This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACCOUNTS of chemical research

Article

pubs.acs.org/accounts

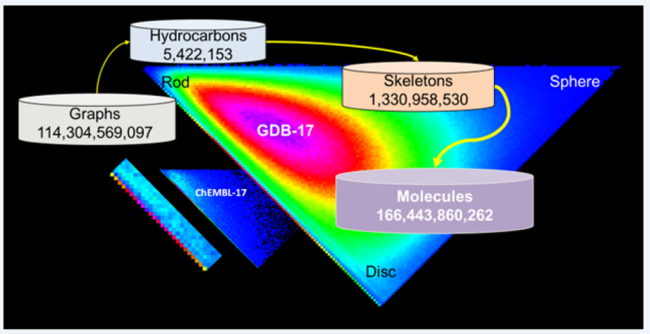
The Chemical Space Project

Published as part of the Accounts of Chemical Research special issue "Synthesis, Design, and Molecular Function".

Jean-Louis Reymond*

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

CONSPECTUS: One of the simplest questions that can be asked about molecular diversity is how many organic molecules are possible in total? To answer this question, my research group has computationally enumerated all possible organic molecules up to a certain size to gain an unbiased insight into the entire chemical space. Our latest database, GDB-17, contains 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens, by far the largest small molecule database reported to date. Molecules allowed by valency rules but unstable or nonsynthesizable due to strained topologies or reactive functional groups were not considered, which reduced the enumeration by at least 10 orders of magnitude and was essential to arrive at a manageable database size. Despite these restrictions, GDB-17 is highly relevant with respect to known molecules.



Beyond enumeration, understanding and exploiting GDBs (generated databases) led us to develop methods for virtual screening and visualization of very large databases in the form of a "periodic system of molecules" comprising six different fingerprint spaces, with web-browsers for nearest neighbor searches, and the MQN- and SMIfp-Mapplet application for exploring color-coded principal component maps of GDB and other large databases. Proof-of-concept applications of GDB for drug discovery were realized by combining virtual screening with chemical synthesis and activity testing for neurotransmitter receptor and transporter ligands. One surprising lesson from using GDB for drug analog searches is the incredible depth of chemical space, that is, the fact that millions of very close analogs of any molecule can be readily identified by nearest-neighbor searches in the MQN-space of the various GDBs. The chemical space project has opened an unprecedented door on chemical diversity. Ongoing and yet unmet challenges concern enumerating molecules beyond 17 atoms and synthesizing GDB molecules with innovative scaffolds and pharmacophores.

INTRODUCTION

Organic molecules are defined by the number, type, topological connectivity, and stereochemistry of atoms described by their structural formula. As of today over one hundred million such organic molecules have been prepared, mostly in the context of medicinal chemistry.[1,2] Cheminformatics provides various computational tools to handle the massive amount of information created by these millions of molecules, in particular to enable database classification and bioactivity prediction.[3−7]

One of the simplest questions that can be asked about molecular diversity is how many organic molecules are possible in total? The so-called "drug-like" chemical space has been estimated at 10^{60} for all molecules obeying Lipinski's rule-of-five for oral bioavailability,[8,9] and at 10^{20}−10^{24} for all molecules up to 30 atoms,[10] in any case a number far too large for practical application.[11] Nevertheless my research group has undertaken the computational enumeration of all possible organic molecules up to a certain size. Our goal was not only to count molecules but also to write them down as SMILES,[4] to understand their diversity, and to test their possible relevance for drug discovery. This "chemical space project" led to the chemical universe databases (GDBs, generated databases) enumerating molecules following criteria for size, chemical stability, and synthetic feasibility. This Account follows previous reviews[12−15] and provides a perspective on our work on database assembly, visualization, and drug discovery.

ENUMERATION

Organic molecules can be derived from mathematical graphs by substituting atoms for graph nodes and chemical bonds for graph edges. In 1875 Cayley, the inventor of graph theory, reported the first application of this principle by estimating the number of possible acyclic branched hydrocarbons as a function of size,[16−18] an approach later followed for other topologies.[19,20] Beyond counting, structure enumeration algorithms such as MOLGEN[21,22] have been produced to enable computer-assisted structure elucidation (CASE)[23−26] by enumerating molecules fitting predefined criteria of elemental composition, mass, and the presence or absence of functional groups. Other types of structure generators such as SPROUT[27] are genetic algorithms that evolve organic molecules for

Received: November 29, 2014
Published: February 17, 2015

 © 2015 American Chemical Society

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

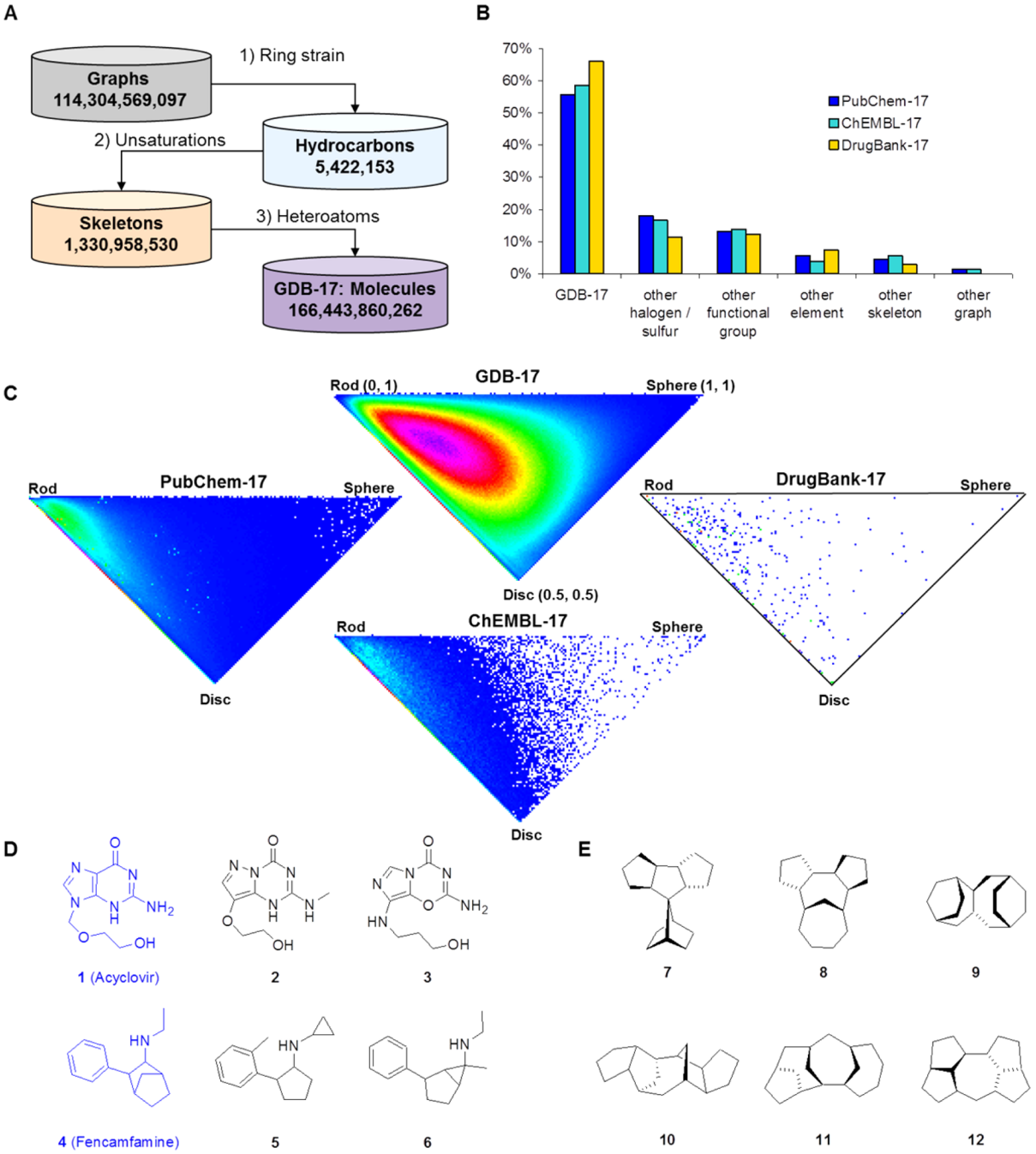


Figure 1. Chemical universe database, GDB-17. (A) Enumeration principle. (B) Percentage of known molecules up to 17 atoms in the public databases PubChem, ChEMBL, and DrugBank present or absent from GDB-17. (C) Occupancy of the shape triangle by molecules up to 17 atoms in GDB, PubChem, ChEMBL, and DrugBank. (D) Drugs (blue) and yet unknown isomers with similar pharmacophores. (E) Yet unknown polycyclic hydrocarbons from GDB-17.

maximum docking to a target protein or highest similarity to a reference molecule.[28] In these algorithms, molecules are often assembled from known building blocks using known coupling reactions, an approach also used to enumerate virtual combinatorial libraries,[29−32] in particular the Pfizer Global Virtual Library, which corresponds to 10 trillion molecules, although the compounds are only enumerated in response to specific searches.[33]

In contrast to the above tailored approaches, we set out to enumerate all possible molecules to gain an unbiased insight into the entire chemical space, taking only simple chemical stability and synthetic feasibility criteria into account. Starting with mathematical graphs produced by the program GENG,[34] graphs suitable to build saturated hydrocarbons were selected taking ring strain and topology into account, for example, excluding hydrocarbons with planar or pyramidal quaternary centers. These graphs were then converted to skeletons by introducing unsaturations following rules for valency, aromaticity, and ring strain, for example, excluding bridgehead double bonds and allenes. Finally molecules were obtained by mutating carbon atoms to N, O, S, and halogens taking functional group stability into account, for example, excluding water-reactive groups such as acyl chlorides, anhydrides, hemiacetals, enols, and enamines, and all heteroatom−heteroatom bonds except

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

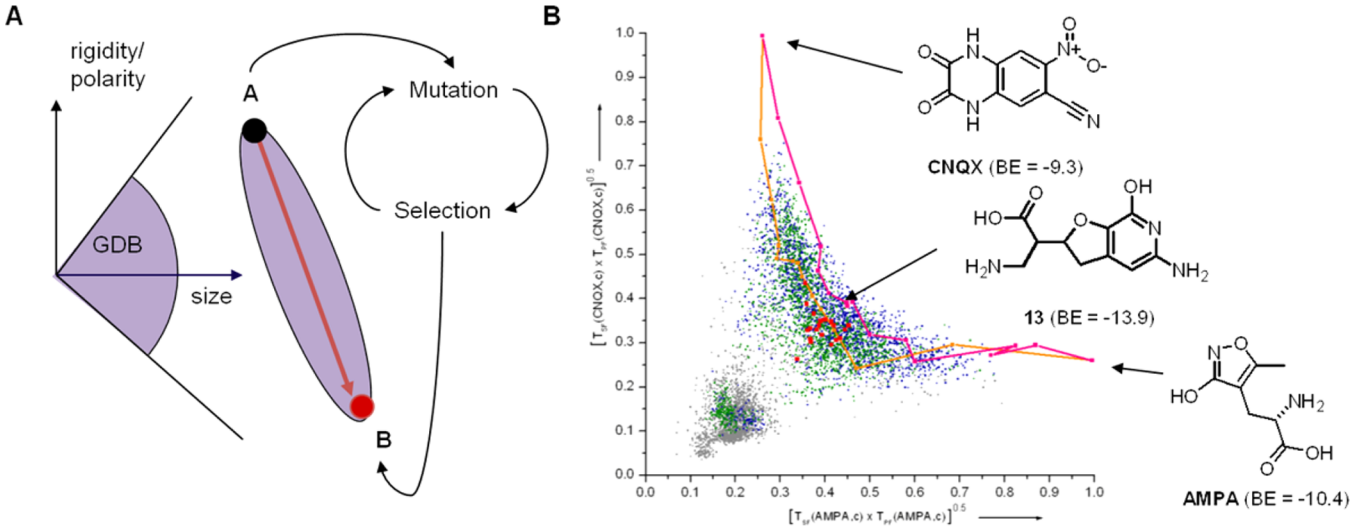


Figure 2. Chemical space travel. (A) Concept of chemical space travel. (B) Application to AMPA receptor ligands.

within aromatic rings, oximes, and hydrazones. Following our initial databases GDB-11[35,36] and GDB-13,[37] we recently enumerated 166.4 billion molecules of up to 17 atoms of C, N, O, S, and halogens collected in GDB-17, by far the largest database of explicitly enumerated small molecules reported to date (Figure 1A).[38,39] Related databases of enumerated aromatic heterocycles have been reported by others.[40,41]

Removing molecules allowed by valency rules but unstable or nonsynthesizable due to strained topologies or reactive functional groups as discussed above reduced the enumeration by at least 10 orders of magnitude and was therefore essential to arrive at a manageable database size. Despite these restrictions, GDB-17 is highly relevant with respect to known molecules. It contains approximately 60% of PubChem, ChEMBL, and Drugbank molecules up to 17 atoms, the remaining 40% featuring mostly molecules with nonenumerated elements (P, B, Si) and functional groups (Figure 1B). The unfiltered database of all valency-allowed C, N, and O compounds up to nine atoms is available for download at www.gdb.unibe.ch under the name DMU9 ("dark-matter universe").

GDB molecules are stored as SMILES representing 2D structures, which can be expanded to stereoisomers using the 3D-generator CORINA.[42] They are mostly stereochemically rich molecules of intermediate polarity with a three-dimensional molecular shape,[43] a property that is rather rare in historical drugs but correlates with clinical success of drug candidates (Figure 1C).[44] Most strikingly, enumeration results in a large diversity of structural types that are otherwise rather difficult to access, as exemplified by millions of isomers of marketed drugs, many of which have a very close pharmacophore and shape yet have never been reported, and by many yet unknown ring systems (compounds **1**−**12**, Figure 1D,E).

Enumerating organic molecules beyond GDB-17 represents an ongoing challenge and probably cannot be done exhaustively. One approach for sampling this larger chemical space consists in mutating known molecules from first principles to generate new structures.[45−47] In our "chemical space travel" (CST) algorithm,[48] iterative cycles of structural mutations on a starting molecule A coupled to selection by similarity to a target molecule B produces trajectory libraries of hundreds of thousands of intermediates representing the chemical space between start and target, an approach suitable for connecting between various molecules up to 50 atoms in a few tens of successive mutations (Figure 2A). CST was used to identify a strongly docking hybrid molecule **13** between AMPA and CNQX as a possible partial agonist of the AMPA receptor (Figure 2B). This approach was recently exploited by others to generate random molecules as a way to explore chemical space.[49]

A PERIODIC SYSTEM OF MOLECULES

The GDB databases complement other large databases of organic molecules of interest in drug discovery, flavors, and fragrances chemistry, which together populate the known and unknown chemical space of organic molecules (Table 1). A classification system would be desirable to efficiently search for analogs and to visualize the diversity of such large databases. To address this challenge, we followed the concept of property spaces by which molecules are assigned numerical descriptor values, collected in a so-called fingerprint,[5] and placed at the corresponding coordinates of a multidimensional space where each dimension represents one of the descriptors.[50] In such property spaces, spatial proximity between molecules measures similarity.[51] Furthermore, principal component analysis (PCA) allows one to represent the property space by projection into the principal component plane (PC1, PC2), the mathematical equivalent of taking a picture from the most favorable angle (Figure 3A).[52−58]

Inspired by the periodic system of the elements, organized according to the atomic and main quantum numbers to form a table such that nearby elements have related properties, we devised the MQN system in the form of a multidimensional grid. Molecules were assigned to grid positions following the values of 42 molecular quantum numbers (MQN) counting features important for chemical structure and biological activity.[73,74] This MQN-system was enabled by a web-browser capable of identifying nearest neighbors of any molecule within seconds.[15,72] The same principle was applied with another five fingerprints providing different insights into molecular structure, namely, the SMILES fingerprint (SMIfp) counting 34 characters appearing in the SMILES of a molecule,[75] the 1024-bit binary Daylight type substructure fingerprint (Sfp)[76] and extended connectivity fingerprint (ECfp4)[77] counting the

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

Table 1. Databases of the Known and Unknown Chemical Space

database	description	size[a]	ref
DrugBank	approved and investigational drugs	7 584	59
SuperScent	scents from literature	2 300	60
Flavornet	volatile compounds from literature	738	61
SuperSweet	carbohydrates and artificial sweeteners	642	62
BitterDB	bitter cpds from literature and Merck index	606	63
PubChem	NIH repository of molecules	63 095 535	64,65
ZINC	commercial small molecules	22 724 825	66,67
ZINC.FL	fragrance-like subset of ZINC	69 724	68
BindingDB	small molecules annotated with bioactivity data	453 657	69,70
ChEMBL	small molecules annotated with bioactivity data	1 411 786	71
GDB-11	molecules of up to 11 atoms of C, N, O, and F	26 434 571	36
GDB-13	molecules of up to 13 atoms of C, N, O, S, and Cl	977 468 314	37
GDB-13.subset	simplicity-selected GDB-13 molecules	43 729 989	72
GDB-13.FL	fragrance-like subset of GDB-13	59 482 898	68
GDB-17	molecules of up to 17 atoms of C, N, O, S, and halogens	166 443 860 262	38

[a]For the latest version of each database as available in November 2014.

presence of specific substructures, the atom pair fingerprint (APfp) encoding molecular shape, and its related category extended atom pair fingerprint (Xfp) encoding pharmacophores.[78] Validation with sets of known bioactive molecules such as the directory of useful decoys (DUD)[79] and shape analogs showed that nearest neighbor searches by city-block distance in each of these six fingerprint spaces efficiently retrieved bioactive analogs. For the case of MQN, SMIfp, APfp, and Xfp fingerprints similarity retrieved "scaffold-hopping" analogs, which are molecules with similar shape, pharmacophores, and bioactivity but very different substructures as measured by Sfp indicating nonobvious and valuable structure–activity relationships.[80]

For MQN and SMIfp chemical spaces, PCA projects over 70% of data variability into the (PC1, PC2)-plane, making this plane suitable as map of chemical diversity. Color-coding according to molecular properties, such as molecule size, rigidity, and polarity, allows visualizing chemical diversity in various databases such as PubChem,[81] DrugBank,[82] and GDB-13.[72] Each pixel in these maps can be inspected by zooming and visualizing the molecules at that position with help of the MQN- and SMIfp-mapplets (Figure 3B). These java applications are freely available at www.gdb.unibe.ch to visualize DrugBank, ChEMBL, ZINC, PubChem, GDB-11, GDB-13, and GDB-17.[75,83] The mapplets also contain molecule localization functions and a link to the web-browser for proximity searching in the parent MQN- and SMIfp-space. A related Fragrance-mapplet application allows inspecting the Flavornet and Superscent databases and the related fragrance-like subsets of ZINC and GDB-13.[68] Taken together, the proximity search browsers and mapplets constitute a "periodic system of molecules" for exploring chemical space that to the best of our knowledge is unprecedented.

DRUG DISCOVERY

GDB contains almost exclusively (>99.9%) new molecules and therefore represents a vast reservoir of opportunities for drug discovery. Remarkably, the majority of GDB molecules fulfill drug-likeness,[8] lead-likeness,[84] and fragment likeness[85] criteria, mostly because graph diversity is highest with polycyclic rigid structures and because the introduction of heteroatoms gives the largest number of possibilities for intermediate ratios of heteroatom to carbon, resulting in many relatively rigid and polar molecules.

Our first projects exploited GDB-11 using a combination of substructure-guided compound selection and high-throughput docking, followed by synthesis and testing. We focused on neurotransmitter receptors and transporters because these targets can be modulated by very small molecules such as those found in GDB. Dipeptide **14** and diketopiperazine **15** were identified from a glycine analog search in GDB-11 as inhibitors of the glycine site of the NMDA receptor,[86,87] and norbornane aspartic acid *rac*-**16** was identified from an aspartate analog search as selective inhibitor of the glutamate transporter GLT-1,[88] in both cases via synthesis and evaluation of 20–30 test compounds (Figure 4). A similar approach was used to select possible modulators of the nicotinic acetylcholine receptor (nAChR) combining the selective enumeration of analogs of PNU-282,987[89] from quinuclidine-like diamines in GDB-11 with docking to the acetylcholine binding protein. Synthesis of over 80 computationally selected analogs and testing led to the discovery of the competitive α7 nAChR inhibitor **17**.[90,91]

Due to the limited throughput of docking, which only allowed evaluation of 5% of the compounds selected from GDB-11 in the case of the nAChR project, a more direct virtual screening approach was envisioned exploiting the concept of chemical space classification discussed above. Analogs of the 3-aminoquinuclidine nucleus of PNU-282,987 were extracted from GDB-13 by constraining MQN values. Remarkably, only 344 quinuclidine-like diamines remained when imposing up to nine carbon atoms, exactly two nitrogen atoms, two cycles size 5–7, no unsaturations, a maximum of two acyclic carbon atoms only as amine substituents, and at least two bonds shared by two rings to enforce a globular shaped bicyclic diamine. Three of these 344 diamines were selected by shape similarity[82,96] to PNU-282,987, novelty, and synthetic feasibility. The synthesis was demanding, but the approach was quite successful. Two of the selected compounds *rac*-**18** and *rac*-**19** together with **20** from the previous approach turned out to be positive allosteric modulators (PAM) of the $\alpha3\beta2$ nAChR, an unprecedented activity type absent from PNU-282,987 but desirable to strengthen muscular contraction in elderly people suffering from sacropenia by reinforcing neurotransmission at the neuromuscular junction.

The suitability of the MQN selection for identifying bioactive analogs of PNU-282,987 from GDB was confirmed in a parallel study with nicotine.[92] Starting from the fact that 322 compounds from GDB-13 also annotated as nicotinic acetylcholine receptor activity in ChEMBL were much closer to nicotine in MQN-space (CBD_{MQN} = 22.8 ± 12.5) than average GDB-13 molecules (CBD_{MQN} = 38.8 ± 11.1), a nearest neighbor selection was performed in the simplicity-selected GDB-13 subset (Table 1), leading to 31 504 nicotine analogs. Sixty of these were purchased from a commercial source and tested against the α7 nAChR, revealing a single agonist as the

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722–730

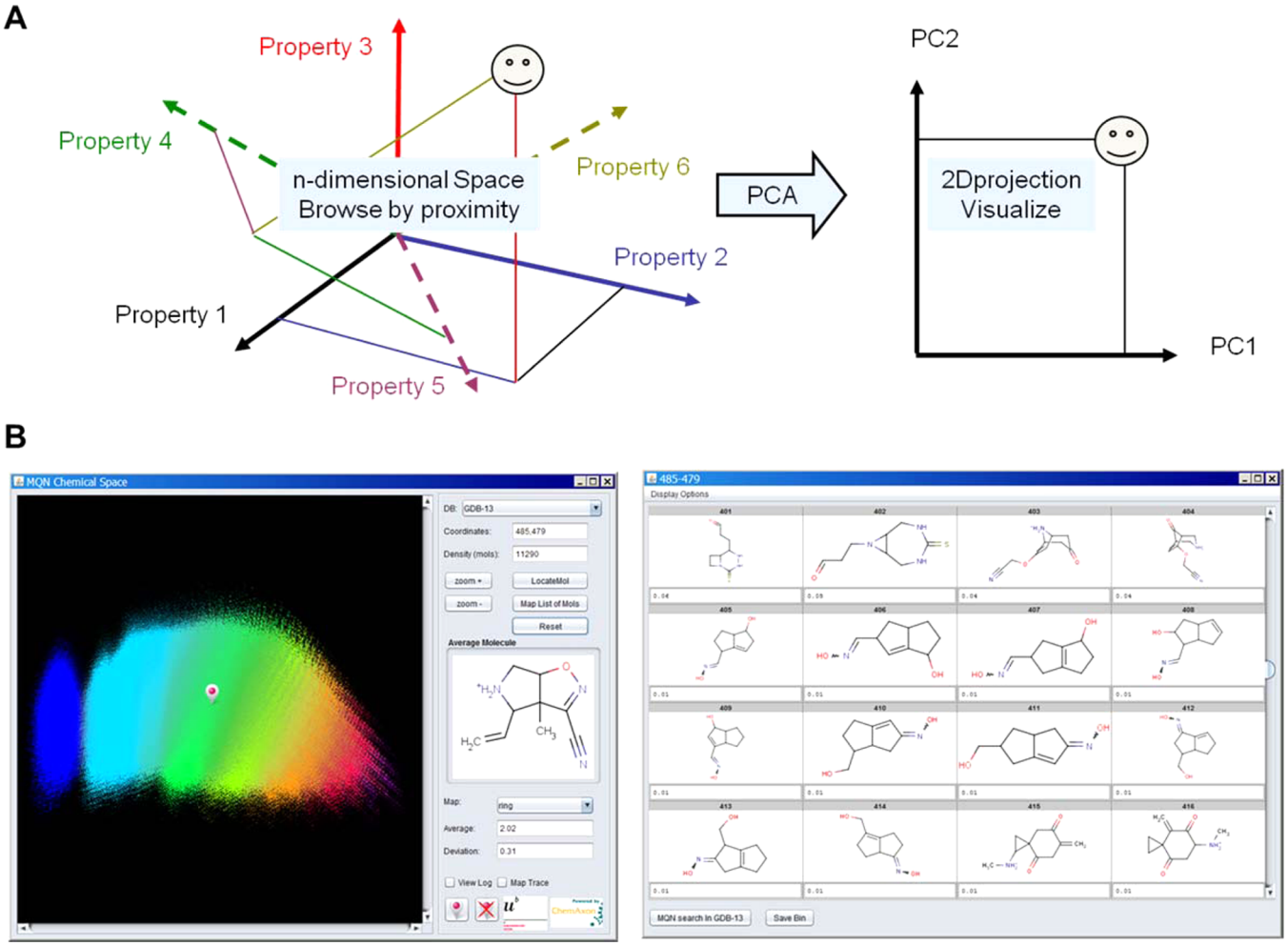
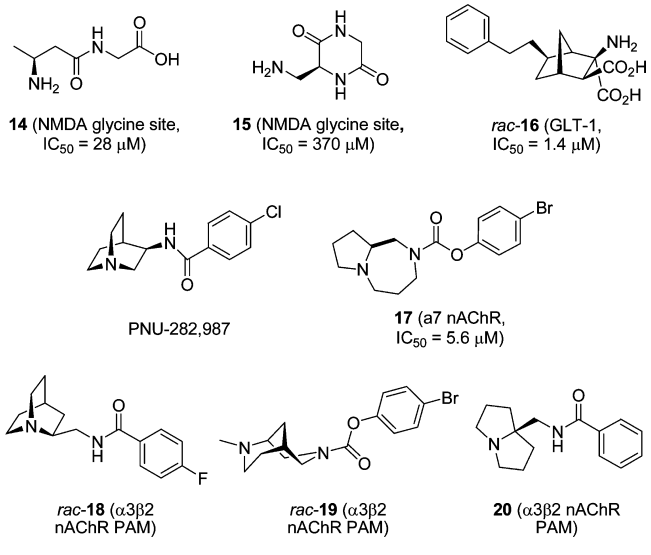


Figure 3. A periodic system of molecules. (A) Multidimensional chemical space and principal component analysis for visualization. (B) Image of the MQN-mapplet showing GDB-13 color-coded by ring count. Part of the content of the flagged pixel is shown at right.



14 (NMDA glycine site, IC_{50} = 28 μM)

15 (NMDA glycine site, IC_{50} = 370 μM)

rac-**16** (GLT-1, IC_{50} = 1.4 μM)

PNU-282,987

17 (α7 nAChR, IC_{50} = 5.6 μM)

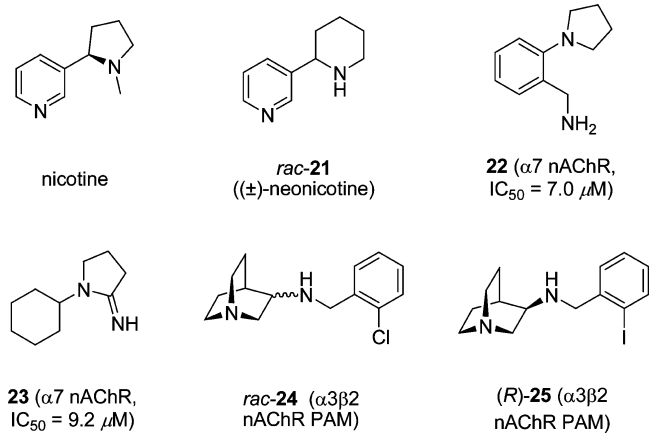
rac-**18** (α3β2 nAChR PAM)

rac-**19** (α3β2 nAChR PAM)

20 (α3β2 nAChR PAM)

Figure 4. Bioactive compounds selected from GDB. The $\alpha3\beta2$ nAChR PAMs enhance the signal induced by 50 μM acetylcholine by 1.5−3-fold at 10 μM.

known neonicotine *rac*-**21**, and several previously unknown inhibitors such as the benzylic amine **22** and the aliphatic amidine **23**, both of which are structurally quite distinct from nicotine (Figure 5).



nicotine

rac-**21** ((±)-neonicotine)

22 (α7 nAChR, IC_{50} = 7.0 μM)

23 (α7 nAChR, IC_{50} = 9.2 μM)

rac-**24** (α3β2 nAChR PAM)

(*R*)-**25** (α3β2 nAChR PAM)

Figure 5. Nearest neighbor searches in chemical space. Bioactive molecules identified by MQN nearest neighbor searches with $CBD_{MQN} \leq 11$ from nicotine in GDB-13 and $CBD_{MQN} \leq 12$ in ChEMBL from PNU-282,987. The $\alpha3\beta2$ nAChR PAM (*R*)-**25** enhances the signal induced by 50 μM acetylcholine by 10-fold at 1 μM.

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

MQN-based nearest neighbor selection as a virtual screening tool recently provided us with an additional striking success in our search for $\alpha3\beta2$ nAChR PAM ligands.[93] Led by the evidence that both agonist and PAM activities on the highly similar $\alpha7$ and $\alpha3\beta2$ nAChRs was best achieved with globular quinuclidine-type tertiary amines, an MQN search was performed in ChEMBL starting from the known $\alpha7$ nAChR agonist PNU-282,987. Applying the distance constraint $CBD_{MQN} \leq 12$ previously found to ensure high pharmacophore and shape similarity gave only 115 analogs, 49 of which were 3-substituted quinuclidines. A visual inspection of these derivatives, which we named "chemical space walk" to highlight the simplicity of the exercise, revealed 2-chlorobenzyl-3-aminoquinuclidine **24** as an interesting yet unexplored compound family for nAChR. Synthesis of this and further derivatives in optically pure form and evaluation by electrophysiology led to 2-iodo derivative (*R*)-**25** as a particularly potent $\alpha3\beta2$ nAChR PAM, which is currently undergoing further pharmacological evaluation.

The above examples illustrate that nearest neighbor searches in MQN and related fingerprint spaces allow the successful exploitation of very large compound databases such as the GDB, or more directly the commercially available compounds in ZINC or the documented bioactive compounds in ChEMBL. We recently established a public web-based multifingerprint browser for the ZINC database, by which nearest neighbors of any query molecule can be retrieved from ZINC in MQN, SMIfp, Sfp, and ECfp4 spaces.[94] This browser is publicly available at www.gdb.unibe.ch and can also cluster nearest neighbors to compose a focused list for purchase and evaluation as valuable help for drug discovery projects.

CONCLUSION AND OUTLOOK

The exploration of chemical space started in the 19th century as a counting game to evaluate its size. The advent of cheminformatics and powerful computers allowed us to perform an actual enumeration of molecules to produce the chemical universe databases, GDBs. The project required chemical expertise to choose a set of criteria to select molecules with likely chemical stability and synthetic feasibility. The size of 166.4 billion structures for the currently largest database GDB-17 was mostly determined by the available computational power, data transfer rates, and memory size. Beyond enumeration, understanding and exploiting GDB led us to develop methods for virtual screening and visualization of very large databases in the form of a "periodic system of molecules" comprising six different fingerprint spaces, with web-browsers for nearest neighbor searches, and the MQN- and SMIfp-Mapplet application for exploring color-coded PC-maps of GDB and other large databases. Further insights into GDB molecules from the point of view of physical chemistry were recently reported by von Lilienfeld et al. in form of calculated electronic properties, which were predicted by machine learning methods.[95−97]

Proof-of-concept applications of GDB for drug discovery were realized by combining virtual screening with chemical synthesis and activity testing for neurotransmitter receptor and transporter ligands. One surprising lesson from using GDB for drug analog searches is the incredible depth of chemical space, that is, the fact that millions of very close analogs of any molecule are possible including scaffold-hopping molecular shape and pharmacophore analogs. These analogs can be readily identified by nearest-neighbor searches in the MQN-space of the various GDBs and may display differentiated pharmacology as exemplified with the discovery of $\alpha3\beta2$ nAChR PAM compounds discussed above.

The chemical space project has opened an unprecedented door on chemical diversity. Ongoing and yet unmet challenges concern enumerating molecules beyond 17 atoms and synthesizing GDB molecules with innovative scaffolds and pharmacophores. In the latter case, a quantum leap in the accuracy of virtual screening predictions would be most welcome since it would greatly increase the attractiveness of challenging GDB molecules for which creative synthetic routes must be designed.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jean-louis.reymond@dcb.unibe.ch. Fax: +41 31 631 80 57.

Notes

The author declares no competing financial interest.

Biography

Jean-Louis Reymond is Professor of Chemistry and Chemical Biology at the University of Bern, Switzerland. He studied chemistry and biochemistry at the ETH Zürich and obtained his Ph.D. in 1989 at the University of Lausanne in the area of natural products synthesis. He then joined the Scripps Research Institute for a postdoctoral appointment and became an assistant Professor there in 1992. In 1997, he joined the Department of Chemistry and Biochemistry at the University of Bern, where he currently serves as Director. His research focuses on computer-aided design of small molecule and peptide drugs.

ACKNOWLEDGMENTS

This work was supported financially by the University of Berne, the Swiss National Science Foundation, and the NCCR TransCure. We thank ChemAxon Pvt. Ltd. for providing free academic and web licenses for their products. The author thanks Heinz Bruggesser, Tobias Fink, Kong Thong Nguyen, Ruud van Deursen, Lorenz Blum, Lars Ruddigkeit, Julian Schwartz, Mahendra Awale, Xian Jin, and Ricardo Visini for GDB cheminformatics, Salahuddin Syed, Erika Lüthi, Noemi Garcia-Delgado, Lise Bréthous, and Justus Bürgi for the synthesis of GDB molecules, and Sonia and Daniel Bertrand for electrophysiology of nicotinic ligands and much advice and encouragement.

REFERENCES

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

(2) 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.

(3) Chen, W. L. Chemoinformatics: Past, present, and future. *J. Chem. Inf. Model.* **2006**, *46*, 2230−2255.

(4) Weininger, D. Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

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

(6) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

(7) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(8) 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.
(9) 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.
(10) Ertl, P. Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of drug-like bioisosteric groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.
(11) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823−823.
(12) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *MedChemComm* **2010**, *1*, 30−38.
(13) 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 DOI: 10.1002/wcms.1104.
(14) Reymond, J. L.; Awale, M. Exploring chemical space for drug discovery using the Chemical Universe Database. *ACS Chem. Neurosci.* **2012**, *3*, 649−657.
(15) 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.
(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) Schiff, H. Zur statistik chemischer verbindungen. *Chem. Ber.* **1875**, *8*, 1542−1547.
(18) Henze, H. R.; Blair, C. M. The number of isomeric hydrocarbons of the methane series. *J. Am. Chem. Soc.* **1931**, *53*, 3077−3085.
(19) Brinkmann, G.; Caporossi, G.; Hansen, P. A survey and new results on computer enumeration of polyhex and fusene hydrocarbons. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 842−851.
(20) Dias, J. R. The polyhex/polypent topological paradigm: regularities in the isomer numbers and topological properties of select subclasses of benzenoid hydrocarbons and related systems. *Chem. Soc. Rev.* **2010**, *39*, 1913−1924.
(21) Wieland, T.; Kerber, A.; Laue, R. Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413−419.
(22) Buchanan, B. G.; Smith, D. H.; White, W. C.; Gritter, R. J.; Feigenbaum, E. A.; Lederberg, J.; Djerassi, C. Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program. *J. Am. Chem. Soc.* **1976**, *98*, 6168−6178.
(23) 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.
(24) Warr, W. A. Computer-assisted structure elucidation. Part II: Indirect database approaches and established systems. *Anal. Chem.* **1993**, *65*, 1087A−1095A.
(25) Steinbeck, C. Recent developments in automated structure elucidation of natural products. *Nat. Prod. Rep.* **2004**, *21*, 512−518.
(26) Elyashberg, M.; Blinov, K.; Molodtsov, S.; Smurnyy, Y.; Williams, A. J.; Churanova, T. Computer-assisted methods for molecular structure elucidation: Realizing a spectroscopist's dream. *J. Cheminf.* **2009**, *1*, No. 3.
(27) Gillet, V. J.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.
(28) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649−663.
(29) Danziger, D. J.; Dean, P. M. Automated site-directed drug design: A general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc. R. Soc. London, Ser. B* **1989**, *236*, 101−113.
(30) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(31) Leach, A. R.; Hann, M. M. The in silico world of virtual libraries. *Drug Discovery Today* **2000**, *5*, 326−336.
(32) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-based approach to de novo design using reaction vectors. *J. Chem. Inf. Model.* **2009**, *49*, 1163−1184.
(33) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) space: creation of readily synthesizable design ideas automatically. *Methods Mol. Biol.* **2011**, *685*, 253−276.
(34) McKay, B. D. Practical Graph Isomorphism. *Congressus Numerantium* **1981**, *30*, 45−87.
(35) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed.* **2005**, *44*, 1504−1508.
(36) 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.
(37) 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.
(38) 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.
(39) Ruddigkeit, L.; Blum, L. C.; Reymond, J. L. Visualization and virtual screening of the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2013**, *53*, 56−65.
(40) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the rings. In silico exploration of ring universe to identify novel bioactive heteroaromatic scaffolds. *J. Med. Chem.* **2006**, *49*, 4568−4573.
(41) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *J. Med. Chem.* **2009**, *52*, 2952−2963.
(42) Sadowski, J.; Gasteiger, J. From atoms and bonds to 3-dimensional atomic coordinates - Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(43) 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.
(44) 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.
(45) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079−1087.
(46) Brown, N.; McKay, B.; Gasteiger, J. The de novo design of median molecules within a property range of interest. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 761−771.
(47) Lameijer, E. W.; Kok, J. N.; Back, T.; Ijzerman, A. P. The molecule evoluator. An interactive evolutionary algorithm for the design of drug-like molecules. *J. Chem. Inf. Model.* **2006**, *46*, 545−552.
(48) van Deursen, R.; Reymond, J. L. Chemical space travel. *ChemMedChem* **2007**, *2*, 636−640.
(49) Virshup, A. M.; Contreras-Garcia, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(50) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Persp. Drug Discovery Des.* **1998**, *9−11*, 339−353.

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

(51) 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.
(52) 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.
(53) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325−330.
(54) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.
(55) Rosen, J.; Gottfries, J.; Muresan, S.; Backlund, A.; Oprea, T. I. Novel chemical space exploration via natural products. *J. Med. Chem.* **2009**, *52*, 1953−1962.
(56) Medina-Franco, J. L.; Martinez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the chemical space in drug discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322−333.
(57) Medina-Franco, J. L.; Martinez-Mayorga, K.; Bender, A.; Marin, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477−491.
(58) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010−1024.
(59) 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.
(60) Dunkel, M.; Schmidt, U.; Struck, S.; Berger, L.; Gruening, B.; Hossbach, J.; Jaeger, I. S.; Effmert, U.; Piechulla, B.; Eriksson, R.; Knudsen, J.; Preissner, R. SuperScent—a database of flavors and scents. *Nucleic Acids Res.* **2009**, *37*, D291−D294.
(61) Arn, H.; Acree, T. E., Flavornet: A database of aroma compounds based on odor potency in natural products. In *Food flavors : formation, analysis, and packaging influences : proceedings of the 9th International Flavor Conference, the George Charalambous Memorial Symposium, Limnos, Greece, 1-4 July 1997*;Contis, E. T., Ho, C.-T., Mussinan, C. J., Parliment, T. H., Shahidi, F., Spanier, A. M., Eds.; Developments in Food Science, Elsevier: Amsterdam, 1998; Vol. *40*, p 27.
(62) Ahmed, J.; Preissner, S.; Dunkel, M.; Worth, C. L.; Eckert, A.; Preissner, R. SuperSweet—a resource on natural and artificial sweetening agents. *Nucleic Acids Res.* **2011**, *39*, D377−D382.
(63) Wiener, A.; Shudler, M.; Levit, A.; Niv, M. Y. BitterDB: A database of bitter compounds. *Nucleic Acids Res.* **2012**, *40*, D413−D419.
(64) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*, Ralph, A. W., David, C. S., Eds. Elsevier: Amsterdam, 2008; Vol. Vol. *4*, Chapter 12, pp 217−241.
(65) 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.
(66) Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(67) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.
(68) Ruddigkeit, L.; Awale, M.; Reymond, J. L. Expanding the fragrance chemical space for virtual screening. *J. Cheminf.* **2014**, *6*, 27−39.
(69) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A web-accessible molecular recognition database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719−725.
(70) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(71) 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.
(72) 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.
(73) 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.
(74) van Deursen, R.; Blum, L. C.; Reymond, J. L. A searchable map of PubChem. *J. Chem. Inf. Model.* **2010**, *50*, 1924−1934.
(75) Schwartz, J.; Awale, M.; Reymond, J. L. SMIfp (SMILES fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules. *J. Chem. Inf. Model.* **2013**, *53*, 1979−1989.
(76) Hagadone, T. R. Molecular substructure similarity searching: Efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515−521.
(77) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(78) Awale, M.; Reymond, J. L. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *J. Chem. Inf. Model.* **2014**, *54*, 1892−1897.
(79) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(80) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(81) 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.
(82) Awale, M.; Reymond, J. L. Cluster analysis of the DrugBank chemical space using molecular quantum numbers. *Bioorg. Med. Chem.* **2012**, *20*, 5372−5378.
(83) Awale, M.; van Deursen, R.; Reymond, J. L. MQN-Mapplet: Visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* **2013**, *53*, 509−518.
(84) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38*, 3743−3748.
(85) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A rule of three for fragment-based lead discovery? *Drug Discovery Today* **2003**, *8*, 876−877.
(86) 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.
(87) 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.
(88) 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.
(89) Bodnar, A. L.; Cortes-Burgos, L. A.; Cook, K. K.; Dinh, D. M.; Groppi, V. E.; Hajos, M.; Higdon, N. R.; Hoffmann, W. E.; Hurst, R. S.; Myers, J. K.; Rogers, B. N.; Wall, T. M.; Wolfe, M. L.; Wong, E. Discovery and structure-activity relationship of quinuclidine benzamides as agonists of alpha7 nicotinic acetylcholine receptors. *J. Med. Chem.* **2005**, *48*, 905−908.

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730

(90) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring α7-nicotinic receptor ligand diversity by scaffold enumeration from the chemical universe database GDB. *ACS Med. Chem. Lett.* **2010**, *1*, 422−426.
(91) 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*-(3*R*)-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.
(92) Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. Discovery of α7-nicotinic receptor ligands by virtual screening of the chemical universe database GDB-13. *J. Chem. Inf. Model.* **2011**, *51*, 3105−3112.
(93) Burgi, J. J.; Awale, M.; Boss, S. D.; Schaer, T.; Marger, F.; Viveros-Paredes, J. M.; Bertrand, S.; Gertsch, J.; Bertrand, D.; Reymond, J. L. Discovery of potent positive allosteric modulators of the α3β2 nicotinic acetylcholine receptor by a chemical space walk in ChEMBL. *ACS Chem. Neurosci.* **2014**, *5*, 346−359.
(94) Awale, M.; Reymond, J. L. A multi-fingerprint browser for the ZINC database. *Nucleic Acids Res.* **2014**, *42*, W234−W239.
(95) Rupp, M.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(96) Grégoire, M.; Matthias, R.; Vivekanand, G.; Alvaro, V.-M.; Katja, H.; Alexandre, T.; Klaus-Robert, M.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, No. 095003.
(97) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, No. 140022.

DOI: 10.1021/ar500432k
Acc. Chem. Res. 2015, 48, 722−730