogens, forming the chemical universe database GDB-17 containing 166.4 billion organic molecules. GDB-17 reaches into molecular sizes compatible with many drugs (367 approved drugs ≤17 atoms) and typical for lead compounds (100 < MW < 350 Da).[32] Millions of isomers of known drugs are readily identified in GDB-17. While molecules

**Received:** August 31, 2012
**Published:** October 22, 2012

   dx.doi.org/10.1021/ci300415d | *J. Chem. Inf. Model.* 2012, 52, 2864−2875

up to 17 atoms in the public databases PubChem,[33] ChEMBL,[34] or DrugBrank[35] are mostly achiral, aromatic, and heteroaromatic compounds with rodlike shapes, GDB-17 molecules are mostly nonaromatic heterocycles with many quaternary centers and stereoisomers. GDB-17 densely populates the third dimension in shape space and represents many more scaffold types than found in PubChem.

## RESULTS AND DISCUSSION

**Enumeration.** The enumeration followed the approach used for GDB-13 starting from the complete list of graphs as given by the program GENG.[36] Graphs corresponding to unstrained hydrocarbons were selected based on geometrical criteria and expanded to ″skeletons″ (unsaturated hydrocarbons) by substituting bonds (single, double, triple bonds) for graph edges. These skeletons were themselves expanded to molecules by substituting atoms (C, N, O, etc.) for graph nodes, respecting valency rules, and eliminating chemically unstable and problematic functional groups (Figure 1). The code used for

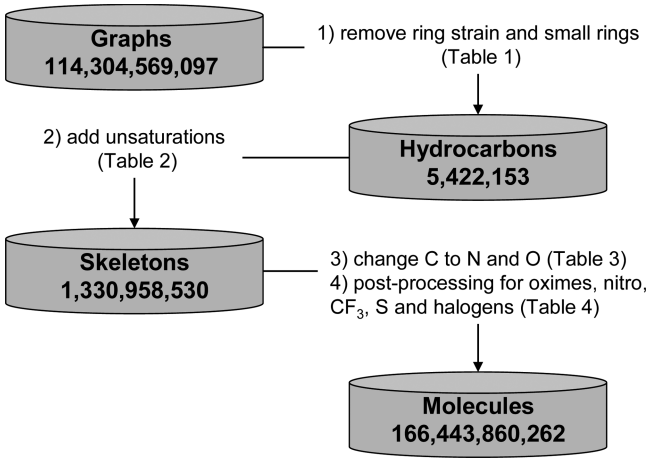



**Figure 1.** Enumeration of GDB-17 starting from mathematical graphs.

GDB-13, which stalled at 14 or more atoms due to inefficient memory usage, was entirely rewritten considering process design and efficiency. New graph and molecule selection criteria were also added to constrain the combinatorial explosion above 13 atoms. The code redesign resulted in a 400-fold increase in computing speed allowing to complete the enumeration up to 17 atoms of C, N, O, S, and halogens within reasonable time, as detailed below.

GENG[36] was set to enumerate all graphs up to 17 nodes with maximum valency of four (quaternary carbon) considering only topologically planar (i.e., eliminating knotted topologies corresponding to the $K_5$ and $K_{3,3}$ graphs), connected (e.g., no catenanes) graphs, which returned 114,304,569,097 graphs. The graphs were then converted to hydrocarbons substituting carbon atoms for graph nodes and carbon−carbon single bonds for graph edges. Hydrocarbons were selected for limited ring strain and topological complexity by applying the hydrocarbon filters H1 to H5 (Table 1), which left 5,422,153 hydrocarbons (0.005% of the graphs) to be considered for molecule generation. These 5.4 million hydrocarbons were converted to ″skeletons″ by introducing double and triple bonds following valency rules and the unsaturation filters S1 to S5, which primarily restrict ring strain and reactive unsaturations (Table 2). The selected skeletons were then run through an aromatization−dearomatization cycle, and duplicates were removed, which eliminated aromatic tautomers. The introduction of unsaturations generated on average 246 skeletons per graphs for a total of 1,330,958,530 skeletons.

The 1.3 billion skeletons were diversified into molecules by combinatorially substituting nitrogen and oxygen for carbon following valency rules but not generating any heteroatom−heteroatom bond. All generated molecules were then checked for undesirable functional groups (FG) to ensure that the molecules have a good probability to be stable and synthetically accessible (FG filters F1−F12, Table 3). These filters followed in part previously reported criteria for removing problematic functional groups from screening libraries.[37,38] This diversification of skeletons into ″CNO″ molecules produced 110.4 billion molecules, corresponding to an average of 83 molecules per skeleton and 20,400 molecules per graph. Postprocessing steps P1−P7 were finally implemented for additional diversity by combinatorial atom type substitutions (Table 4). These postprocessing steps added another 56 billion molecules, resulting in a total of 166.4 billion molecules in the complete GDB-17. The overall molecule generation procedure was always completed in one run for each graph, and all molecules generated from each graph were checked for duplicates to guarantee that all GDB-17 molecules are different (Table 5). The overall computation consumed 100,000 CPU hours, which is only 2.5-fold more than the computing time originally invested for GDB-13.

**Comparing GDB-17 with PubChem, ChEMBL, and DrugBank.** One of the key questions arising from the systematically enumerated chemical space available in GDB-17 is whether this collection significantly differs from the already known chemical space. To perform this comparison, we collected molecules up

**Table 1. Ring Strain and Complexity Filters To Select Hydrocarbon Graphs**

| filter | description | comment |
|---|---|---|
| H1 | SAV (″smallest atomic volume″): all graphs ≤ $C_{11}$ are converted to a 3D-structure, and the volume of the tetrahedron around each C atom is checked for a minimum value. | 95.2% of graphs ≤ $C_{11}$ are discarded due to failed 3D-conversion (using CORINA or ChemAxon molconverter) or due to distorted (planar, pyramidal) centers. Simple polyhedra such as cubane are preserved. See the SI of ref 22 for details. |
| H2 | NA2SR (″no atom shared by two small rings″): removes graphs ≥ $C_{12}$ containing fused or spiro linkages between 3- or 4-membered rings. | The vast majority of fused small ring systems are highly strained and reactive. 96.7% of $C_{12}$- and 97.3% of $C_{13}$-graphs are removed. See also filter H3. |
| H3 | NBH3R (″no bridgehead in 3 rings″): Graphs ≥ $C_{14}$ with three or more ″nonzero″ bridgehead atoms shared by three or more rings are removed. | These multilooped topologies would correspond to molecules of high synthetic complexity. 99.60% of $C_{14}$-, 99.77% of $C_{15}$-, 99.95% of $C_{16}$-, and 99.95% of $C_{17}$-graphs are removed by filters H2+H3. |
| H4 | 1SR (″one small ring″): $C_{15}$ and $C_{16}$ graphs are allowed at most one 3- or 4-membered ring. | 71.89% of the $C_{15}$- and 79.50% the $C_{16}$-graphs that passed H2+H3 are removed. |
| H5 | 0SR (″no small ring″): $C_{17}$-graphs are not allowed any small rings. | 96.95% $C_{17}$-graphs that passed H2+H3 are removed. |


dx.doi.org/10.1021/ci300415d | J. Chem. Inf. Model. 2012, 52, 2864−2875

**Table 2. Unsaturation Filters To Enumerate Skeletons (Unsaturated Hydrocarbons)**

| filter | description | comment |
|---|---|---|
| S1 | no allenes (C=C=C) | Although known and sometime found in bioactive molecules, allenes are usually reactive and quite difficult to prepare but combinatorially extremely frequent. |
| S2 | no unsaturations in 3-membered rings | Cyclopropenes are known but quite reactive and difficult to prepare. Cyclopropynes are unstable. |
| S3 | at most one $sp^2$-center in 4-membered rings | Cyclobutenes and cyclobutynes are not enumerated, but the skeletons leading to $\beta$-lactams and $\beta$-lactones are generated. |
| S4 | triple bonds restrictions | No triple bond in 3- or 4-membered rings, max. One triple bond in rings ≥9 and max two triple bonds in ≥11 rings. Only terminal triple bonds for $C_{17}$-hydrocarbons (allowing to generate nitriles). |
| S5 | bridgehead double bond restrictions | If a "non-zero" bridgehead carbon is $sp^2$, the ring sizes of the smallest set of smallest rings will be checked. At least one ring of this carbon must be ≥8. In case of two such bridgeheads are $sp^2$, the ring size must be ≥10. |

**Table 3. Functional Group Filters To Enumerate CNO Molecules[a]**

| filter | description | comment |
|---|---|---|
| F1 | XCX: only one N or O next to a $sp^3$ carbon or two oxygens if both oxygens are ring atoms. | Aminals, hemiacetals *gem*-diols, and acyclic acetals are not enumerated. Only cyclic acetals are allowed. |
| F2 | Maximum one N or O in small rings. | Allows epoxides, oxiranes, aziridines, azetidines but no cyclic acetals inside 4-membered rings. |
| F3 | Anhydrides (O=C)−O−(C=O) are removed. | Most anhydrides are unstable toward hydrolysis. |
| F4 | Acetal chains O−$Csp^3$−O−$Csp^3$−O are removed. | Although sometimes found, acetal chains are diffult to plan synthetically. |
| F5 | Molecules with a primary amine and a ketone or aldehyde are removed. | This combination often polymerizes. |
| F6 | C=N are removed unless the $sp^2$ carbon is connected to a further N or O atom. | Removes imines which are unstable but retains amidines and guanidines. |
| F7 | (N/O)−C=N−C=(N/O) are removed. | The corresponding (N/O)=C−N−C=(N/O) tautomer is allowed. |
| F8 | Enol/enamine: removes O or N atoms adjacent to a nonaromatic C=C. | Enols, enamines, enol ethers, etc. are almost always unstable toward hydrolysis to the parent carbonyl compound. |
| F9 | Acyclic carbonates C−O−(C=O)−O−C are removed. | Acyclic carbonates are rather unstable toward hydrolysis. |
| F10 | Carbonic acids (O−$CO_2$H), carbamic acids (N−$CO_2$H), and $\beta$-carboxylic acid ((C=O)−C−$CO_2$H) are removed. | These FG decarboxylate spontaneously. |
| F11 | Bridgehead amides: If a "non-zero" bridgehead nitrogen is bound to a nonaromatic $sp^2$ atom, the ring sizes of the smallest set of smallest rings will be checked. At least one ring of the nitrogen must be ≥9. | Such amides are "twisted" and nonconjugated and therefore quite unstable toward hydrolysis. |
| F12 | C=C: Molecules of 17 atoms with nonaromatic carbon−carbon unsaturations are removed. | Nonaromatic C=C are highly frequent but often reactive toward polymerization, cycloadditions, isomerizations, oxidation, or nucleophilic addition. |

[a]No heteroatom−heteroatom bonds are generated at all.

**Table 4. Postprocessing Steps for Aromatic Heterocycles, Oximes, Nitro, $CF_3$, Halogens, and Sulfur**

| step | description | comment |
|---|---|---|
| P1 | Aromatic C to N: aromatic C atoms adjacent to an aromatic N or O atom are converted to N if valency allows. | Aromatic heterocycles with heteroatom−heteroatom bonds are created e.g. 1,2-oxazoles from furans, 1,2,3-triazoles from imidazoles. |
| P2[a] | Ketone oximes C=N−OH: ketones are converted to oximes. | Note that alkylated oximes, hydroxamates, hydrazides, and hydrazone are not considered. |
| P3 | Aromatic halogens: aromatic OH groups are changed to halogens. | Halogen = F, Cl, Br, I, max. Two Br or I per aromatic ring. |
| P4 | Trifluoromethyls: tert-butyl groups are changed to $CF_3$. | |
| P5 | Aromatic nitro groups: aromatic $CO_2$H are converted to $NO_2$. | Aliphatic nitro groups are not considered. |
| P6 | Thiophenes: sulfur was substituted for all heteroaromatic oxygen atoms. | Aliphatic thiols and thioethers are not considered. |
| P7[a] | Sulfones: carbonyl groups (C=O) in ketones, acids, carboxamides, and carbamates are changed to $SO_2$. | Note that C=S and sulfoxides (S=O) are not generated. |

[a]Steps P2 and P7 increase the heavy atom count (hac), generating for example some 17 atoms molecules with small rings. All molecules with hac >17 were removed to avoid a combinatorial explosion.

to 17 atoms in the public archives PubChem,[33] ChEMBL,[34] and DrugBank,[35] to form the reference collections up to 17 atoms PubChem-17 (2,526,453 cpds), ChEMBL-17 (89,156 cpds), and DrugBank-17 (367 cpds, approved drugs only). ChEMBL-17 and DrugBank-17 represent subsets of the larger PubChem-17 which are focused on molecules with biological activities that are reported (ChEMBL-17) respectively clinically approved (DrugBank-17). Unfortunately commercial collections such as the CAS or Beilstein archives could not be obtained. However considering that the 2.5 million molecules in PubChem-17 represent approximately 10% of the 25 million unique molecules listed in PubChem, CAS-17 probably contains around 10% of the entire CAS archive,[1,2] i.e. slightly more than 6 million molecules (2.4-fold larger than PubChem-17).

The comparison of the enumerated chemical space with known molecules starts with considering database size. Thus, GDB-17 (166.4 billion molecules) is much larger than the sum of all known molecules of similar size as found in PubChem-17

**Table 5. Database Generation Statistics**

| HAC | filters[a] | graphs[b] | hydrocarbons[c] | skeletons[d] | molecules[e] | CPU, h[f] |
|---|---|---|---|---|---|---|
| 1 | SAV, FG | 1 | 1 | 1 | 3 | 0 |
| 2 | | 1 | 1 | 3 | 6 | 0 |
| 3 | | 2 | 2 | 4 | 14 | 0 |
| 4 | | 6 | 4 | 12 | 47 | 0 |
| 5 | | 20 | 10 | 32 | 219 | 0 |
| 6 | | 74 | 31 | 119 | 1,091 | 0 |
| 7 | | 321 | 98 | 448 | 6,029 | 0 |
| 8 | | 1,663 | 370 | 2,004 | 37,435 | 0 |
| 9 | | 9,616 | 1,448 | 9,472 | 243,233 | 0 |
| 10 | | 61,840 | 6,325 | 48,721 | 1,670,163 | 0 |
| 11 | | 427,135 | 29,496 | 264,321 | 12,219,460 | 3 |
| 12 | NA2SR | 3,120,002 | 104,165 | 1,188,127 | 72,051,665 | 18 |
| 13 | | 23,722,244 | 651,850 | 7,370,864 | 836,687,200 | 206 |
| 14 | NBH3R | 186,092,397 | 752,277 | 27,419,837 | 2,921,398,415 | 856 |
| 15 | 1SR | 1,496,007,875 | 960,415 | 118,977,963 | 15,084,103,347 | 5,378 |
| 16 | | 12,176,341,897 | 1,331,875 | 213,259,331 | 38,033,661,355 | 14,415 |
| 17 | 0SR, C=C | 100,418,784,003 | 1,583,786 | 962,417,271 | 109,481,780,580 | 79,259 |
| **SUM** | | 114,304,569,097 | 5,422,154 | 1,330,958,530 | 166,443,860,262 | 100,134 |

[a]See Tables 1–4 for details. [b]Graphs produced by GENG for planar, connected graphs up to 17 nodes with maximum node valence of four. [c]Hydrocarbons generated from graphs and passing the filters in Table 1 for limited ring strain and complexity. [d]Unsaturated hydrocarbons generated from hydrocarbons using filters in Table 2. [e]Molecules generated from hydrocarbons by adding heteroatoms (Table 3 and 4), as 2D-structures and stored as SMILES. [f]Computation was parallelized on 360 CPU.

(0.001% of GDB-17). The size of GDB-17 originates in the increase of possible molecules as a function of heavy atom count, which is exponential and much steeper than in the reference databases (Figure 2A). As a consequence the MW range of GDB-17 shows a sharp peak at 240 < MW < 250 Da. The same distribution is observed in the leadlike subset GDBLL-17 (29 billion structures, see below) and leadlike/no small ring subset GDBLLnoSR-17 (22 billion structures, see below), while the MW distribution in the reference databases is more even (Figure 2B).

GDB-17 contains an impressive number of molecules in the area of known drugs. For example, millions of isomers can be identified in GDB-17 for fifteen typical marketed drugs of 14 to 17 atoms selected from DrugBank-17 (Table 6, Figure 3). The examples shown in Figure 3 were selected among isomers with a high shape similarity to the parent drug as measured by the OpenEye scoring function ROCS (Rapid Overlay of Chemical Structures), a well validated virtual screening tool to identify bioactive analogs.[39,40] These isomers include obvious variations of the parent structure such as "methyl walk" analogs, for example structures **5** and **11** as isomers of drugs **4** and **10**, as well as nontrivial changes such as different aromatic heterocycles (**2, 3, 8, 9, 17, 18, 35, 38, 39, 46**), different ring size and connectivity (**15, 23, 24, 26, 27, 32, 33, 39, 41, 44–48**), and different functional groups (**14, 29, 30, 32, 33, 36, 39, 42, 48**).

On the other hand, GDB-17 represents a selective enumeration and therefore does not contain all molecules in the reference databases. Overall 57% of PubChem-17, 60% of ChEMBL-17, and 68% of DrugBank-17 are compatible with the GDB-17 enumeration rules. The molecules found in the reference databases but not considered for GDB-17 contain nonenumerated features such as certain types of halogens (e.g., aliphatic halogens) or sulfurs (thiols, thioethers, thioureas), functional groups (e.g., acyclic acetals, hemiacetals, aminals, azides, aliphatic nitro groups), elements (P, Si, B, etc.), skeletons

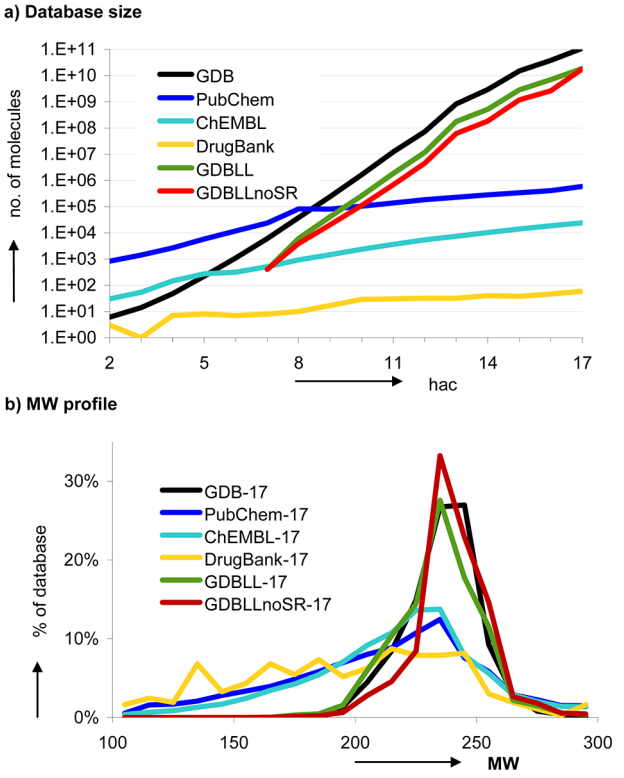



**Figure 2.** Size and MW profiles of the enumerated chemical space in GDB and the reference databases PubChem, ChEMBL, and DrugBank. The size of the leadlike subsets of GDB (GDBLL, GDBLLnoSR) is extrapolated from analyzing a 1% random subset of GDB-17.

(nonaromatic C=C), or graphs (e.g., spiro-fused cyclopropanes) (Figure 4A).

**Small Rings, Topology, and Compound Categories.** The most striking difference between the enumerated chemical


dx.doi.org/10.1021/ci300415d | J. Chem. Inf. Model. 2012, 52, 2864–2875

Table 6. Drug Isomers Found in GDB-17

| drug name[a] | elemental formula | no. of isomer |
|---|---|---|
| Acyclovir | $C_8H_{11}N_5O_3$ | 8,132,952 |
| Aminoglutethimide | $C_{13}H_{16}N_2O_2$ | 183,901,628 |
| Aminophenazone | $C_{13}H_{17}N_3O$ | 97,853,936 |
| Dexmedetomidine | $C_{13}H_{16}N_2$ | 9,721,191 |
| Diethylcarbamazine | $C_{10}H_{21}N_3O$ | 22,409 |
| Ethoxzolamide | $C_9H_{10}N_2S_2O_3$ | 4,563,491 |
| Felbamate | $C_{11}H_{14}N_2O_4$ | 369,751,288 |
| Fencamfamine | $C_{15}H_{21}N$ | 53,917,207 |
| Guanadrel | $C_{10}H_{19}N_3O_2$ | 60,319,220 |
| Procaine | $C_{13}H_{20}N_2O_2$ | 476,975,898 |
| Sulfadiazine | $C_{10}H_{10}N_4SO_2$ | 17,003,297 |
| Tinidazole | $C_8H_{13}N_3SO_4$ | 24,575,941 |
| Tizanidine | $C_9H_8N_5SCl$ | 109,635 |
| Trioxsalen | $C_{14}H_{12}O_3$ | 1,800,849 |
| Varenicline | $C_{13}H_{13}N_3$ | 19,676,640 |

[a]See Figure 3 for structural formula of the drugs and examples of isomers.

space in GDB-17 and known molecules resides in the occurrence of small rings (3- or 4-membered ring). Small rings are very frequent in the systematic enumeration of graphs and the resulting molecules. However they are also relatively difficult to synthesize and often unstable, and indeed they are not found very often in known molecules. While only 4−6% of the compounds in the reference databases contain small rings, the enumerated chemical space up to 16 atoms is to 83% a small ring compound database (Figure 4B). The fraction of small ring compounds in GDB falls to 8.2% at 17 atoms because no small ring were allowed in 17 node graphs (small ring molecules at 17 atoms stem from atom-adding postprocessing steps such as the transformation of carbonyls to sulfonyls and of ketones to oximes, Table 5). Nevertheless the majority (66%) of GDB-17 are molecules with 17 atoms, and the low percentage of small ring compounds at 17 atoms results in an overall 28% of small ring compounds in GDB-17 (25% in GDBLL-17).

In terms of topology, all databases contain approximately two-thirds of molecules with two or three cycles (Figure 4C). Key differences occur in acyclic compounds, which are relatively rare in GDB-17 (1.8%, 3.0 billion molecules) but make up 25% of DrugBank-17. Tri- and polycyclic molecules furthermore combine to 32% of GDB-17 but are much less frequent in the reference databases (PubChem-17: 7%; ChEMBL-17: 16%, DrugBank-17: 6%). The leadlike subset is also rich in tri- and polycyclic compounds (GDBLL-17: 33%), but their proportion is reduced when small rings are removed (GDBLLnoSR-17: 22%).

In terms of compound categories, heteroaromatic compounds make up a large third of GDB-17 and the reference databases (Figure 4D). By contrast aromatics, which also make up a third of reference databases, are quite rare in GDB-17 (0.8%, 1.3 billion molecules). GDB-17 is instead much richer in nonaromatic heterocycles (GDB-17: 57%, GDBLL-17: 41%, GDBLLnoSR-17: 35%) than the reference databases of known compounds (PubChem-17: 12%, ChEMBL-17: 10%, DrugBank-17: 12%).

**Polarity and Leadlikeness.** The histograms of the calculated octanol:water partition coefficient clogP shows that GDB-17 and DrugBank-17 contain more polar molecules than PubChem-17 and ChEMBL-17 (Figure 5A/B). A similar effect is visible in other polarity descriptors such as the number of H-bond donor atoms (Figure 5C/D). The fact that polar molecules often require longer syntheses and are more difficult to purify than apolar ones might explain their lower proportion in PubChem and ChEMBL, which contain mostly synthesized molecules, compared to GDB-17 representing the spectrum of possibilities. The frequency of polar molecules in the systematic enumeration of GDB-17 results in only 18% of GDB-17 being leadlike compounds as defined by the value ranges $1 < \text{clogP} < 3$ and $100 < \text{MW} < 350$ Da.[32] These 18% correspond to 29 billion molecules defining the GDBLL-17 subset, 22 billion of which do not have small rings and form the GDBLLnoSR-17 subset. By comparison approximately half of the compounds from the reference databases are leadlike (PubChem-17: 47%; ChEMBL-17: 49%; DrugBank-17: 36%).

**Molecular Shape.** Organic molecules can be classified in terms of shape by analyzing the principal moments of inertia of their 3D structure, which allows to classify molecules either as rods (linear shape, e.g. stretched alkanes), discs (cyclic planar shape, e.g. benzene), or spheres (globular shape, e.g. cubane or adamantane). This analysis shows that the vast majority of currently used druglike molecules are either rodlike or disklike. Only a minority of the molecules used in medicinal chemistry possess any significant third dimension, leading to shape considerations as a design criteria for screening libraries.[41] Closer analyses of successes and failures show that molecules with a significant third dimension in shape are indeed often more successful in drug development programs, suggesting an "escape out of flatland" as a valuable strategy to search for better drug molecules.[42,43] Nonplanarity is also more pronounced in natural products (NP) and products from diversity-oriented synthesis (DC) compared to commercial screening compounds (CC) and probably contributes to the higher protein binding selectivity of NP and DC compared to CC as observed in small molecule microarray experiments.[44,45]

A 16.7 million random subset of GDB-17 was subjected to the above shape analysis, and the results were compared with the data for the reference databases. The analysis showed that GDB-17 molecules significantly populate the third dimension, which implies that the ″escape out of flatland″ is statistically unavoidable when considering the enumerated chemical space (Figure 6). The shape distribution into the third dimension is similar for the 29 billion leadlike subset GDBLL-17 and the 22 billion leadlike subset without small rings GDBLLnoSR-17. By comparison the known molecules in PubChem-17, ChEMBL-17, and DrugBank-17 are essentially rods and discs with relatively few spherical molecules. This "flatness" is a direct consequence of the abundance of aromatic systems in these databases of known compounds. Conversely, the occurrence of 3D-shaped molecules in GDB-17 results from the low proportion of acyclic and aromatic compounds and the high frequency of saturated heterocycles in the enumerated chemical space. GDB-17 molecules also contain more quaternary carbon centers (qv, Figure 7A/B) and bonds in fused rings (bfr, Figure 7C/D) compared to known compounds, which are features strongly associated with nonplanarity. The decrease in bfr at 14, 15, and 17 atoms in GDB reflects the introduction of the "no bridgehead in 3-rings", "one small ring", and "no small ring" filters which strongly reduce the number of topologies with high bfr. Somewhat unexpectedly, the small-ring filters also reduce the number of quaternary

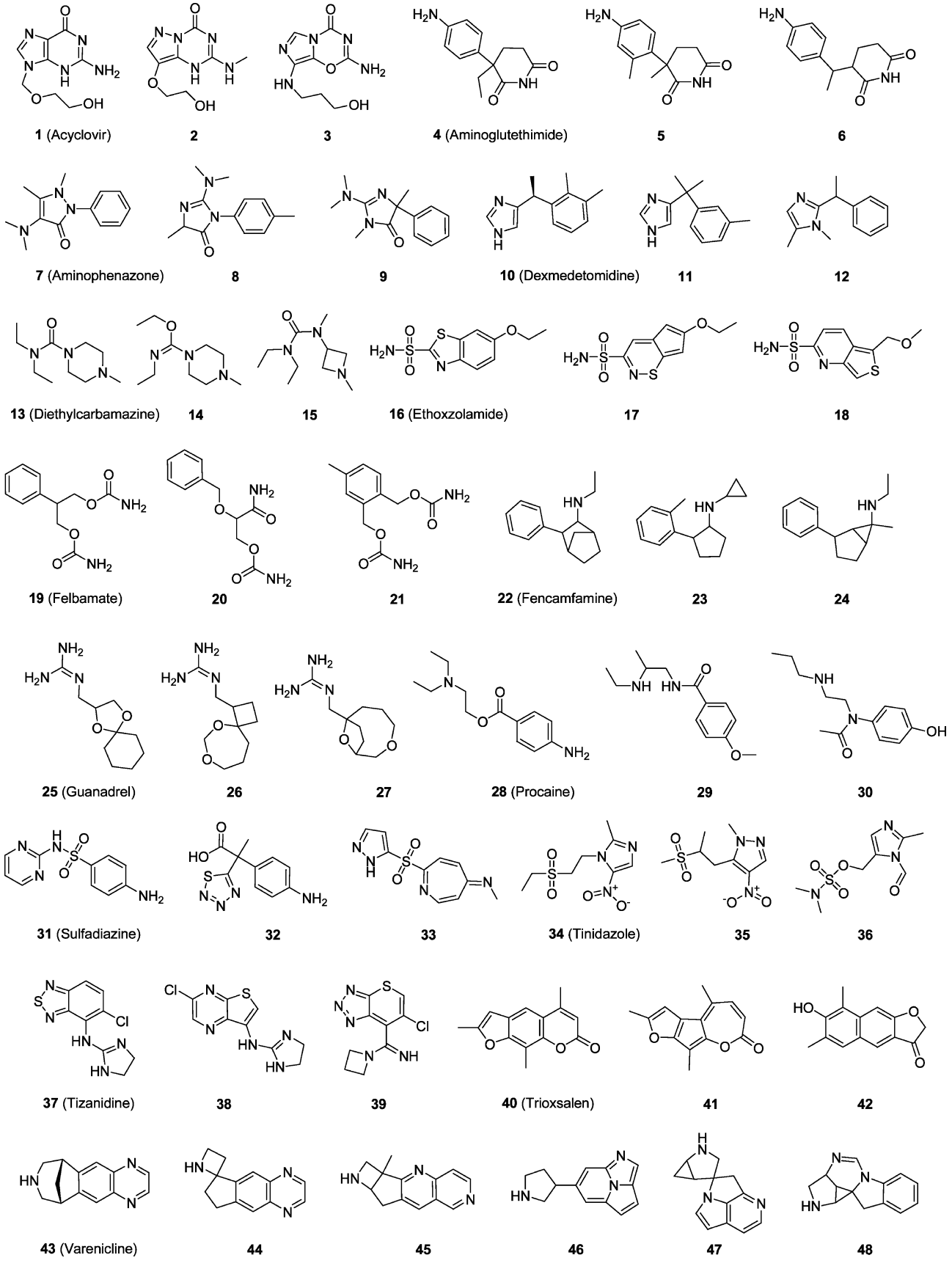



**Figure 3.** Drugs and examples of isomers found in GDB-17. All isomers shown have a shape similarity score ROCS > 1.4. None of the isomers shown are known (Scifinder search). Only acyclovir does not occur in GDB-17 because it contains a hemiaminal (N−Csp$^3$−O), a functional group which is excluded from the enumeration.


dx.doi.org/10.1021/ci300415d | J. Chem. Inf. Model. 2012, 52, 2864−2875

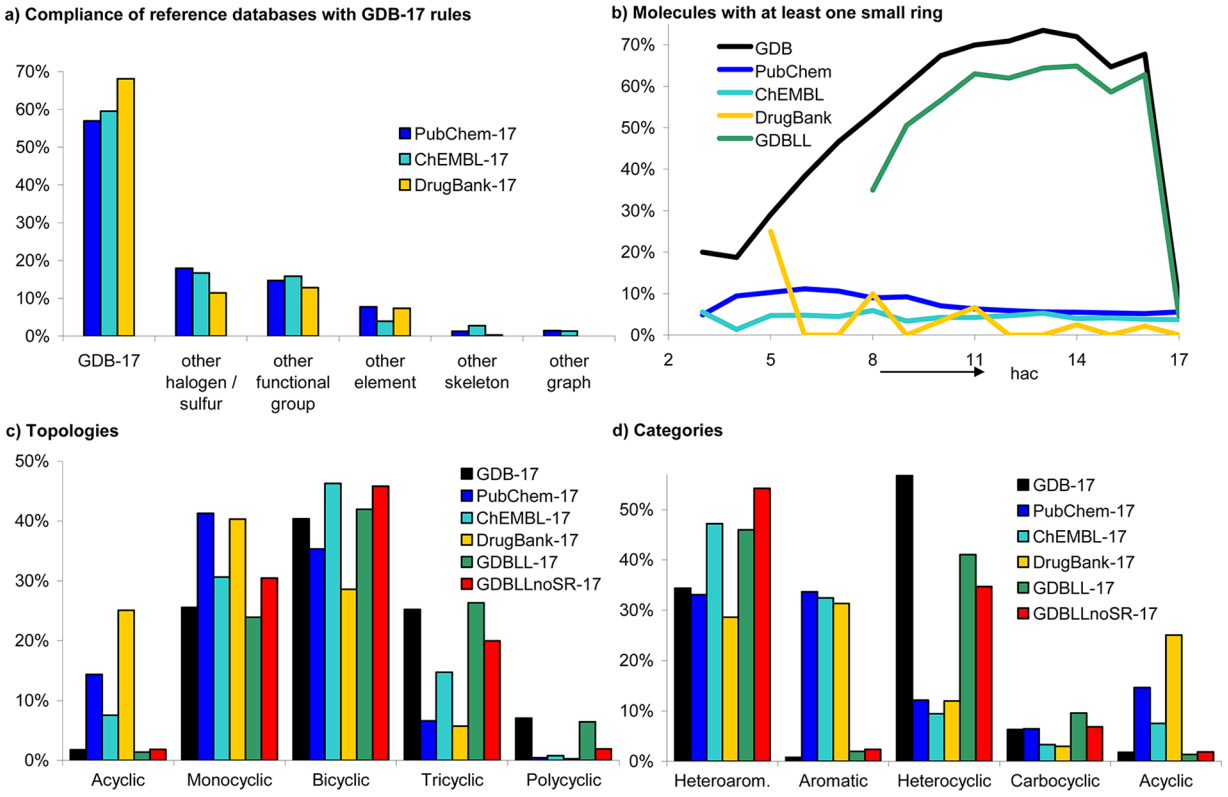



**Figure 4.** Molecule topologies and categories in GDB-17 and reference databases. **A.** Percentage of reference database compatible with GDB-17 enumeration rules or excluded due to nonenumerated halogen (acyl halide, aliphatic halocarbons) or sulfur (thiols, thioethers), functional groups (acyclic acetals, hemiacetals, aminals, azides, aliphatic nitro groups), element (P, Si, B, Bi, Hg, etc.), skeleton (nonaromatic C=C), or graph (e.g., small rings at 17 atoms). **B.** Fraction of compounds with small rings. **C.** Topologies **D.** Database contents as function of molecular categories. Molecules are assigned to one category only with priority order heteroaromatic > aromatic > heterocyclic > carbocyclic > acyclic. The data for GDB-17 and its subsets were computed from a 1% random subset of the database.

centers per molecule, as seen by the fact that the GDBLLnoSR subset contains fewer quaternary centers than GDB and its leadlike subset.

**Stereochemistry.** The much higher frequency of nonplanar molecules in GDB-17 compared to the reference databases should also be reflected in a larger number of stereocenters and hence possible stereoisomers per compound. The number of possible stereoisomers per molecule was determined using the 3D-generator CORINA,[46] which exhaustively generates stereoisomers from 2D structures. CORINA correctly excludes impossible combinations of stereoisomer flipping (e.g., only one stereoisomer for norbornane). CORINA produces enantiomers as pairs with the exception of atropisomers, which are rather rare, implying that molecules for which only a single diastereoisomer is produced are almost always achiral. Their number provides a lower estimate of the number of achiral molecules because *meso* compounds (e.g., 1,4-dimethylcyclohexane or (*R,S*)-2,3-butanediol) and achiral *Z/E* isomer pairs are not singled out.

The stereoisomer counting with CORINA was performed on the 16.7 million subset of GDB-17 and on the reference databases (Figure 8). GDB-17 molecules produced an average of 6.4 stereoisomers per molecule (GDBLL-17: 5.7 stereoisomers/cpd, GDBLLnoSR: 5.1 stereoisomers/cpd), which is three times more than in the reference databases (PubChem-17: 2.0 stereoisomers/cpd, ChEMBL-17: 2.0 stereoisomers/cpd, DrugBank-17: 2.1 stereoisomers/cpd). More than half of the molecules in the reference databases have only one stereoisomer (PubChem-17: 56%, ChEMBL-17: 58%, DrugBank-17: 55%), while only 5% are molecules with eight or more possible stereoisomers (PubChem-17: 4.1%, ChEMBL-17: 4.6%, DrugBank-17: 5.2%). By contrast GDB-17 (respectively GDBLL-17, GDBLLnoSR-17) contains only 22% (respectively 23%, 27%) of molecules with a single stereoisomer but 44% (respectively 38%, 32%) of molecules with eight of more stereoisomers. The smaller average number of stereoisomers per compound as a function of hac in GDBLLnoSR-17 compared to GDBLL-17 shows that the presence of small rings is partly responsible for the larger number of stereoisomers in GDB-17 compared to known compounds.

**Novelty and Scaffolds.** The above analyses show that the novelty of GDB-17 compared to PubChem-17 can be assigned in part to global structural features including the relative rarity of aromatic and acyclic compounds, the frequent occurrence of molecules with small rings and nonaromatic heterocycles, and the higher proportion of polar molecules (clogP < 0). GDB-17 molecules also differ from PubChem-17 molecules in that they contain more structural features leading to 3D-shapes, such as quaternary centers and bonds in fused rings, as well as generally more stereoisomers per molecule. Nevertheless GDB-17 contains impressive numbers of compounds within any constraints, as exemplified with the millions of isomers of known drugs and the size of the leadlike/no small rings subset GDBLLnoSR-17 containing 22 billion molecules. By their number these molecules are

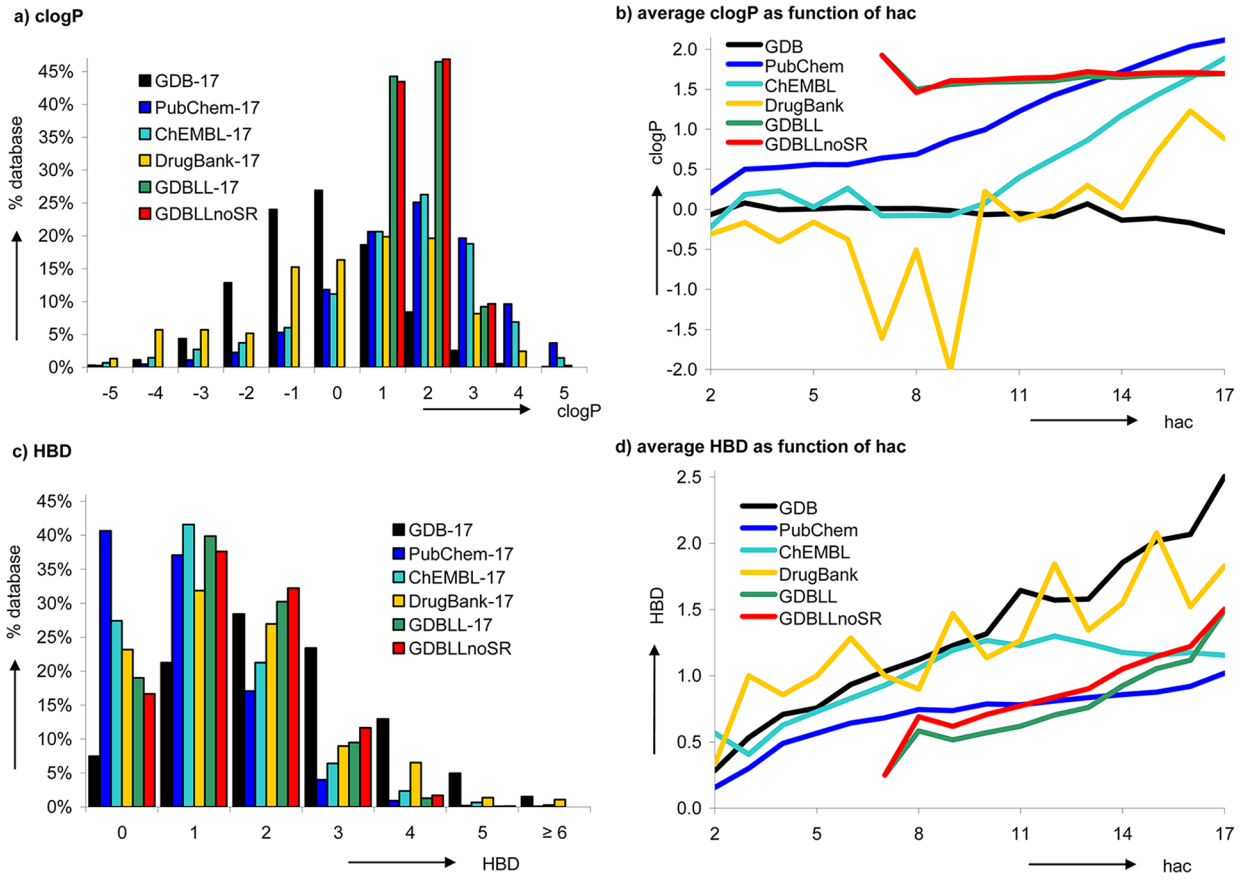



**Figure 5.** Polarity features. **A**. c logP histogram in intervals −5.5 to −4.5, −4.5 to −3.5, etc; **B**. Average clogP as function of hac; **C**. H-bond donor atom (HBD) histogram; **D**. Average HBD as function of hac. The data for GDB-17 and its subsets were computed from a 1% random subset of the database.

necessarily new although they represent variations of known compound types.

One can also analyze the databases for novelty independent of global parameters by focusing on the occurrence of "scaffolds". As "scaffolds" we considered either the "Murcko scaffolds",[47] which are defined as the saturated hydrocarbon graph of a molecule pruned of any terminal atom, or "ring systems" defined as hydrocarbon graphs without acyclic bonds.[21] The analysis was performed for GDB-17 by considering the 5.4 million unique graphs used for molecule generation (Table 1) and extracting graphs corresponding to Murcko scaffolds and ring systems. For PubChem-17 each molecule was converted to its parent saturated hydrocarbon. All terminal atoms were then removed iteratively to produce "Murcko scaffolds", and all acyclic bonds were removed to produce "ring systems". Each resulting list was reduced to unique structures by removing duplicates. Each series was split into three categories by analyzing the smallest set of smallest rings as follows: a) scaffolds containing at least one small ring (″SR″); b) scaffolds containing only 5−7 membered rings (″5−7″); and c) scaffolds without small rings containing at least one 8-membered or larger ring (″8+″). The scaffolds were further subdivided according to the number of quaternary centers (Table 7).

The scaffold and ring system analysis shows that GDB-17 contains 35-fold more Murcko scaffolds and 61-fold more ring systems than PubChem-17. The majority of the imbalance stems from scaffolds containing small rings (Murcko scaffolds: 52-fold excess in GDB-17, ring systems: 109-fold excess in GDB-17), in particular small ring scaffolds with quaternary centers (Murcko scaffolds: 105-fold excess in GDB-17, ring systems: 170-fold excess in GDB-17). If considering only Murcko scaffolds or ring systems with 5- to 7-membered rings and without any quaternary center, which are the easiest to synthesize and most common ring systems, the number of scaffolds is only, but still, 2-fold larger in GDB-17 compared to PubChem-17. Ring systems that are yet unknown even as substructure are readily identified in GDB-17, such as the yet unknown $C_{17}$-hydrocarbon graphs **49**−**55** shown in Figure 9.

## ■ CONCLUSION

In summary the enumeration of organic molecules starting from mathematical graphs was realized up to 17 atoms of C, N, O, S, and halogens, yielding 166.4 billion molecules corresponding to a defined set of functional group and atom types. Compared to the 2.5 million known molecules up to 17 atoms found in PubChem, GDB-17 molecules contain generally more rings, in particular small rings, as well as many non-aromatic heterocycles. On the other hand, cyclic and aromatic compounds form a much smaller fraction of the database compared to PubChem. GDB-17 molecules furthermore contain more quaternary centers and bonds in fused rings than PubChem molecules, resulting in significant 3D-shapes and a larger number of stereoisomers per molecule. GDB-17 molecules are on average also more polar than known molecules


dx.doi.org/10.1021/ci300415d | J. Chem. Inf. Model. 2012, 52, 2864−2875

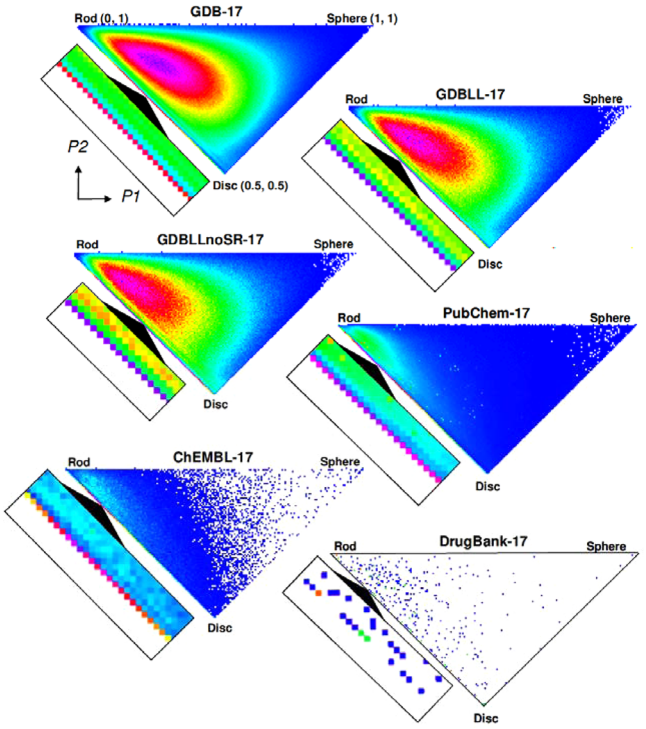



**Figure 6.** Molecular shape analyzed by the principal moments of inertia.[41] Occupancy maps are shown in the (P1,P2)-plane, in which *P1* and *P2* are the normalized ratios of the principal moments of inertia (for details see section Methods), and are colored from blue (1 cpd/pixel) to purple (maximum cpd/pixel for each map: GDB-17: 4,691, GDBLL-17: 889, GDBLLnoSR-17: 684, Pubchem-17: 6202, Chembl-17: 487, Drugbank-17: 4). The inserts show an enlarged view of the lower left edge of each triangle where occupancy is highest for PubChem-17, ChEMBL-17, and DrugBank-17. The GDB-17, GDBLL-17, and GDBLLnoSR-17 were analyzed with a random subset of 16.7 million molecules from GDB-17. For all compounds a single stereoisomer was analyzed as generated by CORINA.

(clogP < 0), although a leadlike subset occupying the range 0 < clogP < 3 still contains 22 billion molecules even when excluding small ring compounds. The structural diversity of GDB-17 is evidenced by the presence of a much larger number of scaffolds compared to known molecules. The abundance of nonplanar molecules suggests that the enumerated chemical space might serve as a rich source of inspiration to design new molecular series for drug discovery.

As to the size of GDB-17, working with 166.4 billion structures is challenging and currently not applicable to advanced virtual screening methods such as shape-based analyses or docking, which are computationally relatively intensive. For such applications a randomly selected subset of GDB-17 of a few hundred thousand to a few million structures is statistically significant and can be used as representative of the whole database. On the other hand, the identification of single molecules such as the selected analogs of known drugs shown in Figure 3, or the examples of polycyclic hydrocarbons shown in Figure 9, can deliver many more interesting results with the complete database because every single molecule is different and identifiable in its own right. The assembly of a searchable version of the entire GDB-17 database and its use for identifying drug analogs by virtual screening represented a challenge of its own and will be described in a separate publication.

## METHODS

**General.** All code packages were written in Java 1.6 with Jchem Libaries from ChemAxon. Every filter was applied to the imported molecule to define bond and atom positions of functional groups. All computations were parallelized on a 360-CPU cluster and manually controlled (100,000 CPU hours corresponds to 11 CPU years). Every step was completed before starting the next. All together around 40,000 single calculations have been done. To preserve disk space every output was compressed directly into gzip either by piping with the bash command gzip or by the implementation of gzip into the GZIPStream in Java BufferedReader/Writer.The complete GDB is more than 400 GB as gzip.

**Enumeration.** *Graphs.* The program Nauty from McKay was used to generate the connectivity tables for graphs, as found under http://cs.anu.edu.au/people/bdm/. The Nauty subprogram GENG was run up to 17 nodes for the generation of all possible graphs/geng -cd1D4 (Number of Nodes) The check was done for planarity of the graphs by PLANARG to avoid molecules with crossed bonds, e.g. Claus' benzene./ planarg

*Hydrocarbons.* The resulting output of GENG is a G6 string for a connection table, which was imported and converted to the corresponding hydrocarbons by exchanging the nodes with carbon atoms and the edges with single bonds. Hydrocarbons were filtered for desirable features (Table 1) because the majority of graphs includes 3- and 4-membered rings or multiple connected globular ring systems.

*Skeletons.* Each single bond was checked for the combinatorially introduction of double bonds (only four bonds per carbon are possible). Additionally every resulting double bond was checked for the combinatorially introduction of triple bonds. The resulting hydrocarbons were aromatized and dearomatized to avoid multiple copies of the same aromatic ring system. Unsaturated hydrocarbons were filtered for desirable features, e.g. the majority of unsaturated hydrocarbons includes allenes (Table 2).

*CNO Molecules.* Every monovalent, divalent, and trivalent carbon position was checked for the substitution with nitrogen following valency rules. Each monovalent and divalent carbon position was then checked for the substitution with oxygen. Before each exchange it was checked if the position is adjacent to a nitrogen or oxygen atom to avoid the generation of heteroatom heteroatom bonds and speeding up computation. Additionally it was checked before if the position is next to a sp atom, in which case no N or O atom was introduced. Symmetric positions were calculated only once. The resulting CNO-molecules were checked for desirable features to avoid unstable functional groups and to reduce the combinatorial explosion (Table 3). Each molecule was converted to unique SMILES strings to check for duplicates before storing the molecule.

*Postprocessing for Oximes, Nitro, $CF_3$, Halogens, and Sulfur.* Postprocessing for introducing additional diversity was performed as described in Table 4.

**Shape Analysis.** The shape analysis was adapted from Sauer and Schwarz[41] and was written in Java 1.6. SMILES were converted into 3D structure using CORINA.[46] The position of the molecule was expressed in a (x,y,z)-coordinate system defined by its principal axes. For each principal axis

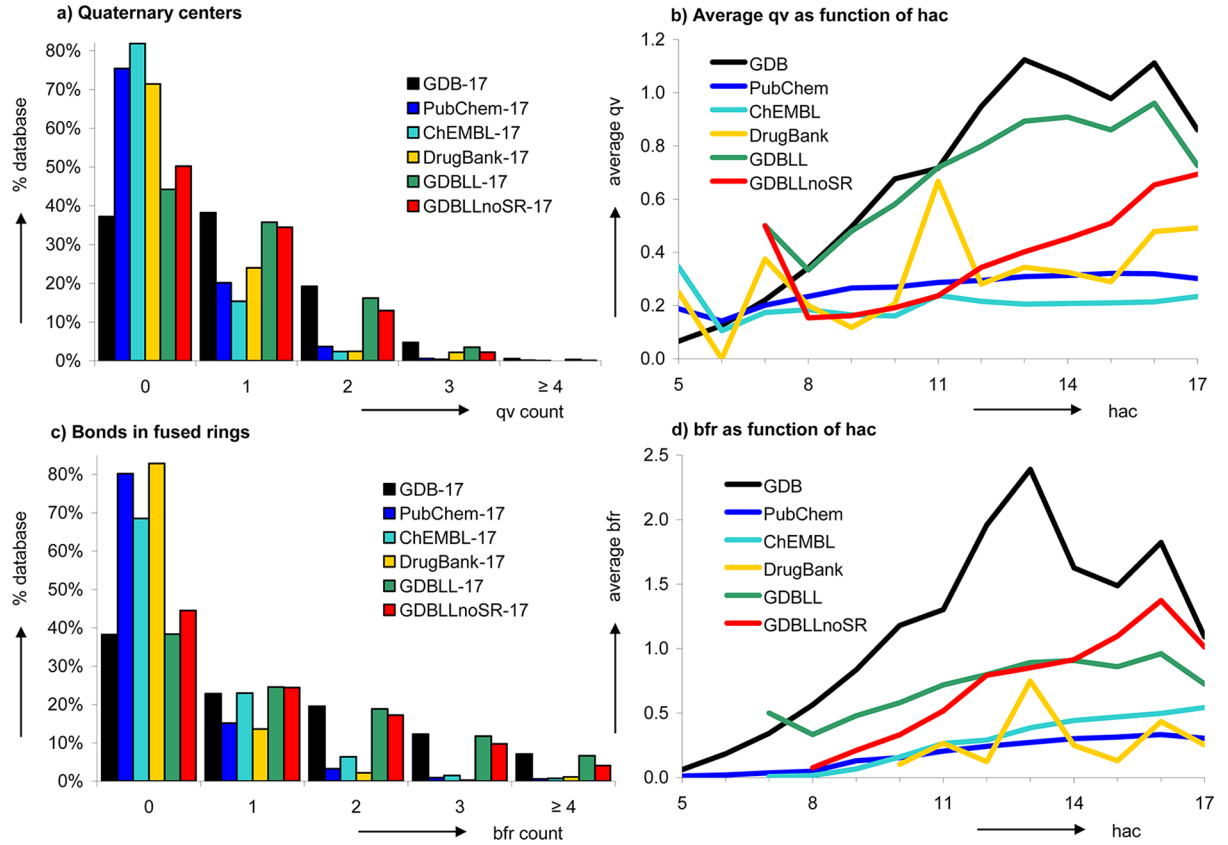



**Figure 7.** Histograms of quaternary centers (qv) and bonds in fused rings (bfr) in the different databases. The data for GDB-17 and its subsets were computed from a 1% random subset of the database.

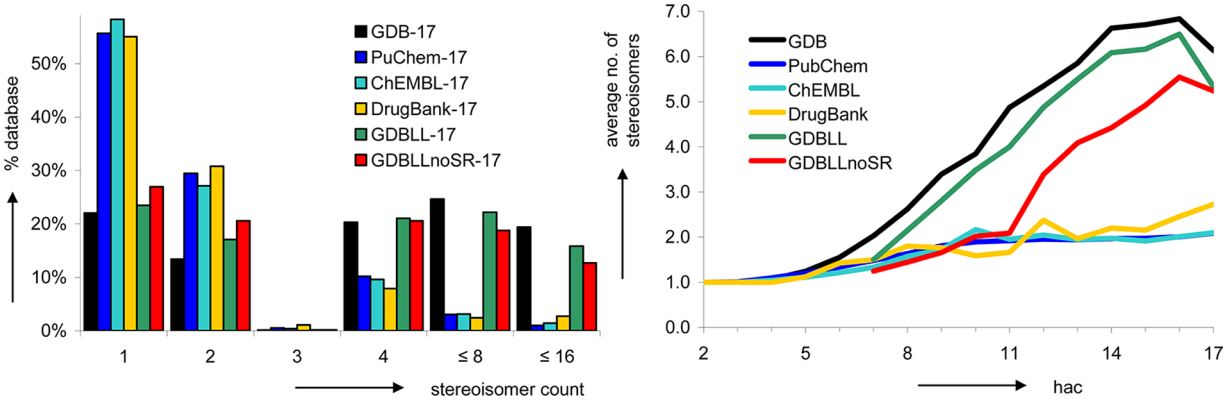



**Figure 8.** Stereochemistry. A Numbers of stereoisomers per compounds. B. Average number of stereoisomer per compound as a function of hac. Stereoisomers were generated from SMILES using CORINA. The data for GDB-17, GDBLL-17, and GDBLLnoSR-17 stem from the analysis of a random 16.7 million subset of GDB-17.

the moment of inertia was calculated using the general equation

$$I = mr^2$$

Specified for each axis it yields

$$Ix = mr_x^2$$

$$Iy = mr_y^2$$

$$Iz = mr_z^2$$

in which the squares of the radii around the axes are defined as $r_x^2 = y^2 + z^2$, $r_y^2 = x^2 + z^2$, and $r_z^2 = x^2 + y^2$. The moments of inertia $I_x$, $I_y$, and $I_z$ were then sorted in ascending order to yield $I_1$, $I_2$, and $I_3$. $I_1$ and $I_2$ were finally divided by the highest moment of inertia $I_3$ to yield the values $P1 = I_1/I_3$ and $P2 = I_2/I_3$.

The (P1,P2)-plane defines a two-dimensional triangular space with distinct boundaries, i.e. structures cannot be found outside the triangle. The triangle also has three distinct edges defining the different dimensionality of molecular shapes: The upper left edge of the triangle (0,1), the lower center edge (0.5,0.5), and the upper right edge (1,1) define 1D rodlike, 2D disklike, and 3D spherical structures, respectively.


dx.doi.org/10.1021/ci300415d | J. Chem. Inf. Model. 2012, 52, 2864−2875

Table 7. Scaffold Analysis of GDB-17 and PubChem-17[a]

| Murcko scaffolds | | | | | |
|---|---|---|---|---|---|
| no. of quat. C | 0 | 1 | 2 | >2 | SUM |
| GDB-17, SR | 19,804 | 56,975 | 65,536 | 42,895 | 185,210 |
| GDB-17, 5−7 | 1,736 | 2,561 | 1,195 | 217 | 5,709 |
| GDB-17, 8+ | 1,113 | 466 | 44 | 0 | 1,623 |
| SUM | 22,653 | 60,002 | 66,775 | 43,112 | **192,542** |
| PubChem-17, SR | 1,997 | 1,114 | 405 | 56 | 3,572 |
| PubChem-17, 5−7 | 960 | 562 | 121 | 3 | 1,646 |
| PubChem-17, 8+ | 307 | 41 | 8 | 0 | 356 |
| SUM | 3,264 | 1,717 | 534 | 59 | **5,574** |
| **Ring Systems** | | | | | |
| no. of quat. C | 0 | 1 | 2 | >2 | SUM |
| GDB-17, SR | 12,607 | 45,419 | 60,720 | 42,293 | 161,039 |
| GDB-17, 5−7 | 1,135 | 2,143 | 1,126 | 217 | 4,621 |
| GDB-17, 8+ | 978 | 426 | 44 | 0 | 1,448 |
| SUM | 14,720 | 47,988 | 61,890 | 42,510 | **167,108** |
| PubChem-17, SR | 600 | 521 | 314 | 45 | 1,480 |
| PubChem-17, 5−7 | 480 | 375 | 111 | 3 | 969 |
| PubChem-17, 8+ | 254 | 36 | 8 | 0 | 298 |
| SUM | 1,334 | 932 | 433 | 48 | **2,747** |

[a]Murcko scaffolds are hydrocarbon graphs without any terminal atoms and ring systems are hydrocarbon graphs without any acyclic bonds. Scaffolds and ring systems are divided into three categories: SR: at least one small (3- or 4-membered) ring; 5−7: containing only 5- to 7-membered rings; 8+: no small ring and at least one 8-membered or larger ring. Rings are analyzed in the smallest set of smallest rings i.e.. bicyclo[2.2.1]heptane (norbornane) contains two 5-membered rings, while its 6-membered ring is not considered.

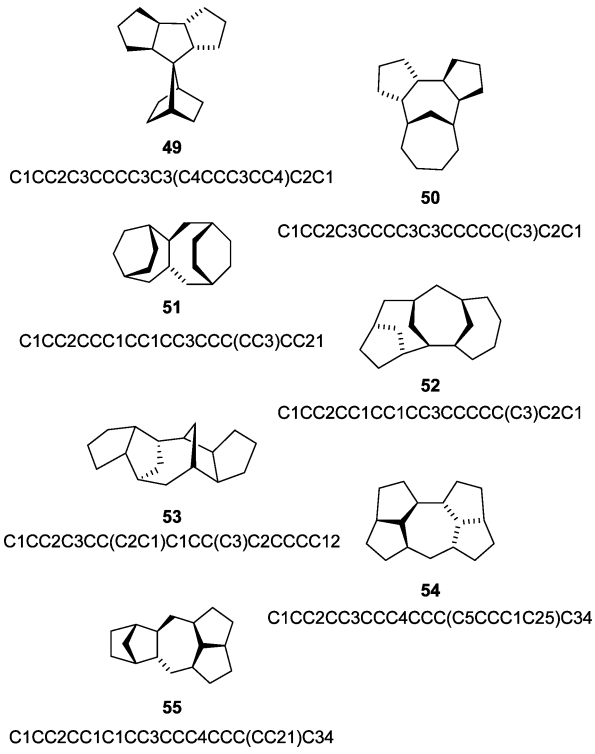



Figure 9. Examples of yet unknown $C_{17}$-ring systems from GDB-17. These hydrocarbons do not give any hits in Scifinder using "any atom" types for carbons and "any bond" for bonds, including substructure searches but locking further ring fusions. Stereochemistry is not considered in these searches. The ring systems are shown as one possible stereoisomer.

**Stereoisomer Counting.** The CORINA command ./corina.lnx -i t=smiles -o t=sdf -d ori,stergen,rs | grep '$$$$' | wc −l was used to count stereoisomers.

**Distribution.** A 50 million random subset of GDB-17 and the leadlike and leadlike/no small ring fraction of this subset are freely available for download as a SMILES list from www.gdb.unibe.ch.

## AUTHOR INFORMATION


### Corresponding Author

*Phone: +41 31 631 43 25. Fax: +41 31 631 80 57. E-mail: jean-louis.reymond@ioc.unibe.ch.


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

This work was supported financially by the University of Berne, the Swiss National Science Foundation, the NCCR TransCure, and the NCCR Chemical Biology.

## REFERENCES


(1) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F.; Schenck, R. J.; Trippe, A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443−4451.

(2) ACS NEWS. *Chem. Eng. News* **2011**, *89*, 38.

(3) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(4) Schreiber, S. L. Small molecules: the missing link in the central dogma. *Nat. Chem. Biol.* **2005**, *1*, 64−66.

(5) Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2009**, *9*, 580−588.

(6) Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H. J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E. Recent trends and observations in the design of high-quality screening collections. *Future Med. Chem* **2011**, *3*, 751−766.

(7) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2004**, *3*, 711−715.

(8) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011**, *2*, 349−355.
(9) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(10) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.
(11) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30−38.
(12) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.* **2011**, *672*, 299−323.
(13) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.
(14) Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **2009**, *20*, 429−36.
(15) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.
(16) Cayley, E. Ueber die analytischen Figuren, welche in der Mathematik Bäume genannt werden und ihre Anwendung auf die Theorie chemischer Verbindungen. *Chem. Ber.* **1875**, *8*, 1056−1059.
(17) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of artificial intelligence for chemical inference. I. Number of possible organic compounds. Acyclic structures containing carbon, hydrogen, oxygen, and nitrogen. *J. Am. Chem. Soc.* **1969**, *91*, 2973−2976.
(18) Steinbeck, C. Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* **2004**, *21*, 512−518.
(19) Reymond, J. L.; Ruddigkeit, L.; Blum, L. C.; Van Deursen, R. The enumeration of chemical space. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 717−733.
(20) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 1504−1508.
(21) 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.
(22) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733.
(23) Blum, L. C.; van Deursen, R.; Reymond, J. L. Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637−647.
(24) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. *ChemMedChem* **2008**, *3*, 1520−1524.
(25) Nguyen, K. T.; Luethi, E.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. 3-(aminomethyl)piperazine-2,5-dione as a novel NMDA glycine site inhibitor from the chemical universe database GDB. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 3832−3835.
(26) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring a7-nicotinic receptor ligand diversity by scaffold enumeration from the Chemical Universe Database GDB. *ACS Med. Chem. Lett.* **2010**, *1*, 422−426.
(27) Luethi, E.; Nguyen, K. T.; Burzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J. L. Identification of selective norbornane-type aspartate analogue inhibitors of the glutamate transporter 1 (GLT-1) from the chemical universe generated database (GDB). *J. Med. Chem.* **2010**, *53*, 7236−7250.
(28) Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. Discovery of alpha7-nicotinic receptor ligands by virtual screening of the Chemical Universe Database GDB-13. *J. Chem. Inf. Model.* **2011**, *51*, 3105−3112.
(29) Brethous, L.; Garcia-Delgado, N.; Schwartz, J.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Synthesis and nicotinic receptor activity of chemical space analogues of N-(3R)-1-azabicyclo[2.2.2]oct-3-yl-4-chlorobenzamide (PNU-282,987) and 1,4-diazabicyclo[3.2.2]nonane-4-carboxylic acid 4-bromophenyl ester (SSR180711). *J. Med. Chem.* **2012**, *55*, 4605−4618.
(30) Reymond, J. L.; Awale, M. Exploring chemical space for drug discovery using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.
(31) Foloppe, N. The benefits of constructing leads from fragment hits. *Future Med. Chem.* **2011**, *3*, 1111−1115.
(32) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 3743−3748.
(33) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623−W633.
(34) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(35) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−D1041.
(36) McKay, B. D. Practical graph isomorphism. *Congressus Numerantium* **1981**, *30*, 45−87.
(37) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.
(38) Rishton, G. M. Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discovery Today* **2003**, *8*, 86−96.
(39) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(40) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.
(41) Sauer, W. H.; Schwarz, M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987−1003.
(42) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52*, 6752−6756.
(43) Ritchie, T. J.; Macdonald, S. J.; Young, R. J.; Pickett, S. D. The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discovery Today* **2011**, *16*, 164−171.
(44) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787−18792.
(45) Clemons, P. A.; Wilson, J. A.; Dancik, V.; Muller, S.; Carrinski, H. A.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6817−6822.
(46) Sadowski, J.; Gasteiger, J. From atoms and bonds to 3-dimensional atomic coordinates - automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(47) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.