J Comput Aided Mol Des (2011) 25:637–647
DOI 10.1007/s10822-011-9436-y

Visualisation and subsets of the chemical universe database GDB-13 for virtual screening

Lorenz C. Blum · Ruud van Deursen · Jean-Louis Reymond

Received: 22 February 2011 / Accepted: 13 May 2011 / Published online: 27 May 2011

Abstract The chemical universe database GDB-13, which enumerates 977 million organic molecules up to 13 atoms of C, N, O, S and Cl following simple chemical stability and synthetic feasibility rules, represents a vast reservoir for new fragments. GDB-13 was classified using the MQN-system discussed in the preceding paper for the analysis of PubChem fragments. Two hundred and fifty-five subsets of GDB-13 were generated by the combinatorial use of eight restrictive criteria, including fragment-like ("rule of three") and scaffold-like (no acyclic carbon atoms) filters. Virtual screening for analogs of 15 commercial drugs of 13 non-hydrogen atoms or less shows that retrieving MQN-neighbors of a query molecule from GDB-13 or its subsets provides on average a 38-fold enrichment in structural analogs (Daylight-type substructure fingerprint Tanimoto $T_{SF} > 0.7$), and a 75-fold enrichment in shape-similar analogs (ROCS TanimotoCombo score > 1.4). An MQN-searchable version of GDB-13 is provided at www.gdb.unibe.ch.

Keywords Databases · Virtual screening · Chemical space · Enumeration · Fragments

L. C. Blum · R. van Deursen · J.-L. Reymond (✉)
Department of Chemistry and Biochemistry, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland
e-mail: jean-louis.reymond@ioc.unibe.ch

L. C. Blum · R. van Deursen · J.-L. Reymond
Swiss National Center of Competence in Research, NCCR-TransCure, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland

Introduction

The discovery of innovative chemotypes is one of the key chemical problems in small molecule drug discovery, in particular at the level of fragments, a size range which also includes many drugs [1–8]. Beyond all fragment-sized molecules that are already known, such as those collected in the public access database PubChem [9] as discussed in the preceding paper in this issue [10], one might want to consider all molecules that could ever be possibly synthesized. Along these lines, we recently reported the enumeration of all molecules up to a size of 13 non-hydrogen atoms following predefined chemical stability and synthetic feasibility rules, which produced the chemical universe database GDB-13 containing 977 million virtual molecules of C, N, O, S, Cl [11]. This database was an extension of a previous version GDB-11 containing 26.4 million virtual molecules up to 11 non-hydrogen atoms of C, N, O, F [12, 13], which was shown to provide a useful starting point for designing bioactive synthetic ligands [14–17]. GDB-13 exceeds the number of known molecules of similar size by several orders of magnitude and represents a vast and mostly unexploited reservoir for innovation [18].

The meaningful exploration of GDB-13 requires efficient virtual screening tools to identify compounds of biological interest for synthesis and testing. At present however such exploration is limited by the currently available virtual screening methods, which typically process at most a few million structures within resonable computing time. To address this limitation, we recently reported a molecule classification method for large databases called the MQN-system [19]. This system places organic molecules in a chemical space [18, 20–22] on the basis of 42 integer value descriptors for structural and topological features, called MQNs (Table 1). The MQNs can be determined visually from the structural formula by anyone with basic knowledge in organic chemistry, such that MQN-space is readily accessible to non-specialists. The analysis produces meaningful overviews of large molecular databases such as ZINC [23] and PubChem [9] as color-coded maps derived from principal component analysis (PCA) of the MQN data [24]. Furthermore, the MQN-system computes very fast and performs remarkably well in virtual screening as exemplified previously for the enrichment of bioactives from the DUD dataset from the entire PubChem database [24, 25].

Herein we report the visualisation and an efficient virtual screening approach for the entire GDB-13 database based on the MQN-system. The database was subdivided into 255 subsets defined by the combinatorial use of eight different criteria limiting structural complexity and functional groups. MQN nearest-neighbour searches performed on the entire GDB-13 or on any of its subsets are shown to rapidly identify structural and shape analogs of any query molecule. Structural similarity between two compounds is typically measured by the substructure fingerprint Thus, scoring MQN-nearest neighbors of a query molecule by substructure fingerprint (SF) similarity (measured by the Tanimoto value T_{SF}) [26] or by shape similarity (measured by the ROCS TanimotoCombo score) [27] as indicators of bioactivity probability shows that MQN-nearest neighbors are strongly enriched in structural analogs and shape-similar analogs of the query molecule. An MQN-searchable version of GDB-13 is provided at www.gdb.unibe.ch, and should greatly facilitate the exploitation of GDB-13 for the identification of new medicinally relevant small molecules for synthesis and testing.

Table 1 The 42 molecular quantum numbers (MQNs)

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	Sulphur	ctb	Cyclic triple bonds
p	Phosphorous	rbc	Rotatable bond count
an	Acyclic nitrogen	*Topology counts[b] (17)*	
cn	Cyclic nitrogen	asv	Acyclic monovalent nodes
ao	Acyclic oxygen	adv	Acyclic divalent nodes
co	Cyclic oxygen	atv	Acyclic trivalent nodes
hac	Heavy atom count	aqv	Acyclic tetravalent nodes
Polarity counts[a] (6)		cdv	Cyclic divalent nodes
hbam	H-bond acceptor sites	ctv	Cyclic trivalent nodes
hba	H-bond acceptor atoms	cqv	Cyclic tetravalent nodes
hbdm	H-bond donor sites	r3	3-membered rings
hbd	H-bond donor atoms	r4	4-membered rings
neg	Negative charges	r5	5-membered rings
pos	Positive charges	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 repectively bonds shared by at least two rings

Results and discussion

Visualisation of GDB-13

The 42 MQN-values were obtained for 975,821,779 molecules in GDB-13, resulting in 37,249,813 different MQN-combinations. The most occupied MQN-bin contained 12,589 molecules, while 4,959,920 MQN-bins contained only one molecule (Fig. 1) [19]. Principal component analysis (PCA) was performed to gain an insight into the data structure. PC1 (51%) represented mostly structural rigidity, with strongly positive loadings in cyclic descriptors e.g. cyclic single bonds (csb) and strongly negative loadings in acyclic descriptors e.g. acyclic single bonds (asb). PC2 (12%) reflected H-bonding behaviour and polarity, with strong positive loadings in hydrogen bond acceptors (hba, hbam), and strongly negative loadings in hydrogen bond donors (hbdm) and carbon counts (c) (Fig. 2).

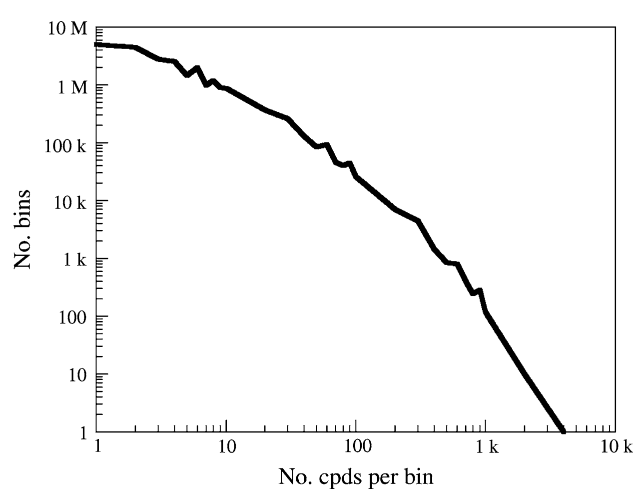


Fig. 1 Distribution of MQN-bins as a function of bin-occupancy for GDB-13

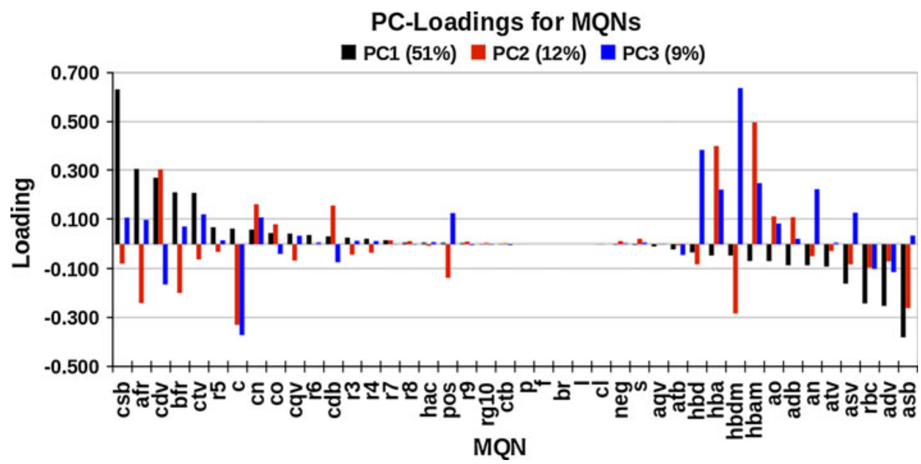


Fig. 2 Loadings of the first three principal components in the PCA of MQNs for GDB-13. MQNs are sorted by decreasing value of PC1 (*black bar*). See Table 1 for listing of MQNs

In the (PC1, PC2) plane, GDB-13 appears as a series of overlapping vertically elongated islands, each containing compounds with increasing numbers of rings and ring atoms (Fig. 3). Polar molecules with a high proportion of H-bond acceptor atoms occupy the northern portion of the map, while apolar molecules with mostly carbon atoms occupy the south. This layout corresponds to the (PC1, PC3) view obtained in the analysis of the PubChem fragments presented in the preceding paper. Indeed molecular size, which determined PC2 in the PubChem fragment analysis, does not significantly impact variance in GDB-13 because 87% of the database contains molecules of exactly 13 non-hydrogen atoms.

Subsets of GDB-13

GDB-13 is produced from an exhaustive enumeration starting from mathematical graphs [28] using filters removing many unstable and/or synthetically undesirable functional groups [11], which circumvents some of the limitations of the enumeration algorithms used in computer-aided structure elucidation [29, 30]. Despite of this careful selection of functional groups, the database still contains a large fraction of problematic molecules in the perspective of synthetic and medicinal chemistry [31]. For example, 35% of GDB-13 molecules contain one or more non-aromatic N–N or N–O bond (in an oxime or hydrazone), 29% contain at least one ester, aldehyde, carbonate, sulfate, epoxide or aziridine, 63% contain at least one non-aromatic carbon–carbon double or triple bond, and 54% contain at least one 3- or 4-membered ring.

To facilitate the identification of molecules with the least problematic structural features and the most relevance for drug discovery in GDB-13, subsets A–H were formed by removing non-aromatic cyclic and acyclic heteroatom-heteroatom bonds (subsets A and B), problematic functional groups (subset C), non-aromatic cyclic and acyclic CC-unsaturations (subsets D and E), small cycles (subset F). Further restrictions were taken by applying the “rule of three” for fragment-likeness (subset G) [32], and finally excluding acyclic carbon atoms (defined here as scaffold-likeness, subset H) (Table 2). The combinatorial application of these criteria defined 255 subsets of decreasing size, with the smallest subset ABCDEFGH applying all criteria cumulatively containing 1.47 million structures, which is still 2.4-fold larger than the 619,675 molecules of that size available in public databases (Fig. 4).

Virtual screening

GDB-13 is far too large for applying advanced virtual screening tools such as docking or shape-based analyses, which are too resource intensive to perform on more than a few million structures [33]. Therefore a virtual screening strategy for GDB-13 should start with a first rapid enrichment step. We showed recently that distances between molecules in MQN-space measured by the city-block distance (CBD_{MQN}) provide a useful similarity measure for virtual screening [24, 34, 35]. The CBD_{MQN} between two compounds is simply the sum of the 42 absolute values of the differences between MQN-values across the 42 pairs. Ranking PubChem by CBD_{MQN} relative to a reference bioactive compound was shown to strongly enrich other actives for the same target for most of the 40 classes listed in the DUD-dataset [24, 25]. To test if a similar MQN-based enrichment strategy would be applicable for GDB-13, we searched for analogs of 15 known reference bioactive compounds of 12 or 13 non-hydrogen atoms. An MQN-subset of 150,000 structures was assembled containing the 10,000 MQN-nearest neighbors of each of these 15 bioactive compounds in GDB-13. We then used scoring functions to estimate bioactivity probability in place of actual bioactivity measurements because synthesizing and testing any significant fraction of GDB-13 was not a practical option.

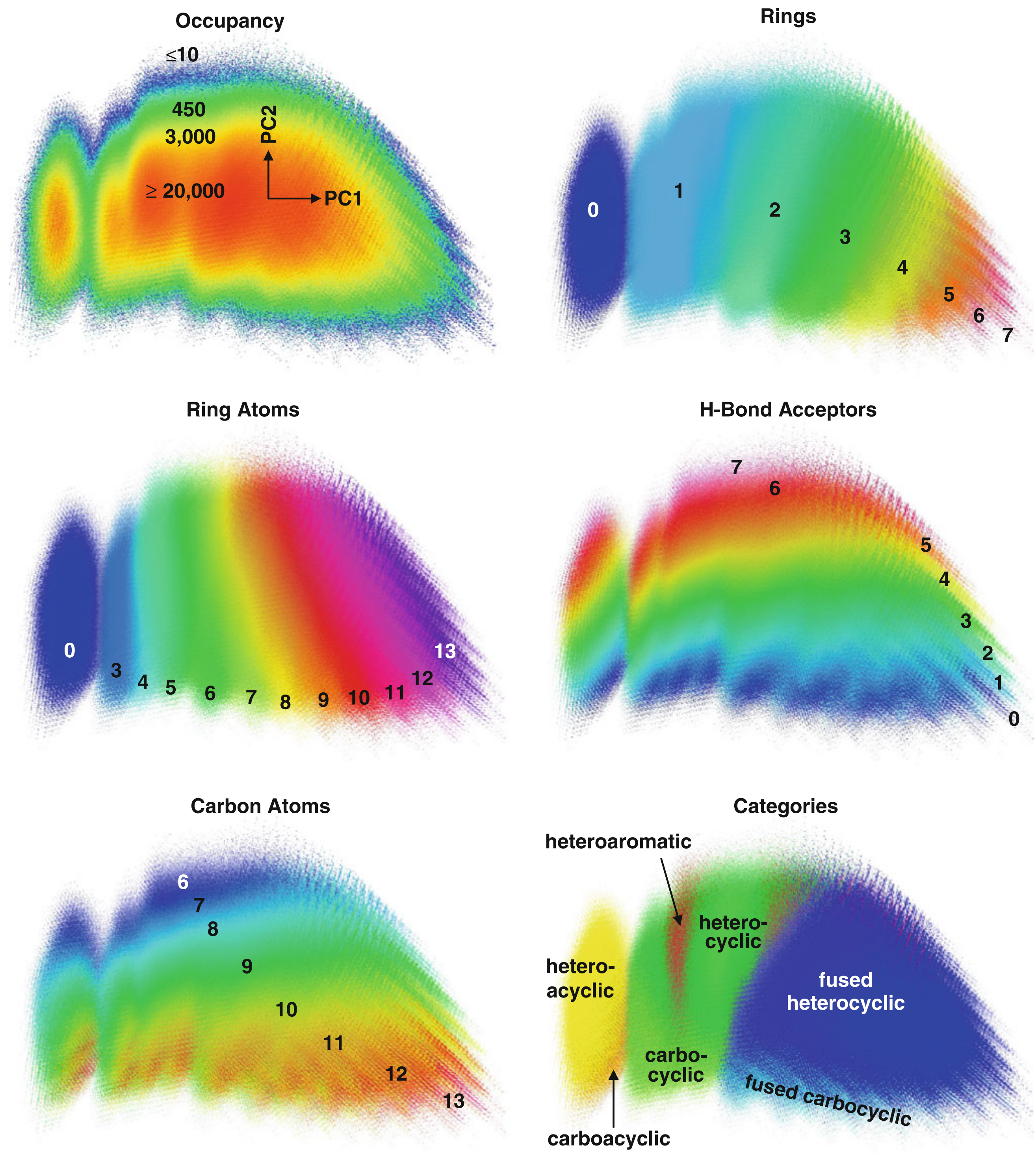


Fig. 3 MQN-maps of the (PC1, PC2) plane for GDB-13. PC1 codes for molecular rigidity and PC2 codes for polarity, see Fig. 2. The surface is hashed in 1,000 × 700 pixels. Each pixel is colored according to the occupancy or to the average value in that pixel, following the values indicated on the map on the corresponding color. Saturation to *grey* indicates the standard deviation for that value in the pixel up to ± 1 (rings), ± 2.1 (ring atoms and H-bond acceptors), and ± 2.8 (carbon atoms). The lightness scale (fading to *white*) encodes the occupancy in a logarithmic scale between 0 (*white*) and 200 (*full color*). For the category map molecules were assigned to categories in the priority order heteroaromatic (*red*) >aromatic (*purple*, not visible) >fused heterocycles (*blue*) >fused carbocycles (*cyan*) >heterocycles (*green*) >carbocycles (*green-yellow*) >heteroacyclic compounds (acyclic molecules with interrupted carbon chain, *yellow*) >carboacyclic compounds (acyclic molecules with continuous carbon chain, *orange*), and pixels were colored following the most frequent category in that pixel with fading to *grey* indicating category purity in the pixel

Table 2 Subsets of GDB-13

Criteria	Subset	Size	Cumulated	Size
GDB-13	–	975,821,779	–	–
No cyclic HetHet Bond[a]	A	801,013,244	–	–
No acyclic HetHet Bond[b]	B	779,957,069	AB	635,647,478
Stable FG[c]	C	693,944,404	ABC	441,084,370
No cyclic C=C and C≡C bonds[d]	D	662,075,045	ABCD	277,628,675
No acyclic C=C and C≡C bonds[e]	E	565,872,718	ABCDE	140,606,518
No small rings[f]	F	449,553,758	ABCDEF	43,729,989
Fragment-like[g]	G	353,200,314	ABCDEFG	12,899,741
Scaffold-like[h]	H	77,489,370	ABCDEFGH	1,470,284

[a] excludes non-aromatic cyclic NN and NO bonds

[b] excludes acyclic NN and NO bonds, mostly from oximes and hydrazones

[c] excludes aldehydes, esters, carbonates, sulfates, epoxides, aziridines

[d] excludes non-aromatic CC double and triple bonds inside cycles

[e] excludes acyclic CC double and triple bonds

[f] excludes three- and four-membered rings

[g] “rule of 3” according to Congreve [32]

[h] excludes acyclic carbon atoms

GDB molecules were first scored using the Tanimoto similarity coefficient of a 1,024-bit Daylight-type substructure fingerprint (T_{SF}) [26]. Substructure fingerprints are binary fingerprints in which bits are turned on whenever a particular substructure is present in a molecule, with substructures defined as groups of atoms connected by bonds up to a given maximum topological length, in our case up to 7 bonds. Therefore, the similarity coefficient T_{SF} reflects structural similarity but is strongly correlated with bioactivity because structural analogs often share similar bioactivities. The T_{SF} values of all 977 million GDB-13 molecules to each of the 15 reference bioactive compounds was computed. A threshold value of $T_{SF} > 0.7$ was used for hit identification by structural similarity, which gave a hit rate of 0.12% across the entire GDB-13. Similar values were observed in the different subsets A–H (hit rate = 0.14 ± 0.05% across the eight subsets), indicating that substructure limitations had relatively little impact on the hit rate. Interestingly, the MQN-nearest neighbor subset showed a hit rate of 4.5%, indicating that MQN-nearest neighbours were enriched 38-fold for high similarity analogs over the entire database (Table 3).

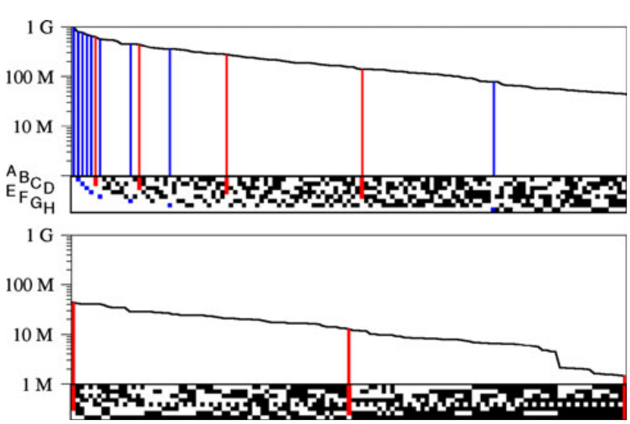


Fig. 4 Size of the 255 GDB-13 subsets obtained by the combinatorial use of criteria A–H (Table 2). Subsets are ordered by decreasing size. *Blue bars*: pure subsets (one criterion only). *Red bars*: cumulated subsets (one to eight criteria). Subset criteria are shown in the *bar code* below each point. Each *line* of the bar code corresponds to one of the criteria A through H as indicated at *left* in the *upper* plot

Table 3 Structural similarity scores of GDB-13 and its subsets relative to 15 reference bioactive compounds

Subset	Size[a]	$T_{SF(max)} > 0.7$[b]	%
GDB-13	975,821,779	1,171,324	0.12
A	801,013,244	1,171,324	0.15
B	779,957,069	1,171,258	0.15
C	693,944,404	1,120,167	0.16
D	662,075,045	1,171,272	0.18
E	565,872,718	1,169,816	0.21
F	449,553,758	188,036	0.04
G	353,200,314	592,208	0.17
H	77,489,370	61,306	0.079
Top MQNs[c]	150,000	6,748	4.5

[a] see also Table 2

[b] maximum T_{SF} value against the 15 bioactive compounds. T_{SF} is the Tanimoto similarity coefficient for a 1,024-bit Daylight-type substructure fingerprint

[c] containing the 10,000 MQN-nearest neighbors of each of the 15 reference bioactive compounds

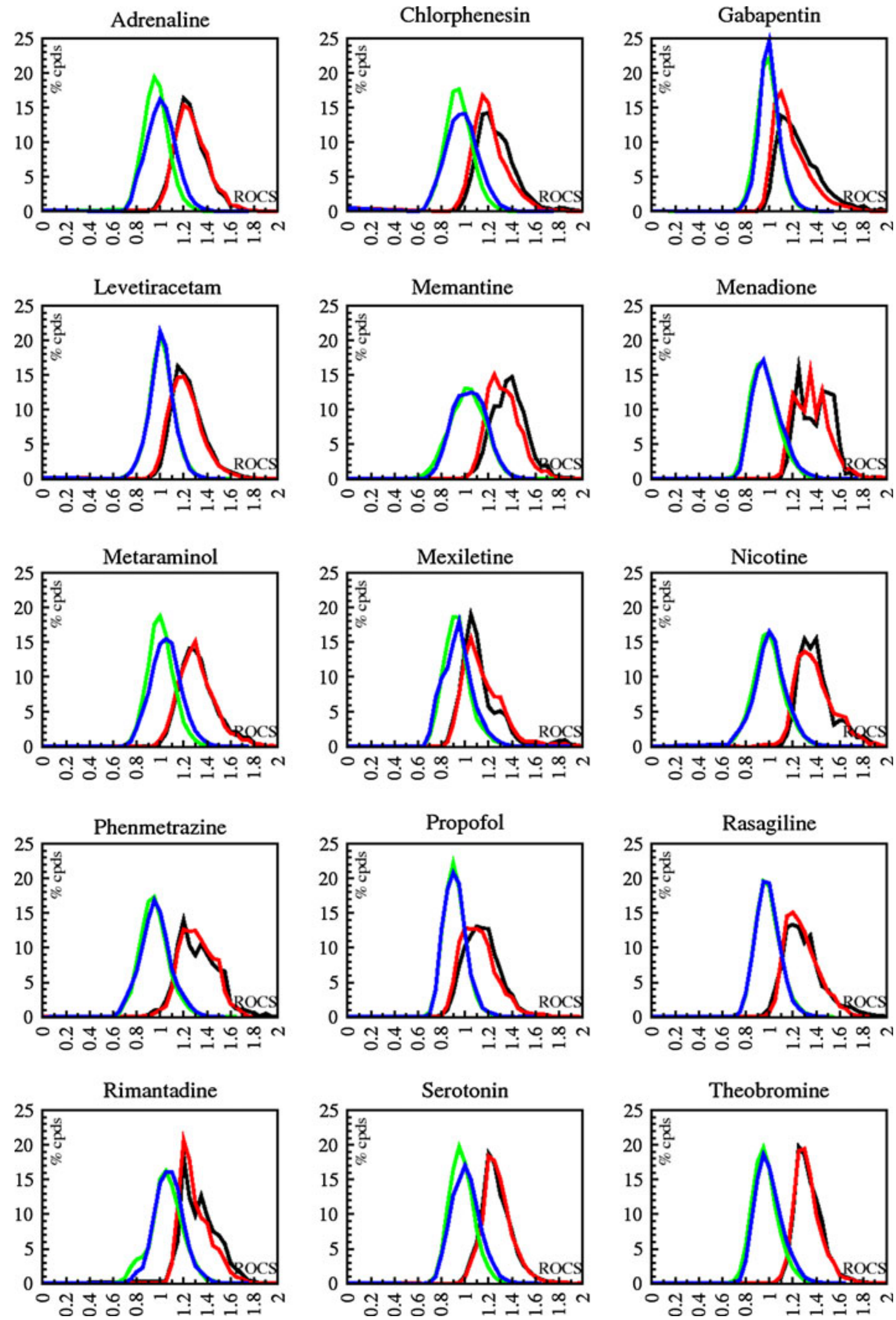


Fig. 5 Gaussian distribution of ROCS scores in ranking various subsets. *Green line*: 10,000 randomly selected compounds from GDB-13. *Blue line*: 10,000 randomly selected compounds from subset ABCDE. *Black line*: 10,000 MQN-nearest neighbors of the respective bioactive compound taken from GDB-13. *Red line*: 10,000 MQN-nearest neighbors of the respective bioactive compound taken from subset ABCDE

In a second approach, GDB-13 and its subsets were scored using the ROCS TanimotoCombo score. The ROCS (Rapid Overlay of Chemical Structures) TanimotoCombo score measures the similarity between 3D shapes of molecules by maximizing an overlap function between molecular shapes, considering these shapes as continuous functions constructed from atom-centered electrostatic and volume Gaussians [27]. The score is maximized by comparing various conformers of both query and reference molecule. This 3D shape-based approach is well-validated for ligand-based virtual screening [36]. ROCS was applied to search for shape-similar compounds of each of the 15 selected bioactive compounds among four different sets: (1) 10,000 randomly selected compounds from GDB-13, or (2) from subset ABCDE (Table 1), (3) 10,000 MQN-nearest neighbors of the respective bioactive compound taken from GDB-13, or (4) from subset ABCDE. For each structure, all possible diastereoisomers were generated and

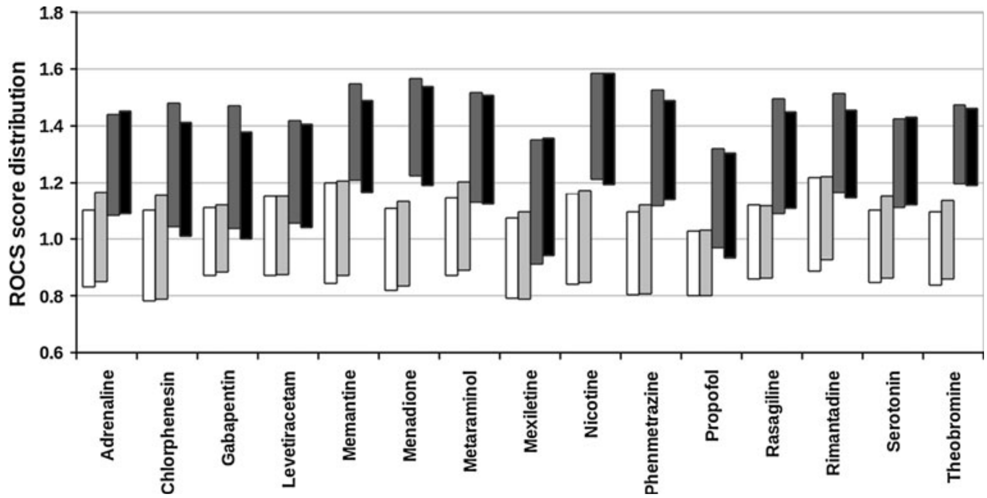


Fig. 6 Shape similarity scores of GDB-13 and its subsets against analogs of 15 bioactive compounds. "*Floating bar*" plot of max ROCS TanimotoCombo score showing in each case the the full width at half maximum range (FWHM, corresponds average $\pm$ 1.177 stddev for (1) 10,000 compounds randomly selected from GDB-13 (*white bars*); (2) 10,000 compounds randomly selected from subset ABCDE (*light grey bars*); (3) 10,000 MQN-nearest neighbors from GDB-13 (*dark grey bars*); (4) 10,000 MQN-nearest neighbors from subset ABCDE (*black bars*)

the score of the highest scoring stereoisomer was retained. In each case a Gaussian distribution of ROCS scores was observed (Fig. 5). Strikingly, the MQN-nearest neighbor series (3) and (4) ranked on average 0.29 ± 0.06 units higher than the non-MQN selected series (1) and (2) (Fig. 6). On average $22 \pm 11\%$ of each MQN-nearest neighbor subset (3) and (4) scored higher than 1.4, which can be considered as an indicator of similar bioactivity [37]. By comparison the random selections (1) and (2) contained only $0.29 \pm 0.24\%$ of compounds with ROCS > 1.4. The MQN-nearest neighbors thus contained 75-fold more high-ROCS compounds than the random selection.

A closer analysis of the T_{SF} and ROCS scores as a function of CBD_{MQN} from the reference drugs showed that MQN-neighbors ($CBD_{MQN} \leq 15$) consistently provided many high T_{SF} and high ROCS scoring compounds, while structures at larger MQN-distance ($CBD_{MQN} > 15$) had generally low scores (Fig. 7). The T_{SF} and ROCS scores were however only weakly correlated. The "high-ROCS, low T_{SF}" compounds are of particular interest since these represent scaffold-hopping analogs [38]. Examples of analogs with high scores with only one of the similarity measures are shown in Fig. 8. Note that these examples were taken mostly from subset ABCDE, which excludes in particular non-aromatic CC unsaturations (criteria DE) and thereby avoids compounds with cyclohexadiene and cyclopentadiene analogs of aromatic rings. Such cyclic dienes are indeed highly similar in shape to aromatic rings, but are not advisable as synthetic targets due to their reactivity, in particular towards oxidative aromatization (cyclohexadienes), electrocyclic isomerization or dimerization (cyclopentadienes).

Conclusion

The chemical universe database GDB-13 was analyzed by MQN and subdivided into 255 subsets by the combinatorial use of eight different restrictive criteria eliminating problematic functional groups and structural elements. Virtual screening for analogs of fifteen bioactive compounds of 12 or 13 non-hydrogen atoms showed that selection of MQN-nearest neighbors of any query molecule (using CBD_{MQN} as distance measure) provides subsets that are enriched in high-scoring compounds in terms of both structural similarity (T_{SF}) and shape similarity (ROCS TanimotoCombo score). The automatic retrieval of MQN-nearest neighbors from GDB-13 or its subsets is facilitated by a search tool available at www.gdb.unibe.ch. The method should greatly facilitate the exploitation of GDB-13 for the identification of new medicinally relevant small molecules for synthesis and testing.

Methods

MQNs

MQNs were calculated using the previously reported calculator source code (Supporting Information in Ref. [19])

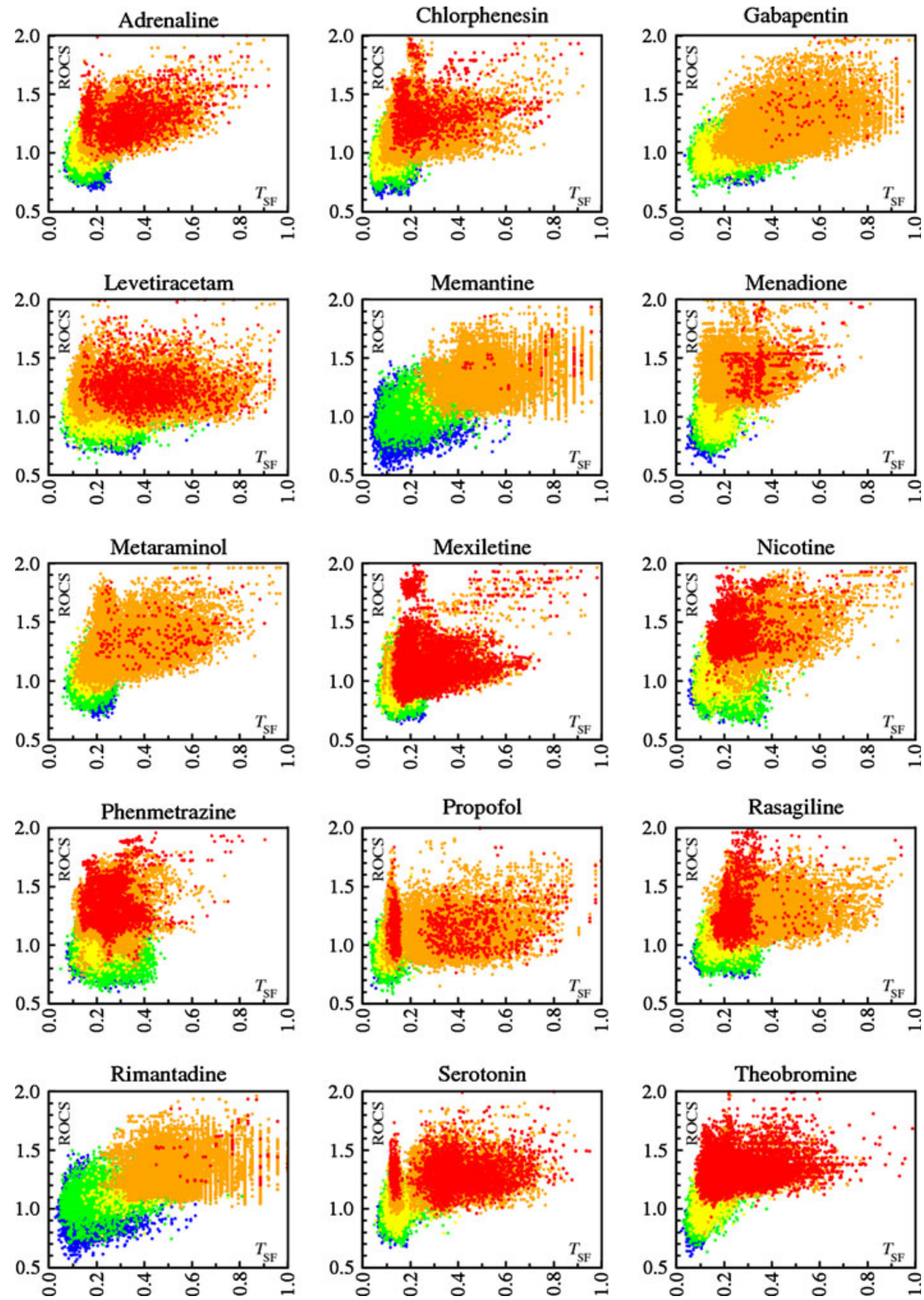


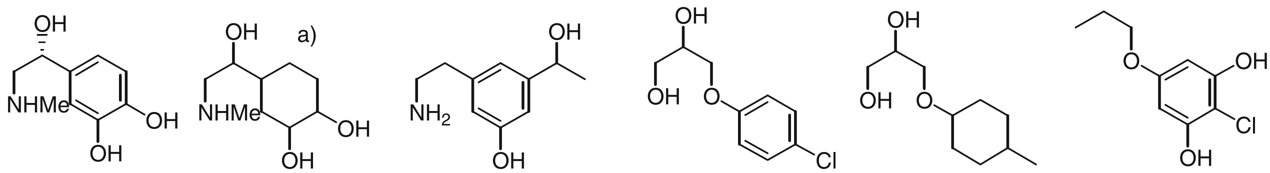
Fig. 7 For all 15 bioactive compounds: Scatter plot of ROCS TanimotoCombo score vs. substructure fingerprint Tanimoto (T_{SF}) of the 10'000 CBD_{MQN}-nearest neighbors and 10'000 randomly selected compounds from GDB-13 (= 20'000 data points per plot). Color of the points is according to CBD_{MQN} to reference: *red* = 0–5, *orange* = 6–15, *yellow* = 16–30, *green* = 31–50, *blue* ≥ 50. Levels in the plot are overlayed with priority *red* > *orange* > *yellow* > *green* > *blue*

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 [39] was done by using an in-house developed Java application using Jsci (http://jsci.sourceforge.net). The source code is based on the tutorial of Lindsay I. Smith (http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf) and made parallalelizable to reduce calculation time. 1,646,535 SMILES (0.17% of GDB-13) dropped out during the MQN-determination process. CBD_{MQN} calculations were computed at approximately 175,000 comparisons per minute per CPU.

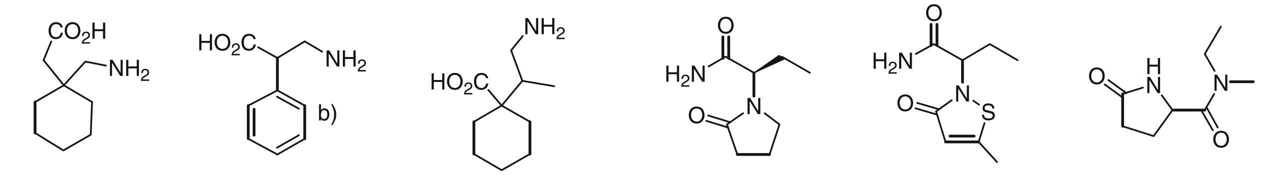
Fig. 8 Structural formula of the 15 reference bioactive compounds and ▶ their analogs with their CBD_{MQN}, ROCS-score and T_{SF}-score relative to the reference drug. The analogs are not yet known with the following exceptions: *a* known; *b* purchasable; *c* known as substructure

Substructure fingerprints

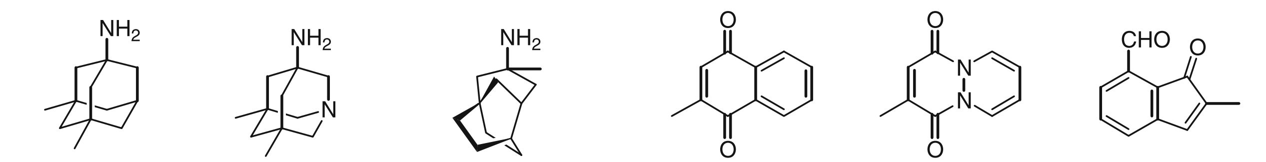
For substructure similarity calculation a Daylight-type 1,024-bit hashed fingerprint from ChemAxon was used. T_{SF}-similarity calculations were computed at approximately 86,000 comparisons per minute per CPU.



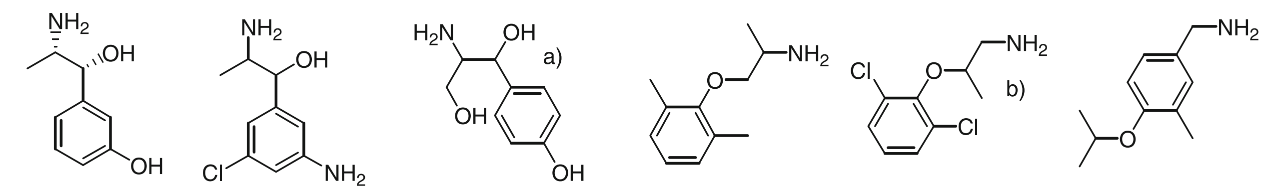
Adrenaline	6 / **1.83** / 0.25	5 / 1.22 / **0.81**	Chlorphenesin	8 / **1.94 /** 0.29	7 / 1.07 / **0.76**



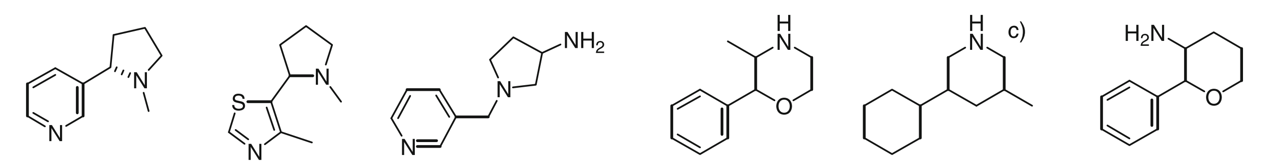
Gabapentin	10 / **1.74** / 0.28	6 / 1.17 / **0.93**	Levetiracteam	8 / **1.95** / 0.19	2 / 1.04 / **0.9**



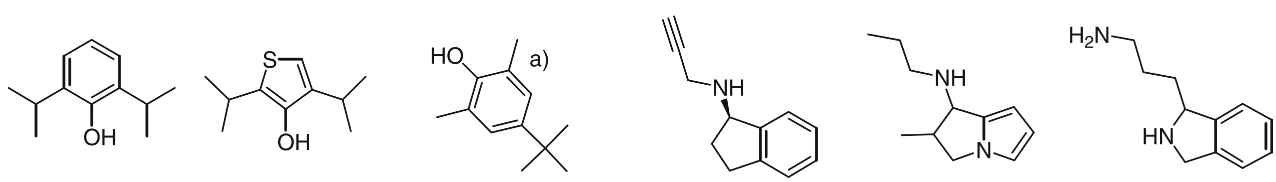
Memantine	4 / **1.85** / 0.54	12 / 1.02 / **1.0**	Menadione	6 / **1.99 /** 0.15	7 / 1.39 / **0.7**



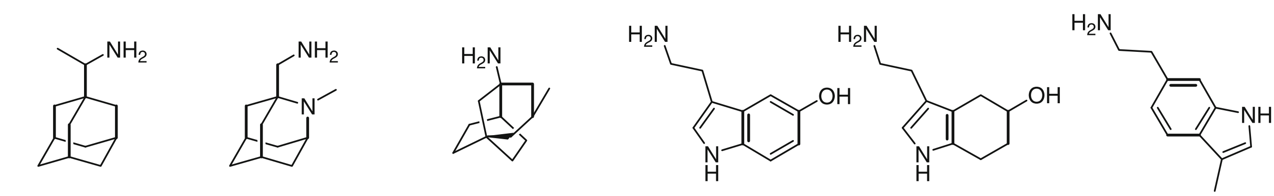
Metaraminol	9 / **1.96** / 0.56	9 / 1.21 / **0.86**	Mexiletine	4 / **1.93** / 0.47	0 / 1.10 / **0.72**



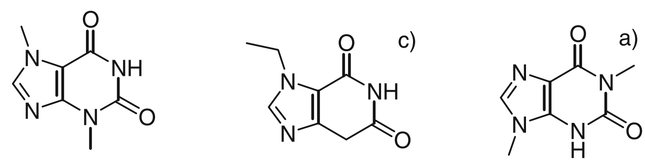
Nicotine	10 / **1. 9/** 0.39	7 / 1.3 / **0.72**	Phenmetrazine	4 / **1.72** / 0.18	3 / 1.37 / **0.75**



Propofol	8 / **1.91** / 0.38	7 / 1.09 / **0.91**	Rasagiline	13 / **1.82** / 0.31	10 / 1.17 / **0.84**



Rimantadine	8 / **1.80** / 0.58	9 / 1.16 / **0.96**	Serotonin	4 / **1.99** / 0.42	6 / 1.29 / **0.83**



Theobromine	7 / **1.94** / 0.22	0 / 1.28 / **0.94**

ROCS

For the ROCS calculations, the stereo information of the 15 reference bioactive compounds was added as found in DrugBank or Pubchem (see Fig. 7, no information found for Chlorphenesin, Mexiletine and Phenmetrazine). All queries and target molecules were sent to Omega to create a maximum of 200 lowest energy 3D structures including various stereoisomers and their conformers. For all ROCS runs the "TanimotoCombo" overlap score was used. ROCS TanimotoCombo score calculations were computed at approximately 15 comparisons per minute per CPU.

Acknowledgments This work was supported financially by the University of Berne, the Swiss National Science Foundation and the Office Fédéral Suisse de l'Education et de la Science.

References

1. Coyne AG, Scott DE, Abell C (2010) Drugging challenging targets using fragment-based approaches. Curr Opin Chem Biol 14:299–307
2. Schulz MN, Hubbard RE (2009) Recent progress in fragment-based lead discovery. Curr Opin Pharmacol 9:615–621
3. Hartenfeller M, Schneider G (2011) De novo drug design. Methods Mol Biol 672:299–323
4. Venhorst J, Nunez S, Kruse CG (2010) Design of a high fragment efficiency library by molecular graph theory. ACS Med Chem Lett 1:499–503
5. Carr RA, Congreve M, Murray CW, Rees DC (2005) Fragment-based lead discovery: leads by design. Drug Discov Today 10:987–992
6. Rees DC, Congreve M, Murray CW, Carr R (2004) Fragment-based lead discovery. Nat Rev Drug Discov 3:660–672
7. Hajduk PJ, Greer J (2007) A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov 6:211–219
8. Boyd SM, de Kloe GE (2010) Fragment library design: efficiently hunting drugs in chemical space. Drug Discov Today 7:e173–e180
9. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623–W633
10. van Deursen R, Blum LC, Reymond JL (2011) Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem. J Comput Aided Mol Des. doi:10.1007/s10822-011-9437-x
11. Blum LC, Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131:8732–8733
12. Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 daltons. Angew Chem Int Ed Engl 44:1504–1508
13. Fink T, Reymond JL (2007) 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 47:342–353
14. Nguyen KT, Syed S, Urwyler S, Bertrand S, Bertrand D, Reymond JL (2008) Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. ChemMedChem 3:1520–1524
15. Nguyen KT, Luethi E, Syed S, Urwyler S, Bertrand S, Bertrand D, Reymond JL (2009) 3-(aminomethyl)piperazine-2, 5-dione as a novel NMDA glycine site inhibitor from the chemical universe database GDB. Bioorg Med Chem Lett 19:3832–3835
16. Luethi E, Nguyen KT, Burzle M, Blum LC, Suzuki Y, Hediger M, Reymond JL (2010) 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 53:7236–7250
17. Garcia-Delgado N, Bertrand S, Nguyen KT, van Deursen R, Bertrand D, Reymond J-L (2010) Exploring a7-nicotinic receptor ligand diversity by scaffold enumeration from the chemical universe database GDB. ACS Med Chem Lett 1:422–426
18. Reymond JL, Van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Med Chem Commun 1:30–38
19. Nguyen KT, Blum LC, van Deursen R, Reymond JL (2009) Classification of organic molecules by molecular quantum numbers. ChemMedChem 4:1803–1805
20. Pearlman RS, Smith KM (1998) Novel software tools for chemical diversity. Perspect Drug Discov Des 9–11:339–353
21. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216
22. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330
23. Irwin JJ, Shoichet BK (2005) ZINC—A free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
24. van Deursen R, Blum LC, Reymond JL (2010) A searchable map of PubChem. J Chem Inf Model 50:1924–1934
25. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
26. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996
27. Rush TS III, Grant JA, Mosyak L, Nicholls A (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. J Med Chem 48:1489–1495
28. McKay BD (1981) Practical graph isomorphism. Congr Numerantium 30:45–87
29. Warr WA (1993) Computer-assisted structure elucidation. Part II: indirect database approaches and established systems. Anal Chem 65:1087A–1095A
30. Steinbeck C (2004) Recent developments in automated structure elucidation of natural products. Nat Prod Rep 21:512–518
31. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1:8
32. Congreve M, Carr R, Murray C, Jhoti H (2003) A rule of three for fragment-based lead discovery? Drug Discov Today 8:876–877
33. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK (2009) Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol 20:429–436
34. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11:1046–1053
35. Khalifa AA, Haranczyk M, Holliday J (2009) Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. J Chem Inf Model 49:1193–1201
36. Hawkins PC, Skillman AG, Nicholls A (2007) Comparison of shape-matching and docking as virtual screening tools. J Med Chem 50:74–82

37. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B (2010) Molecular shape and medicinal chemistry: a perspective. J Med Chem 53: 3862–3886
38. Schneider G, Neidhart W, Giller T, Schmid G (1999) "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed Engl 38:2894–2896
39. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York