JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

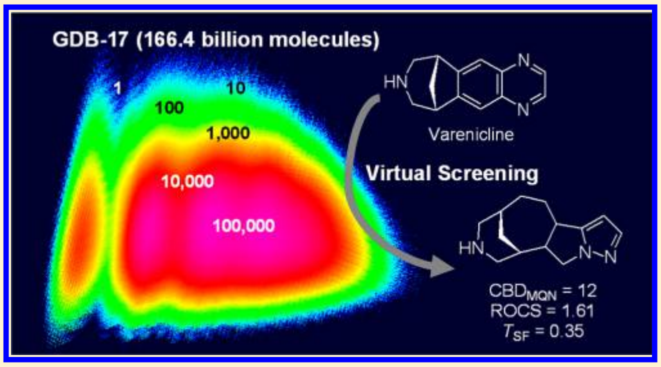
Visualization and Virtual Screening of the Chemical Universe Database GDB-17

Lars Ruddigkeit, Lorenz C. Blum, and Jean-Louis Reymond*

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

S *Supporting Information*

ABSTRACT: The chemical universe database GDB-17 contains 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens obeying rules for chemical stability, synthetic feasibility, and medicinal chemistry. GDB-17 was analyzed using 42 integer value descriptors of molecular structure which we term "Molecular Quantum Numbers" (MQN). Principal component analysis and representation of the (PC1, PC2)-plane provided a graphical overview of the GDB-17 chemical space. Rapid ligand-based virtual screening (LBVS) of GDB-17 using the city-block distance $\mathrm{CBD}_{\mathrm{MQN}}$ as a similarity search measure was enabled by a hashed MQN-fingerprint. LBVS of the entire GDB-17 and of selected subsets identified shape similar, scaffold hopping analogs (ROCS > 1.6 and $T_{\mathrm{SF}} < 0.5$) of 15 drugs. Over 97% of these analogs occurred within $\mathrm{CBD}_{\mathrm{MQN}} \leq 12$ from each drug, a constraint which might help focus advanced virtual screening. An MQN-searchable 50 million subset of GDB-17 is publicly available at www.gdb.unibe.ch.



INTRODUCTION

The drug-like chemical space, comprising organic small molecules (MW < 500 Da) of intermediate polarity,[1] is very large and might contain on the order of 10^{60} possible molecules.[2,3] De novo drug design consists in exploring this chemical space in search for bioactive compounds by generating and scoring virtual molecules prior to their synthesis,[4−7] whereby scoring relies on a variety of virtual screening (VS) approaches such as ligand-based and target-based methods.[8−10] The advantage of de novo drug design is that many more molecules can be evaluated than what is possible if considering only already synthesized molecules, a strategy which might help to uncover new molecular series and eventually improve clinical success for new drugs.[11−15]

Most de novo drug design approaches use stochastic methods to generate virtual molecules, such as genetic algorithms consisting of iterative cycles of virtual molecule generation and VS. Recently we proposed that de novo drug design might also be possible by first exhaustively enumerating the entire chemical space in the form of a database of virtual molecules, which can then be searched by VS to identify compounds for synthesis and testing. The advantage of this approach is that fundamental insights can be gained on the composition of the entire chemical space and that an exhaustive search can be performed that might in principle identify the best possible VS hit. Our initial studies focused on the chemical universe databases GDB-11 listing 26.4 million molecules up to 11 atoms of C, N, O, F,[16,17] and GDB-13 listing 977 million molecules up to 13 atoms of C, N, O, Cl, and S.[18] Although GDB-11 and GDB-13 contain only fragment-size molecules,[19] successful examples of ligand discovery by VS, synthesis, and testing were reported for both databases.[20−26] To expand this approach to larger molecules, we recently enumerated the chemical universe database GDB-17, which contains 166.4 billion molecules up to 17 atoms of C, N, O, S, and halogens obeying simple rules for chemical stability, synthetic feasibility, and medicinal chemistry.[27] GDB-17 reaches into molecular sizes compatible with many drugs (367 approved drugs ≤ 17 atoms) and typical for lead compounds (100 < MW < 350 Da).[28] In comparison to the database PubChem up to 17 atoms,[29] GDB-17 contains fewer aromatic and acyclic molecules but many more nonaromatic heterocycles and stereochemically rich and 3D-shaped molecules, and therefore should be particularly favorable for ″escaping out of flatland″ in drug discovery.[30−32]

Herein we report the visualization and ligand-based virtual screening (LBVS) of GDB-17 in MQN-space, a property space[33] defined by 42 integer value descriptors of molecular structure which we term "Molecular Quantum Numbers" (MQN, Table 1).[34] Similarly to previous examples with PubChem,[35−37] GDB-13,[38] and DrugBank,[39] maps of the GDB-17 chemical space were produced by principal component analysis (PCA) of the MQN data set and color-coded representation of the (PC1, PC2)-plane, showing that the chemical space covered by GDB-17 is broader than the

Received: November 8, 2012
Published: December 23, 2012

Table 1. 42 Molecular Quantum Numbers

atom counts (12)		bond counts (7)	
c	carbon	asb	acyclic single bonds
f	fluorine	adb	acyclic double bonds
cl	chlorine	atb	acyclic triple bonds
br	bromine	csb	cyclic single bonds
i	iodine	cdb	cyclic double bonds
s	sulfur	ctb	cyclic triple bonds
p	phosphorus	rbc	rotatable bond count
an	acyclic nitrogen		
cn	cyclic nitrogen		
ao	acyclic oxygen		
co	cyclic oxygen		
hac	heavy atom count		
polarity counts[a] (6)		**topology counts[b] (17)**	
hbam	H-bond acceptor sites	asv	acyclic monovalent nodes
hba	H-bond acceptor atoms	adv	acyclic divalent nodes
hbdm	H-bond donor sites	atv	acyclic trivalent nodes
hbd	H-bond donor atoms	aqv	acyclic tetravalent nodes
neg	negative charges	cdv	cyclic divalent nodes
pos	positive charges	ctv	cyclic trivalent nodes
		cqv	cyclic tetravalent nodes
		r3	3-membered rings
		r4	4-membered rings
		r5	5-membered rings
		r6	6-membered rings
		r7	7-membered rings
		r8	8-membered rings
		r9	9-membered rings
		rg10	≥10 membered rings
		afr	atoms shared by fused rings
		bfr	bonds shared by fused rings

[a] Polarity counts consider the ionization state predicted for the physiological pH = 7.4. hbam counts lone pairs on H-bond acceptor atoms, and hbdm counts H-atoms on H-bond donating atoms. [b] All topology counts refer to the smallest set of smallest rings. afr and bfr count atoms and bonds, repectively, shared by at least two rings.

known chemical space up to 17 atoms as covered by PubChem,[29] ChEMBL,[40] or DrugBank.[41] Efficient LBVS of GDB-17 by city-block distance $\mathrm{CBD_{MQN}}$ as a similarity measure was enabled by a hashed MQN-fingerprint following general concepts of fast database search strategies.[42−49] LBVS for analogs of fifteen approved and marketed drugs of 14 to 17 atoms was performed by searching for nearest neighbors of these drugs in MQN-space in the entire GDB-17 and in selected subsets. In agreement with previous studies with GDB-13,[24,38] searching using the city-block distance $\mathrm{CBD_{MQN}}$ as similarity measure was extremely fast and compatible with the large size of GDB-17. $\mathrm{CBD_{MQN}}$-similarity searching pointed to scaffold-hopping[50] analogs having high shape similarity as measured by the OpenEye scoring function ROCS (Rapid Overlay of Chemical Structure, threshold value ROCS > 1.6),[51] and low substructure similarity as defined by the Tanimoto similarity coefficient of a 1024-bit Daylight type substructure fingerprint (threshold value $T_{\mathrm{SF}} < 0.5$). Over 97% of all shape similar analogs occurred within the distance $\mathrm{CBD_{MQN}} \leq 12$ from the reference drug. This $\mathrm{CBD_{MQN}}$ distance constraint can be used to narrow down VS from 166.4 billion to a limited set of 0.25−55 million MQN-neighbors of each drug on which to apply more sophisticated scoring functions such as ROCS. To the best of our knowledge, the present study represents the first documented example of virtual screening applied to more than 100 billion molecules.

■ RESULTS AND DISCUSSION

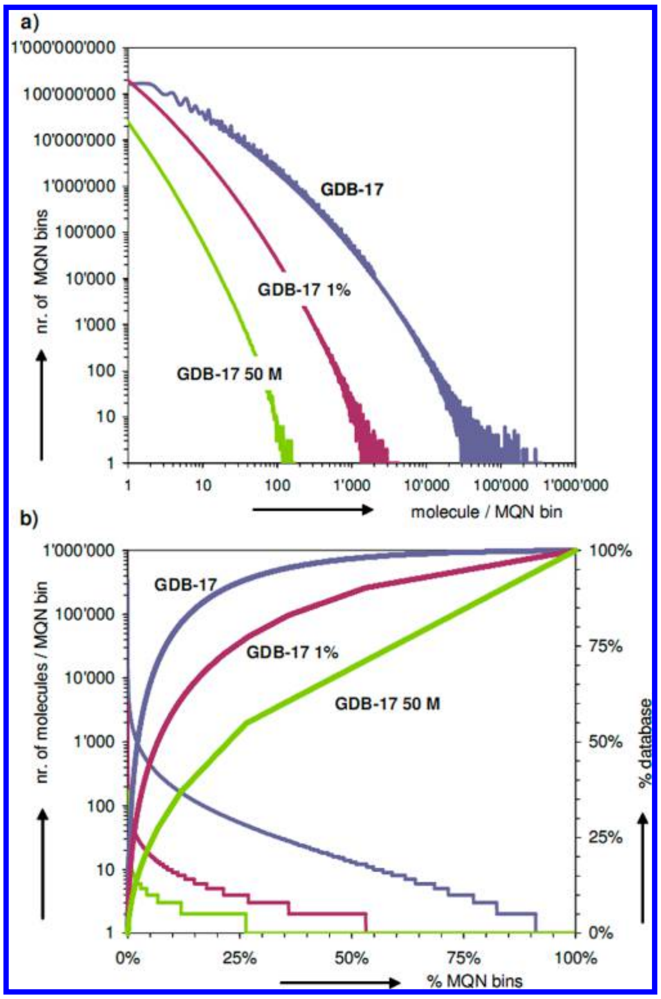


Figure 1. Distribution of GDB-17 and a 1% and 50 M random subset in MQN-bins. (A) Number of occupied MQN-bins as a function of the number of molecules per MQN-bin. (B) MQN-bin occupancy (lower curves, left axis) and database coverage by MQN-bins (upper curve, right axis) as a function of the percentage of MQN-bins ordered by decreasing bin size.

MQN-Space. The 42 MQN were determined as described previously.[34] The MQN-annotation of GDB-17 succeeded for 166 359 433 957 of the 166 369 433 581 SMILES in GDB-17 (99.994%) and consumed approximately 630 000 CPU h (with 2 GB RAM/CPU), which is 6-fold longer than for the database generation itself. The 166.4 billion molecules in GDB-17 were distributed in 1 875 717 681 different MQN-value combinations, here called MQN-bins, following an exponential distribution spanning from 351 905 molecules in the most occupied MQN-bin to 167 285 690 MQN-bins containing only a single molecule (Figure 1A). In total 50% of the database was found in the first 2.19% of the MQN-bins, corresponding to an average of 142 000 molecules per MQN-bin in the most densely populated part of the GDB-17 MQN-space (Figure 1B). The MQN-bin occupancy was lower for the 1% and 50 M random subsets of GDB-17, with fewer high-occupancy bins due to the only partial coverage of these smaller subsets.

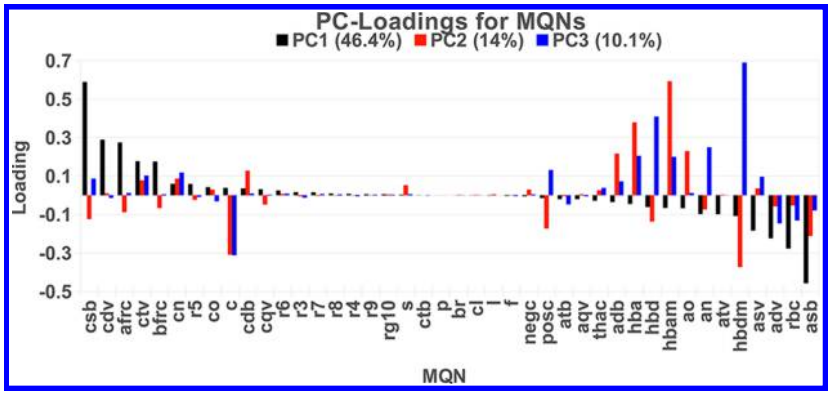


Figure 2. PC loadings of the GDB-17 MQN-space. See Table 1 for the definition of the MQN descriptors.

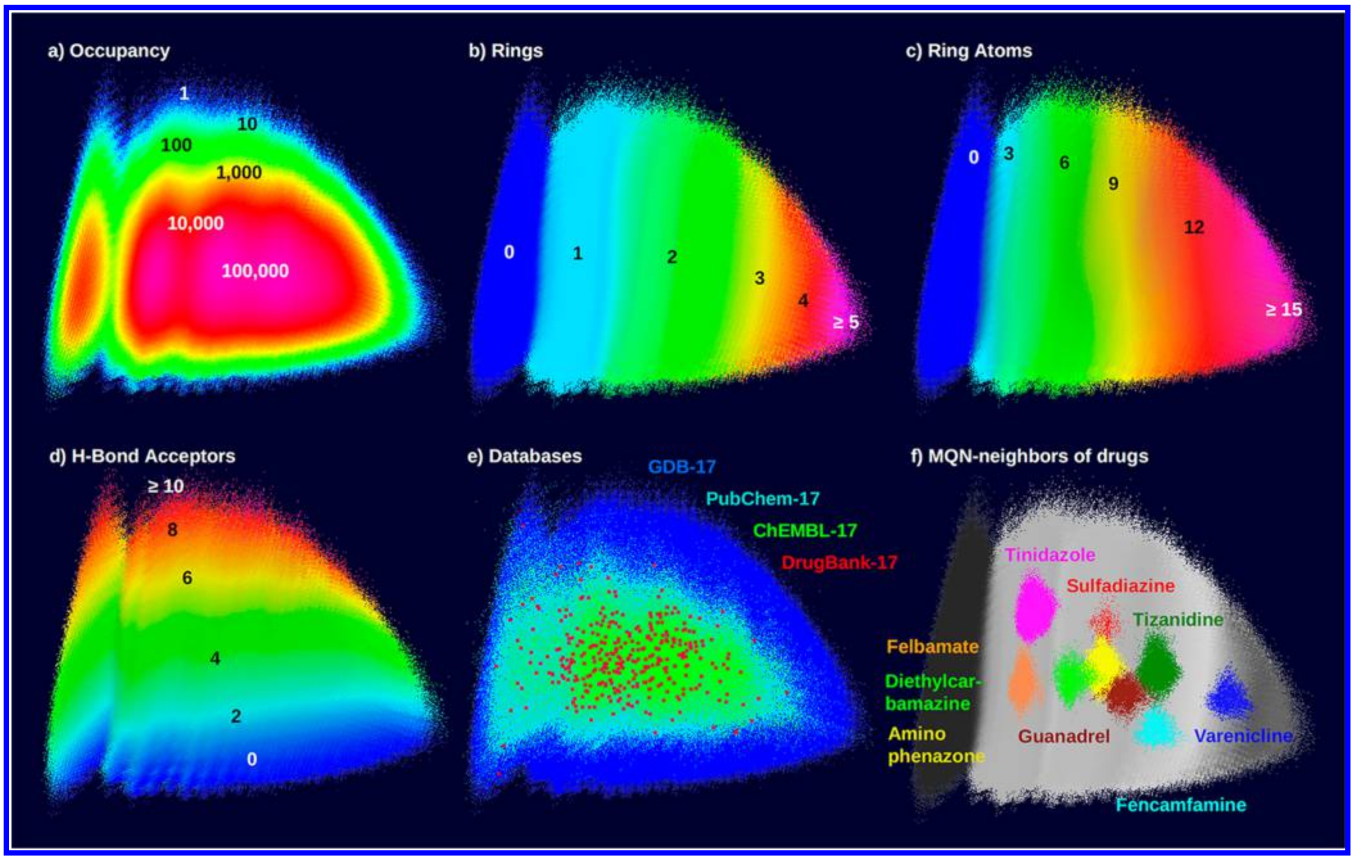


Figure 3. Color-coded MQN-maps of a 1% random subset of GDB-17 in the (PC1, PC2)-plane. (a) Occupancy heat map. (b) Number of cycles per molecules. (c) Number of cyclic atoms per molecule. (d) Number of H-bond acceptor atoms per molecule. (e) Color coding of pixels by database occupancy by priority DrugBank-17 > ChEMBL-17 > PubChem-17 > GDB-17. (f) MQN-Neighbors of selected drugs up to $\mathrm{CBD_{MQN}} \leq 12$. The structural formula of the drugs are shown in Figure 6. The background map is a greyscale rendering of part b.

Overall the distribution of the molecules of GDB-17 in MQN-bins corresponds to a power law distribution, which was comparably observed in previous MQN-analyses of GDB-11,[34] PubChem,[35−37] and GDB-13.[38]

MQN-Maps. Principal component analysis (PCA) of the MQN data set was carried out for a 1.6 billion (1%) subset of GDB-17. As shown by the PC-loadings, PC1 covered 46.4% of the variance and separated molecules according to cyclic features (Figure 2) such as cyclic single bonds (csb) and acyclic single bonds (asb). PC2 and PC3 covered 14% and 10.1%, respectively, of the variance and separated molecules according to polarity descriptors such as H-bond donors, H-bond acceptors, and carbon atoms. The variance covered, and the loadings of the first three PCs were comparable to those observed previously with GDB-11[34] and GDB-13,[38] reflecting the similar enumeration strategy used to generate these databases. MQN-maps were created by representing the (PC1,PC2)-plane as a 500 × 500 pixel image. Each pixel was color-coded according to the number of molecules in that pixel (occupancy heat map, Figure 3A) or by the average value of selected descriptors for the molecules in that pixel. These MQN-maps spread the GDB-17 molecules from west to east according to structural rigidity (e.g., increasing number of cycles and cyclic atoms, Figure 3B/C) and from south to north according to increasing polarity (e.g., increasing numbers of H-bond acceptor atoms, Figure 3D), providing images very

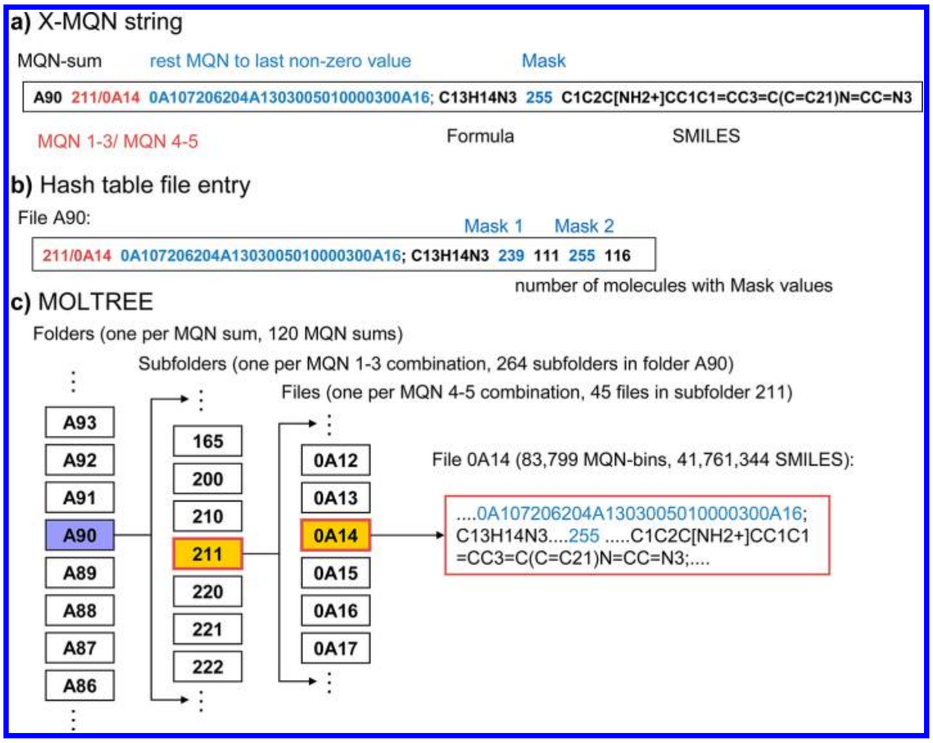

a) X-MQN string
MQN-sum
rest MQN to last non-zero value
Mask
A90 211/0A14 0A107206204A1303005010000300A16; C13H14N3 255 C1C2C[NH2+]CC1C1=CC3=C(C=C21)N=CC=N3
MQN 1-3/ MQN 4-5
Formula
SMILES
b) Hash table file entry
File A90:
Mask 1
Mask 2
211/0A14 0A107206204A1303005010000300A16; C13H14N3 239 111 255 116
number of molecules with Mask values
c) MOLTREE
Folders (one per MQN sum, 120 MQN sums)
Subfolders (one per MQN 1-3 combination, 264 subfolders in folder A90)
Files (one per MQN 4-5 combination, 45 files in subfolder 211)
A93
A92
A91
A90
A89
A88
A87
A86
165
200
210
211
220
221
222
0A12
0A13
0A14
0A15
0A16
0A17
File 0A14 (83,799 MQN-bins, 41,761,344 SMILES):
....0A107206204A1303005010000300A16;
C13H14N3....255 .....C1C2C[NH2+]CC1C1
=CC3=C(C=C21)N=CC=N3;....

Figure 4. X-MQN system exemplified with the drug varenicline. (a) X-MQN string. (b) Hash table file entry for the database. Varenicline is one of 116 GDB-17 molecules with identical X-MQN. (c) MOLTREE file structure. For compression, MQN values are assumed to be one digit numbers, two-digit numbers when preceded by A, and 3-digit numbers when preceded by B. The MQN-values are ordered starting with four preselection criteria followed by the remaining MQN in order of decreasing variance in GDB-13 as follows: hba-hbd-pos/neg-csb asb-cdv-afr-hbdm-adv-ctv-hbam-rbc-bfr-c-asv-cn-adb-an-cdb-ao-r5-atv-cqv-co-r3-r6-atb-r4-hac-r7-s-r8-aqv-r9-ctb-rg10-cl-I-br-f-p (see Table 1 for definitions). For compression, MQN are written only up to the last nonzero value. The mask is an integer encoding eight bits true (1) or false (0) for eight filtering criteria (see main text).

comparable to those obtained previously with GDB-11[34] and GDB-13,[38] reflecting again the fact that these databases were generated by similar enumeration principles. The occupancy heat map showed that GDB-17 contained the largest number of molecules in the area of molecules with 1−3 rings, 4−10 ring atoms, and 3−5 H-bond acceptor atoms. Mapping of the reference databases PubChem-17, ChEMBL-17, and DrugBank-17 on the MQN map of GDB-17 showed that the mostly densely occupied region of the GDB-17 MQN-map overlaps with the area covered by these databases of known compounds; however, GDB-17 spread over a significantly broader area (Figure 3E). Mapping of MQN-nearest neighbors of typical drugs showed that these drugs spread over most of the MQN-map (Figure 3F).

X-MQN System. One of the key features of MQN-space is that MQN-similarity searching by $\mathrm{CBD_{MQN}}$ allows us to identify analogs of known drugs with similar molecular shape and bioactivities but quite different substructures in retrospective[35,36,38] or prospective[24] studies. The $\mathrm{CBD_{MQN}}$ is one of the simplest similarity measures to compare fingerprints[52] and consists of the sum of the absolute differences between MQN pairs across the 42 MQN. The search for MQN-nearest neighbors in the 42-dimensional MQN-space of any query molecule can be performed very rapidly with a preorganized database using an extended MQN-fingerprint (X-MQN). X-MQN is a hashed MQN-fingerprint also allowing additional search criteria beyond $\mathrm{CBD_{MQN}}$ nearest neighbor searching.[38] The X-MQN string lists the sum of all 42 MQN values (MQN sum as hash function), followed by the 42 MQN values themselves in an optimized order, the elemental formula, a mask value, and the SMILES (Figure 4A and legend). The mask is an integer encoding eight bits true (1) or false (0) for the following filtering criteria: (1) no acyclic carbon ("scaffold-like");[38] (2) follows Congreve's fragment-likeness principles (the so-called "Rule of 3");[53] (3) no 3- or 4-membered ring; (4) no acyclic CC unsaturation; (5) no cyclic CC unsaturation; (6) no aldehydes, epoxides, aziridines, esters, carbonates, or sulfonic esters; (7) no acyclic heteroatom−heteroatom bonds; (8) no cyclic heteroatom−heteroatom bonds. To enable LBVS, a hash table is created as a series of separate files, one for each MQN sum, listing the ordered MQN values, the elemental formula, and the mask value with the number of molecules per X-MQN rather than the actual SMILES (Figure 4B). The database itself is organized by grouping molecules with similar X-MQN in the MOLTREE, which is an organized structure of folders, subfolders, and files mirroring the hash table and containing the SMILES (Figure 4C).

LBVS for MQN-nearest neighbors of any query molecule is extremely fast because to find all molecules within $\mathrm{CBD_{MQN}} \leq N$, one only needs to search in folders with MQN-sum difference $\Delta_{\mathrm{MQNsum}} \leq N$, subfolders with partial $\mathrm{CBD_{MQN}}$ for the first three MQN $\mathrm{CBD_{MQN(1-3)}} \leq N$, and files with partial $\mathrm{CBD_{MQN}}$ for the fourth and fifth MQNs $\mathrm{CBD_{MQN(4-5)}} \leq (N - \mathrm{CBD_{MQN(1-3)}})$. Within each file, the $\mathrm{CBD_{MQN}}$ for the rest MQN values must fufill $\mathrm{CBD_{restMQN}} \leq (N - \mathrm{CBD_{MQN(1-3)}} - \mathrm{CBD_{MQN(4-5)}})$. For example all molecules with $\mathrm{CBD_{MQN}} = 0$ to a given query molecule must have the same MQN-sum defining a single folder, and the same combination of the first five MQN defining a single subfolders and a single file within this subfolder.

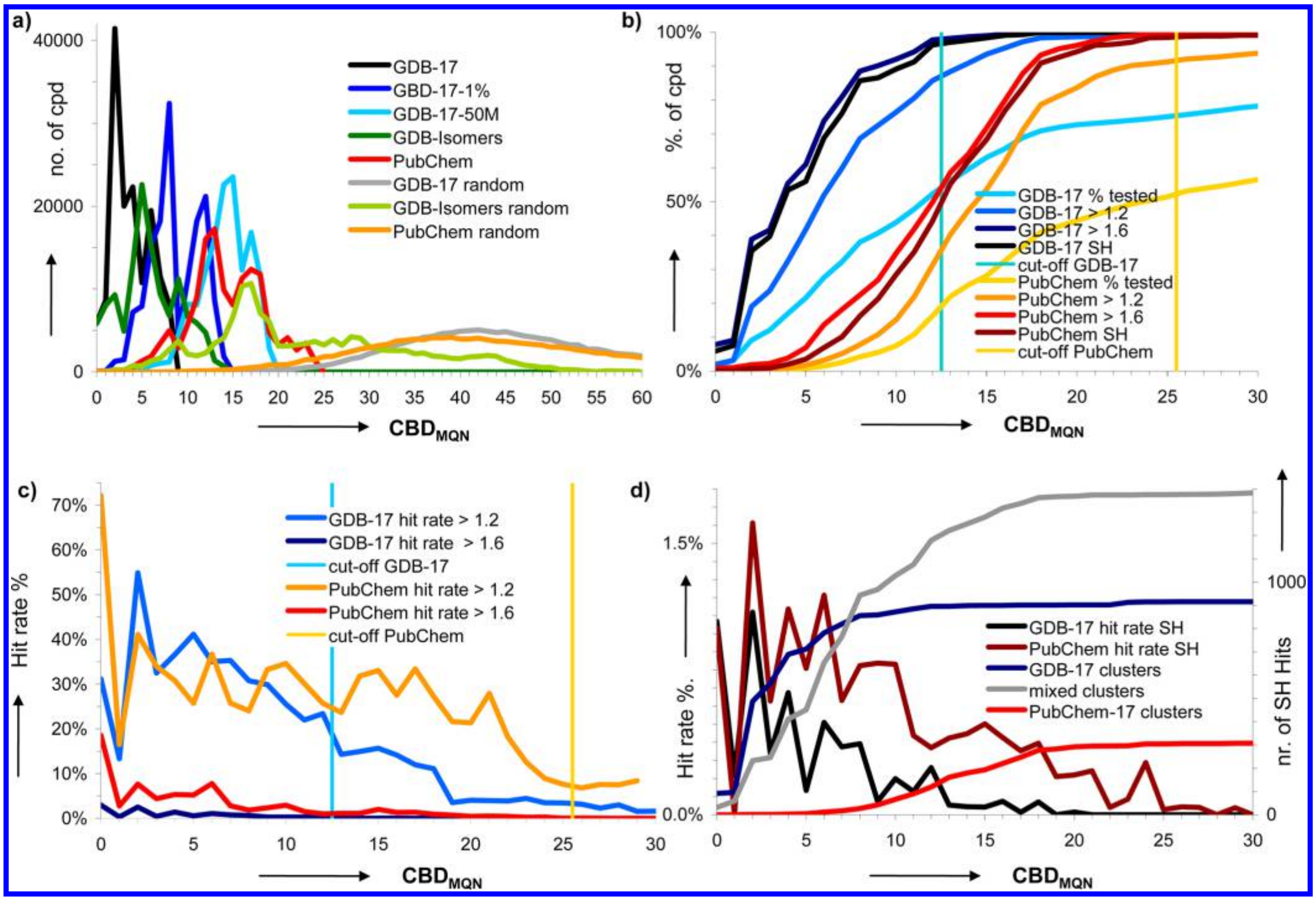


Figure 5. VS of GDB-17 and PubChem for shape similar, scaffold-hopping drug analogs. (A) Number of molecules as function of CBD$_{MQN}$ to the reference drug, cumulated over all 15 drugs, for the following sets of (15×) 10 000 molecules: CBD$_{MQN}$-nearest neighbors in GDB-17 (black), its 1%- (blue) and 50M-subset (cyan), the drug isomers (excluding cpds with nonaromatic CC unsaturations, dark green), and PubChem (red); 10 000 molecules selected randomly from GDB-17 (gray), the drug isomers (light green), and PubChem (orange). (B) Percentage of molecules as a function of CBD$_{MQN}$ to the reference drug, cumulated over all 15 drugs, for the following sets: all molecules subjected to ROCS evaluation in GDB-17 (cyan) and PubChem (yellow), molecules with shape similarity to their reference drug ROCS > 1.2 in GDB-17 (blue) and PubChem (orange), ROCS > 1.6 in GDB-17 (dark blue) and PubChem (red), and scaffold-hopping hits (ROCS > 1.6 AND T_{SF} < 0.5) in GDB-17 (black) and PubChem (brown). (C) Hit rates for ROCS > 1.2 and ROCS > 1.6. The cutoff lines indicate the recommended CBD$_{MQN}$ to retrieve the majority of ROCS > 1.6 and scaffold-hopping hits from GDB-17 (cyan, CBD$_{MQN}$ ≤ 12) and PubChem (yellow, CBD$_{MQN}$ ≤ 25). (D) Scaffold hopping hit rates (left *y*-scale) for GDB-17 (black line) and PubChem-17 (brown line) and number of SH hits (right *y*-scale) belonging to GDB-17 only clusters (blue line), PubChem only clusters (red line), and mixed clusters (gray line).

The search for MQN-nearest neighbors of a query molecule consists in calculating its X-MQN string and searching the hash table files in the order of increasing difference in MQN-sum until all hash table file entries have been found up to a preset CBD$_{MQN}$ distance to the query molecule or a preset total number of compounds. Optionally, one can select for compounds retaining the number of H-bond acceptor atoms (fixed first MQN), H-bond donor atoms (fixed second MQN), positive charges (fixed third MQN), or negative charges (fixed fourth MQN) at neutral pH, or the elemental formula (isomer search) of the query molecule. The mask value can be used to select molecules fulfilling up to eight criteria (see above and legend of Figure 4). The organized look-up of the hash table accelerates the search by a factor of approximately 5000-fold compared to a simple linear search (calculating the CBD$_{MQN}$ of all GDB-17 molecules to the query using the precalculated fingerprint values). Once all hash table file entries corresponding to the search have been identified, the molecules are retrieved by copying the corresponding SMILES from the MOLTREE files.

The organization of GDB-17 in a searchable X-MQN system was computationally quite intensive and required approximately 760 000 h CPU time using 16 GB RAM per CPU. The same operation was carried out for a 1% random subset (GDB-17−1%) and a 50 million random subset (GDB-17−50M) of GDB-17. Once organized, GDB-17 and its subsets could be searched efficiently for MQN-similarity. The identification of the first 50 million CBD$_{MQN}$-neighbors of a query molecule in the hash table of the entire GDB-17 (166.4 billion molecules) required approximately 15 min on a single hexacore desktop machine. Copying the molecule SMILES from the MOLTREE files to an output file required an additional 30 min. Searches on the 1% and 50 million random subsets were shorter than 1 min.

Virtual Screening. As a proof-of-concept study for LBVS in the entire GDB-17 by MQN-similarity, we set out to search for shape similar, scaffold-hopping analogs of 15 known drugs of 14−17 atoms selected from the marketed drugs in that size range listed in DrugBank. Molecular shape is known to correlate strongly with biological activity.[54] Shape similar analogs of existing drugs with very different substructures

Table 2. Virtual Hits Identified in GDB-17 and PubChem

	GDB-17			PubChem		
drug	ROCS > 1.2	ROCS > 1.6	scaffold hopping	ROCS > 1.2	ROCS > 1.6	scaffold hopping
aciclovir	14966	30	27	2666	312	53
aminoglutethimide	5425	53	10	1055	30	0
aminophenazone	17512	436	427	4354	135	70
dexmedetomidine	14232	505	113	3016	30	6
diethylcarbamazine	5801	101	45	4220	132	105
ethoxzolamide	11791	147	109	2266	102	60
felbamate	12085	176	88	1430	41	22
fencamfamine	12767	426	158	4803	162	67
guanadrel	14960	608	262	870	8	7
procaine	5921	66	29	1858	93	40
sulfadiazine	5167	182	61	3644	473	97
tinidazole	12561	172	101	1758	47	1
tizanidine	11260	475	318	1584	66	26
trioxsalen	26634	920	261	6679	368	70
varenicline	13909	205	135	3827	98	26
total	184991	4502	2144	44030	2097	650

("scaffold-hopping" analogs)[50] are highly relevant because they may lead to new molecular series with similar activity on the target but possible improvement on pharmacology and ADMET properties.

LBVS was carried out by extracting the 10 000 MQN-nearest neighbors of each drug from the following five groups: (1) the entire GDB-17, (2) the 1.6 billion random subset of GDB-17, (3) the 50 million random subset of GDB-17, (4) all isomers of each drug found in GDB-17 (isomers: compounds with the same molecular formula but different structural formulas), (5) PubChem considering molecules up to 17 atoms (PubChem-17). In addition, three groups of 10 000 randomly selected molecules were extracted from the following series: (6) GDB-17, (7) the isomers of each drug in GDB-17, (8) PubChem-17. In the case of GDB-17, the search excluded molecules with nonaromatic carbon−carbon double bonds to avoid trivial solutions to the scaffold-hopping problem consisting in drug analogs with unsaturated equivalents of saturated carbon chains. Indeed such unsaturated analogs often show high shape similarity to the parent drug and low substructure similarity, but lack innovative character.

Each of the above eight groups of molecules (six groups for GDB-17, two groups for PubChem) represented different selections probing MQN-space at different CBD_{MQN}-distances from the reference drug (Figure 5A). For each of the fifteen drugs, each of the above eight groups of 10 000 molecules was scored for shape-similarity by calculating the ROCS score (Rapid Overlay of Chemical Structure) to the parent drug. ROCS is a relatively slow but well-validated shape similarity function.[51] The Tanimoto similarity coefficient of a 1024 bit Daylight-type substructure fingerprint (T_{SF}) was also calculated in each of these 10 000 molecule sets as a measure of substructure similarity (Supporting Information Figure S1−S8).[52,55]

Three categories of LBVS hits were defined as follows: (1) all molecules with ROCS > 1.2, which corresponds to a broad criterion of shape similarity;[56] (2) all molecules with ROCS > 1.6, defined here as a strong selection criterion for high degree of shape similarity to the parent drug; (3) ROCS > 1.6 AND $T_{SF} < 0.5$, used here to define scaffold-hopping, highly shape similar analogs of the parent drugs. From a total of 900 000 ROCS comparisons from the various GDB-17 subsets, 184 991 molecules (20.6%) showed moderate shape similarity to one of the reference drugs (ROCS > 1.2), 4502 molecules (0.50%) showed strong shape similarity (ROCS > 1.6), and 2144 (0.24%) showed scaffold-hopping properties (ROCS > 1.6 AND $T_{SF} < 0.5$). The hit rates were comparable to those obtained for the 300 000 ROCS comparisons performed with PubChem molecules, which delivered 44 030 molecules (14.7%) with ROCS > 1.2, 2097 molecules (0.70%) with ROCS > 1.6, and 650 molecules (0.22%) with ROCS > 1.6 AND $T_{SF} < 0.5$ (Table 2). This showed that GDB-17 was of comparable value to PubChem for delivering shape-similarity analogs of the drugs, with the added value that almost all GDB-17 molecules are previously unknown.

Although the various sets of 10 000 compounds sampled GDB-17 and PubChem at various CBD_{MQN}-distances from each drug, the vast majority of the LBVS hits occurred in relatively close proximity to the drugs (Figure 5A/B). For GDB-17, 85.7% of the ROCS > 1.2 hits, 97.7% of the ROCS > 1.6 hits, and 96.2% of scaffold-hopping hits were found at $CBD_{MQN} \leq 12$ from each drug. The hit rates were highest at $CBD_{MQN} \leq 2$ to the reference drug and sharply decreased with increasing CBD_{MQN}-distance (Figure 5C/D). The cutoff value $CBD_{MQN} \leq 12$ might be used as a fast measure to narrow down VS from the entire GDB-17 to a much smaller number of molecules. For instance only 0.25 to 55 million molecules are found in GDB-17 within $CBD_{MQN} \leq 12$ from each drug. The hit-rate found within the 460 000 molecules within $CBD_{MQN} \leq 12$ and scored by ROCS was quite significant: ROCS > 1.2 36%; ROCS > 1.6 1.1%; scaffold-hopping 0.48%. Extrapolating from these numbers, one might obtain hundreds of thousands to several millions of VS hits if scoring all GDB-17 molecules within $CBD_{MQN} \leq 12$.

Compared to GDB-17, the LBVS with PubChem showed a higher distance cutoff value $CBD_{MQN} \leq 25$ to identify 91.3% of the ROCS > 1.2 hits, 99.4% of the ROCS > 1.6 hits, and 98.5% of the scaffold-hopping hits. Indeed PubChem contained only very few molecules within the $CBD_{MQN} \leq 12$ cutoff where most GDB-17 hits were found, and the PubChem MQN-nearest neighbor sets occupied mostly the range $10 \leq CBD_{MQN} \leq 20$. The hit rates were higher in PubChem than in GDB-17 for ROCS > 1.2 and ROCS > 1.6 (Figure 5C); however, the scaffold-hopping hit rates were comparable in both databases

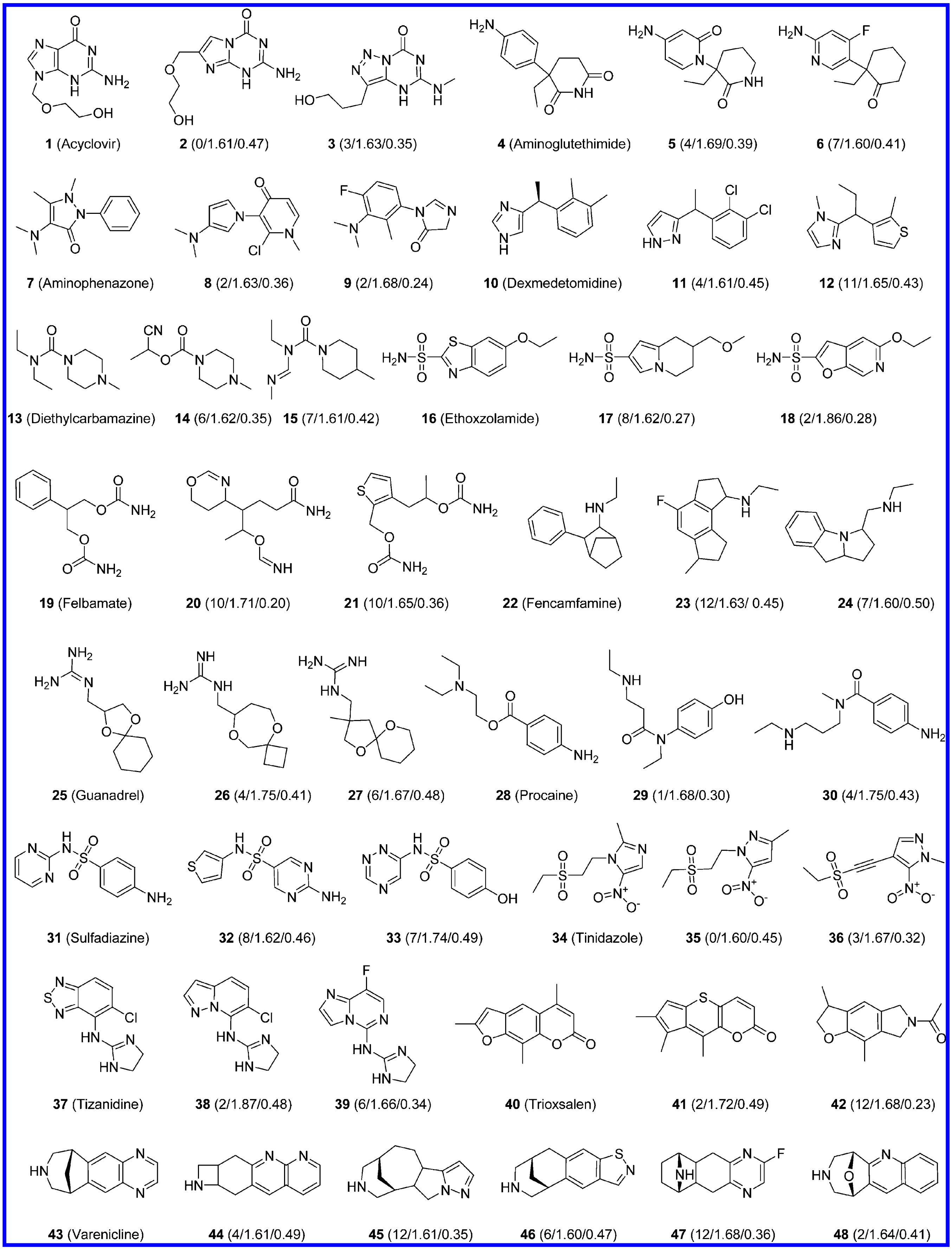


Figure 6. Structures of the 15 drugs used for VS in alphabetical order, accompanied with two examples each of scaffold-hopping VS hits from GDB-17. The drug name or the scores $\mathrm{CBD}_{\mathrm{MQN}}/\mathrm{ROCS}/T_{\mathrm{SF}}$ to the parent drug are given in parentheses next to the compound number.

(Figure 5D). This suggests that PubChem contains mostly substructure analogs of the 15 drugs studied here, while GDB-17 offers mostly yet unexplored shape-similar analog with very different substructures (scaffold-hopping).

The LBVS hits were characterized closer with respect to their structural diversity. The 2781 scaffold-hopping virtual hits from GDB-17 and PubChem were combined and clustered by affinity propagation considering T_{SF} as similarity measure, which yielded 383 different clusters. There were 132 clusters containing only GDB-17 molecules, which were almost exclusively located at $\mathrm{CBD_{MQN}} \leq 12$ to their reference drug. There were 59 clusters containing only PubChem molecules, which were located at $8 < \mathrm{CBD_{MQN}} < 25$ to their reference drugs. Twenty of these PubChem only clusters contained molecules with functional groups that are not enumerated in GDB-17 (e.g., aliphatic halogens, thiols, thioethers, thioureas, aminals). The other 39 PubChem only clusters contained molecules which belong to GDB-17 but occur beyond the 10 000 $\mathrm{CBD_{MQN}}$ nearest neighbors of each drug, corresponding to a region that was only very partially sampled in the present study. Another 192 clusters were mixed (containing molecules from both GDB-17 and PubChem) and featured molecules covering the range $0 < \mathrm{CBD_{MQN}} < 20$ to their reference drug (Figure 5D). A selection of scaffold-hopping LBVS hits from GDB-17 stemming from various clusters is shown in Figure 6. Compared to the examples of shape similar (ROCS > 1.4), isomers of the same drugs shown in Figure 3 of ref 27, the present LBVS hits cover a range of different molecular formula, yet show substantially higher shape similarity to the parent drugs. As for the previous isomer examples,[27] the LBVS hits in Figure 6 are all yet unknown as by Scifinder search. LBVS scaffold-hopping hits of another 15 drugs up to 13 atoms had been similarly identified by MQN-similarity searching in GDB-13,[38] showing the robustness of the approach.

The scaffold-hopping hit list contained heterocyclic analogs of the parent drugs with identical graphs (**18**, **48**) including isosteric replacements of substituents (**11**, **33**, **35**, **38**, **41**), or analogs with regioisomeric attachment points of similar substituents (**2**, **5**, **8**, **39**). Ring size analogs were often found, including the exchange of 5- and 6-membered rings (**12**, **21**, **32**, **46**, **47**), and other ring sizes (**26**, **44**, **45**). In several cases saturated carbo- or heterocycles replaced aromatic rings (**17**, **20**, **42**, **47**). Some scaffold-hopping hits featured completely different graphs compared to the parent drug, for example the fencamfamine analogs **23** and **24**. Such hits could represent the true value of shape-based virtual screening in GDB-17 to identify new molecular series.

■ CONCLUSION

The data above describe the first example of VS applied to more than 100 billion molecules. The 166.4 billion molecules in the chemical universe database GDB-17 were classified using the 42 MQN descriptors, allowing visualization of the entire database in the form of color-coded MQN-maps. Rapid LBVS of GDB-17 using the city-block distance $\mathrm{CBD_{MQN}}$ as a similarity search measure was enabled by a hashed MQN-fingerprint. GDB-17 and selected subsets of this database were sampled at various $\mathrm{CBD_{MQN}}$ distances from 15 reference drugs by extracting sets of 10 000 compounds at random or sorted by $\mathrm{CBD_{MQN}}$. In total 900 000 GDB-17 molecules were evaluated for shape similarity to these drugs using ROCS, which revealed that almost all highly shape similar LBVS hits (ROCS > 1.6) occurred within $\mathrm{CBD_{MQN}} \leq 12$ from the reference drug. This criterion may be used as a first criterion to identify series with a tractable number of molecules on which to apply more advanced scoring functions such as ROCS or docking. This VS strategy is remarkably fast and should be generally useful to handle large databases.

■ METHODS

All code packages were written in Java 1.6 with Jchem Libaries from ChemAxon. All computations were parallelized on a 360-CPU cluster and manually controlled. To preserve disc space every output was compressed directly into gzip by the implementation of gzip into the GZIPStream in Java BufferedReader/Writer.

MQNs were calculated using the previously reported calculator source code[34] written in Java using the JChem library from ChemAxon, Ltd. Prior to MQN-calculation, the ionization state of each structure was adjusted to pH 7.4 using the Jchem API. PCA was done by using an in-house developed Java application using Jsci (http://jsci.sourceforge.net). The parallelized source code is based on the tutorial of Lindsay I. Smith (http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf).

X-MQN were calculated with a modified version of the published source code for MQN.[34] The MQNs and the elemental formula are calculated for the ionization state of the molecules at pH 7.4. The hash table files and MOLTREE were created for each MQN sum separately. The computation required unusually high RAM and was performed with a modified cluster assigning 16 GB RAM/CPU. For some of the highly occupied MQN sums, the calculation was splitted into individual subfolders and subfiles to avoid memory overload.

For substructure similarity calculation a Daylighttype 1024-bit hashed fingerprint from ChemAxon was used. For the ROCS calculations, the stereo information of the 15 reference drugs was added only if found in DrugBank. All queries and target moelcules were sent to Omega to create a maximum of 200 lowest energy 3D structures including various stereo-isomers and their conformers. For all ROCS runs the "TanimotoCombo" overlap score was used.

An MQN-searchable version of the 50 million random subset of GDB-17 is freely accessible at www.gdb.unibe.ch.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Scatter plots and histograms of T_{SF} and ROCS values obtained in virtual screening (Figure S1−S8). This information is available free of charge via the Internet at http://pubs.acs.org.

■ 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 and the Swiss National Science Foundation. We thank ChemAxon Ltd. and OpenEye Scientific Ltd. for the free academic licenses to their softwares.

REFERENCES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(3) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824−828.

(4) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.

(5) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **2009**, *42*, 724−733.

(6) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30−38.

(7) Hartenfeller, M.; Schneider, G. De novo drug design. *Methods Mol. Biol.* **2011**, *672*, 299−323.

(8) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.

(9) 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.

(10) 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.

(11) 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.

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

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

(14) Hann, M. M. Molecular obesity, potency and other addictions in drug discovery. *MedChemComm* **2011**, *2*, 349−355.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) Foloppe, N. The benefits of constructing leads from fragment hits. *Future Med. Chem.* **2011**, *3*, 1111−1115.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) 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.

(26) Reymond, J. L.; Awale, M. Exploring Chemical Space for Drug Discovery Using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.

(27) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.

(28) 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.

(29) 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.

(30) 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.

(31) 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.

(32) 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.

(33) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Persp. Drug Discovery Des.* **1998**, *9−11*, 339−353.

(34) Nguyen, K. T.; Blum, L. C.; van Deursen, R.; Reymond, J.-L. Classification of Organic Molecules by Molecular Quantum Numbers. *ChemMedChem* **2009**, *4*, 1803−1805.

(35) van Deursen, R.; Blum, L. C.; Reymond, J. L. A searchable map of PubChem. *J. Chem. Inf. Model.* **2010**, *50*, 1924−1934.

(36) van Deursen, R.; Blum, L. C.; Reymond, J. L. Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 649−662.

(37) Reymond, J. L.; Blum, L. C.; Van Deursen, R. Exploring the Chemical Space of Known and Unknown Organic Small Molecules at www.gdb.unibe.ch. *Chimia* **2011**, *65*, 863-867.

(38) 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.

(39) Awale, M.; Reymond, J. L. Cluster analysis of the DrugBank chemical space using molecular quantum numbers. *Bioorg. Med. Chem.* **2012**, *20*, 5372−5378.

(40) 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.

(41) 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.

(42) Wang, X.; Wang, J. T. L. Fast Similarity Search in Three-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 442−451.

(43) Dutta, D.; Guha, R.; Jurs, P. C.; Chen, T. Scalable Partitioning and Exploration of Chemical Spaces Using Geometric Hashing. *J. Chem. Inf. Model.* **2005**, *46*, 321−333.

(44) Swamidass, S. J.; Baldi, P. Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952−964.

(45) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098−2109.

(46) Nisius, B.; Vogt, M.; Bajorath, J. r. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback−Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347−1358.

(47) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(48) Nasr, R.; Hirschberg, D. S.; Baldi, P. Hashing Algorithms and Data Structures for Rapid Searches of Fingerprint Vectors. *J. Chem. Inf. Model.* **2010**, *50*, 1358−1368.

(49) Nasr, R.; Vernica, R.; Li, C.; Baldi, P. Speeding Up Chemical Searches Using the Inverted Index: The Convergence of Chemoinformatics and Text Search Methods. *J. Chem. Inf. Model.* **2012**, *52*, 891−900.

(50) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.

(51) Rush, T. S., 3rd; 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.

(52) Khalifa, A. A.; Haranczyk, M.; Holliday, J. Comparison of Nonbinary Similarity Coefficients for Similarity Searching, Clustering and Compound Selection. *J. Chem. Inf. Model.* **2009**, *49*, 1193−1201.

(53) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.

(54) 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.

(55) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(56) AbdulHameed, M. D.; Chaudhury, S.; Singh, N.; Sun, H.; Wallqvist, A.; Tawa, G. J. Exploring polypharmacology using a ROCS-based target fishing approach. *J. Chem. Inf. Model.* **2012**, *52*, 492−505.