doi:10.2533/chimia.2017.661 Chimia 71 (2017) 661–666

Chemical Space: Big Data Challenge for Molecular Diversity

Mahendra Awale, Ricardo Visini, Daniel Probst, Josep Arús-Pous, and Jean-Louis Reymond*

Abstract: Chemical space describes all possible molecules as well as multi-dimensional conceptual spaces representing the structural diversity of these molecules. Part of this chemical space is available in public databases ranging from thousands to billions of compounds. Exploiting these databases for drug discovery represents a typical big data problem limited by computational power, data storage and data access capacity. Here we review recent developments of our laboratory, including progress in the chemical universe databases (GDB) and the fragment subset FDB-17, tools for ligand-based virtual screening by nearest neighbor searches, such as our multi-fingerprint browser for the ZINC database to select purchasable screening compounds, and their application to discover potent and selective inhibitors for calcium channel TRPV6 and Aurora A kinase, the polypharmacology browser (PPB) for predicting off-target effects, and finally interactive 3D-chemical space visualization using our online tools WebDrugCS and WebMolCS. All resources described in this paper are available for public use at *www.gdb.unibe.ch*.

Keywords: Chemical space

Introduction

In principle, small molecule drugs are quite simple. They consist of atoms connected by covalent bonds, generally around 20 to 30 heavy atoms (non-hydrogen atoms: mostly carbon, nitrogen, oxygen, sulfur and halogens) connected by single, double or triple bonds forming combinations of rings and branches. What defines the identity and the chemical and biological properties of each molecule is the type and number of atoms involved as well as their connectivity pattern and stereochemistry. This is where things get complicated, because the connection possibilities between atoms are almost endless. Quantitative estimates range from 10E24 possible organic molecules with up to 30 atoms using known functional groups[1] to 10E60 for all drug-like molecules up to a molecular weight of 500 Daltons.[2] On the other hand, almost 200 years of synthetic organic chemistry have produced more than 100 million different compounds, most of them in the last 30 years following the industrialization of combinatorial and parallel synthesis in support of drug discovery.[3] This represents a very large number of molecules in absolute terms when considering the experimental resources needed for synthesis, but still a very small number compared to the possibilities mentioned above.

Taken together, all these molecules form the so-called chemical space, part of which is available in public databases ranging from thousands to billions of compounds (Table 1). Chemical space also describes multi-dimensional conceptual spaces in which dimensions represent properties of molecules calculated from the molecular structure and grouped in a feature vector, or fingerprint (Table 2). In such chemical spaces, each molecule is placed at the coordinates corresponding to its properties, and the distance between two molecules measures their similarity. Chemical space provides a powerful concept to address the diversity of organic molecules and exploit this diversity for various applications.

Here we discuss how to use chemical space to assist drug discovery with recent examples from our laboratory extending beyond our previous reviews.[4] First, we address the question of understanding chemical space as the ensemble of all molecules by performing an exhaustive enumeration to form the chemical universe databases (GDB: generated databases), which opens the way to unknown classes of molecules that have not yet been considered for synthesis. We highlight the need to consider relevant subsets of this very large chemical space and to focus on molecules that are structurally simple and readily accessible by chemical synthesis, such as fragrance-like and fragment-like molecules. Second, we consider chemical space as a tool to visualize and search large molecular databases. The idea is quite simple: since the physicochemical and biological properties of molecules are determined by their molecular structures, one can use the distribution of molecules in chemical space to guide a search for a particular property. Most often one starts with a reference molecule and scans a large database to identify its nearest neighbors, which are the molecules with the most similar properties to the reference molecule.

These approaches represent a typical big data problem where the limiting factor is computational power, data storage and data access capacity. Indeed, molecular databases are very large, therefore the computations necessary to organize molecules in chemical spaces and to perform proximity searches are quite resource intensive. The computational tools and databases from our group discussed here are available for public use at *www.gdb.unibe.ch*.

The Chemical Universe Databases GDB

We have undertaken a computational enumeration to understand the number of possible organic molecules and hence the scope of organic chemistry. This problem was addressed for the first time back in the

*Correspondence: Prof. Dr. J.-L. Reymond
Department of Chemistry and Biochemistry
National Center of Competence in Research NCCR TransCure
University of Bern
Freiestrasse 3, CH-3012 Bern
E-mail: jean-louis.reymond@dcb.unibe.ch

Table 1. Publicly available databases of small molecules[a].

Database	Description	Size	web addresses
DrugBank[5]	Collection of approved and experimental drugs	7895	*https://www.drugbank.ca/*
CTD[6]	Toxicogenomics database	12 K	*http://ctdbase.org/about/dataStatus.go*
NCI[7]	National cancer institute chemical database	265 K	*https://cactus.nci.nih.gov/*
BindingDB[8]	Bioactive small molecules annotated with experimental data	600 K	*https://www.bindingdb.org/bind/index.jsp*
ChEMBL[9]	Bioactive small molecules annotated with experimental data	1.7 M	*https://www.ebi.ac.uk/chembldb*
SureChEMBL[10]	Collection of patented compounds	17 M	*https://www.surechembl.org/search/*
eMolecules	Commercial small molecules for screening	7 M	*https://www.emolecules.com/*
ChemSpider	Collection of compounds from various institutions and commercial companies	58 M	*http://www.chemspider.com/*
PubChem[11]	NIH repository of molecules	93 M	*http://pubchem.ncbi.nlm.nih.gov*
ZINC 15[12]	Commercial small molecules for screening	378 M	*http://zinc15.docking.org/*
GDB-11[13]	Possible small molecules up to 11 atoms of C, N, O, F	26 M	*http://gdb.unibe.ch*
GDB-13[14]	Possible small molecules up to 13 atoms of C, N, O, S, Cl	980 M	*http://gdb.unibe.ch*
GDB-13.FL[15]	Fragrance-like subset of GDB-13	59 M	*http://gdb.unibe.ch*
GDB-17[16]	Possible small molecules up to 17 atoms of C, N, O, S and halogens	166 B	*http://gdb.unibe.ch*
FDB-17[17]	Fragment like subset of GDB-17	10 M	*http://gdb.unibe.ch*

[a]Compound numbers as of 24 June 2017

19th century with efforts to count the number of possible acyclic hydrocarbons.[22] Our aim was not only to count but also to generate the structures of all possible molecules, including cyclic and functional ones, which would be of interest for drug discovery. By combining cheminformatics tools with chemical insights, we assembled the chemical universe databases GDB listing all possible molecules following defined rules for chemical stability and synthetic feasibility. Thanks to increasing computational resources and smarter programming over the years, we expanded our initial enumeration of 26.4 million molecules up to 11 atoms (GDB-11)[23] to 977 million molecules up to 13 atoms (GDB-13)[14] and finally to 166.4 billion molecules up to 17 atoms (GDB-17).[16] Although these large databases are difficult to handle, we also succeeded in classifying them on the basis of the MQN-system (Table 2) such that similarity searches can be performed.[24]

Table 2. Fingerprints used to generate chemical spaces.

Fingerprint	Feature perceived	Description
APfp[15]	Shape	Atom-pair fingerprint. 20-dimensional scalar fingerprint, each dimension counts the number of atom pairs at one particular topological distance between 1 and 20 bonds, normalized by HAC
SMIfp[18]	Composition	SMILES fingerprint. 34-dimensional scalar fingerprint, counts 34 characters appearing in the SMILES notation of molecules
MQN[19]	Composition	Molecular Quantum Numbers. 42-dimensional scalar fingerprint, counts 42 Molecular Quantum Numbers (MQN) counting atom types, bond types, polar groups and topologies
Xfp[15]	Pharmacophore	Atom category extended atom-pair fingerprint. 55-dimensional scalar fingerprint, category extended version of APfp counting the number of category atom pairs at one particular topological distance between 0 and 10 bonds, normalized by the number of category atoms, for categories: hydrophobic atoms, H-bond donor atoms, H-bond acceptor atoms, sp2 hybridized atoms, and HBA/HBD cross-pairs
Sfp[20]	Substructure	Daylight type substructure fingerprint. 1024-dimensional binary fingerprint, perceives the presence of substructures
ECfp4[21]	Substructure	Extended connectivity fingerprint. 1024-dimensional binary fingerprint, perceives the presence of extended connectivity elements up to 4 bonds around each atom

The combinatorial enumeration procedure used to assemble the GDB databases produces the largest number of possible molecules for the biggest, most functionalized, structurally and stereochemically most complex molecular structures. By contrast, during experimental syntheses of GDB molecules we generally selected the smaller and simpler molecules to ensure rapid synthetic success.[25] Rather than to enumerate even more molecules, we have therefore recently taken steps to select GDB-subsets to propose smaller collections of molecules enriched with the simplest and synthetically most accessible compounds.

In our first implementation of this idea we addressed fragrances, which are typically small volatile molecules containing one or two functional groups with oxygen only. We defined a set of 'fragrance-likeness' criteria to constrain molecules within that property range and filtered the 977 million structures in GDB-13, which yielded a fragrance-like subset of only 59 million compounds, called GDB-13.FL.[15] In a second and most recent application of this idea we have defined a subset of fragment-like molecules from our largest database GDB-17. The very large database GDB-17 had been assembled using our typical enumeration procedure starting from mathematical graphs, whereby we first select graphs corresponding to relatively unstrained hydrocarbons, followed by introduction of unsaturation considering again ring strain criteria such as the avoidance of bridgehead double bonds, and finally substitute heteroatoms for carbons taking functional group and chemical stability criteria into account. Our fragment-like subset, called FDB-17, was then obtained by applying complexity reduction and fragment-likeness[26] criteria to GDB-17, resulting in a subset of 4.6 billion molecules. This subset was further reduced to only 10 million structures by sampling molecules evenly across molecular size, polarity and stereochemical complexity to enrich the smaller, less functionalized and stereochemically simplest structures representing the more realistic synthetic targets in GDB-17 (Fig. 1a).[17]

Starting with a known molecule of interest, one can readily search these subsets by nearest neighbor searches in various chemical spaces and identify interesting, possibly yet unknown and synthetically tractable analogs predicted to have very similar properties. We exemplify this idea for the MQN-nearest neighbors of menthone (**1**) in GDB-13.FL which comprise many related cyclic aliphatic ketones, and for a similarity search for new analogs of gabapentin (**2**) in FDB-17 involving MQN-nearest neighbors combined with 3D-shape similarity comparisons, which returns for example the yet unknown pharmacophore analogs **3**–**6** (Fig. 1b). The fragrance database GDB-13.FL and the fragment database FDB-17 are freely available for download and interactively searchable on our website.

Ligand-based Virtual Screening (LBVS) by Nearest Neighbor Searches

Virtual screening (VS) consists in applying computational models to large databases of molecules to select a limited number of compounds, typically tens to hundreds of molecules, on which to focus experimental evaluation.[27] VS saves time and resources compared to classical high-throughput screening (HTS) and can therefore be applied in projects for which large support cannot be committed, such as to assist the identification of tool compounds for targets that are not yet validated. VS can also address a much larger number and broader range of molecules than HTS, including virtual molecules that have not been synthesized yet, such as the GDB databases and their subsets.

In ligand-based virtual screening (LBVS) one performs similarity searching to identify analogs of one or several known reference compounds, typically because these reference compounds already possess the desired activity, as exemplified above for GDB-13.FL and FDB-17 (Fig. 1b).[28] LBVS is particularly well-suited for large compound databases when using the concept of nearest neighbors in a multi-dimensional chemical space. We have produced web-based searchable versions of our GDB databases and of several public databases in Table 1 by placing these databases in the chemical spaces defined by the fingerprints listed in Table 2. These web tools allow one to input a reference molecule and retrieve its nearest neighbors by similarity according to a selected fingerprint. We have programmed a particularly advanced tool for the ZINC database, which lists commercially available screening compounds from various providers.[29] In this so-called multi-fingerprint browser for the ZINC database search results can be clustered to produce small subsets of test compounds, as illustrated here for purchasable adrenaline analogs identified by ECfp4 similarity (Fig. 1c).

Virtual Screening with ZINC

While our main research interest is to use GDB for drug discovery, we have recently performed LBVS projects exploiting the ZINC database because compounds from ZINC can be purchased directly. In two recently published cases related to collaboration with the NCCR TransCure and the NCCR Chemical Biology, we focused on approximately 900,000 compounds from two vendors (Princeton Pvt. Ltd. And Otava Ltd.), and started with known reference compounds for the targets of interest. For similarity searching we used an in-house developed virtual screening algorithm called xLOS, which compares the 3D-shape and pharmacophore of molecules. This comparison involves generating the 3D-structures of molecules and optimizing the spatial overlap between the reference and each database compound by an iterative sampling and scoring procedure. This 3D similarity search, which conceptually corresponds to a nearest neighbor search in a non-Euclidean high-dimensional chemical space, is much more demanding than fingerprint-based comparisons, but is still manageable for even millions of molecules with an optimized workflow.

In the first instance, we addressed the question of discovering a submicromolar and selective inhibitor of TRPV6, a calcium channel overexpressed in various cancers, to clarify whether pharmacological blocking of this channel might offer an option to control cell growth as suggested by si-RNA experiments.[30] Starting with the known but very weak and non-selective TRPV6 inhibitor **7**, we performed two successive rounds of LBVS and identified a new lead series based on a cyclohexylpiperazine based scaffold, which we optimized by classical medicinal chemistry. Our best compound **8** showed high selectivity for TRPV6 among other calcium channels (Fig. 1d). Comparison with its less active *trans*-stereoisomer showed that **8** selectively but only marginally reduced cell growth in TRPV6 overexpressing cancer cells.

In the second instance, we used the same approach as above to identify a new and selective inhibitor of the anti-cancer target Aurora A kinase starting from known kinase inhibitors.[31] We hoped to exploit the fact that 3D-shape and pharmacophore similarity searching ignores the structural details of the molecules and therefore allows to identify 'scaffold-hopping' analogs.[32] Our LBVS workflow allowed us to identify a thiazolidinone hit compound with submicromolar inhibition of Aurora A. We then solved the crystal structure of its complex with Aurora A to identify its binding mode, and designed the optimized analog **9** with a single digit nanomolar potency against Aurora A (Fig. 1e). Our inhibitor **9** was highly selective for Aurora kinases among over 400 other human kinases, and induced a selective Aurora A inhibition phenotype in cells. Interestingly **9** was flagged as being a so-called PAINS (pan-assay interference)[33] compound due

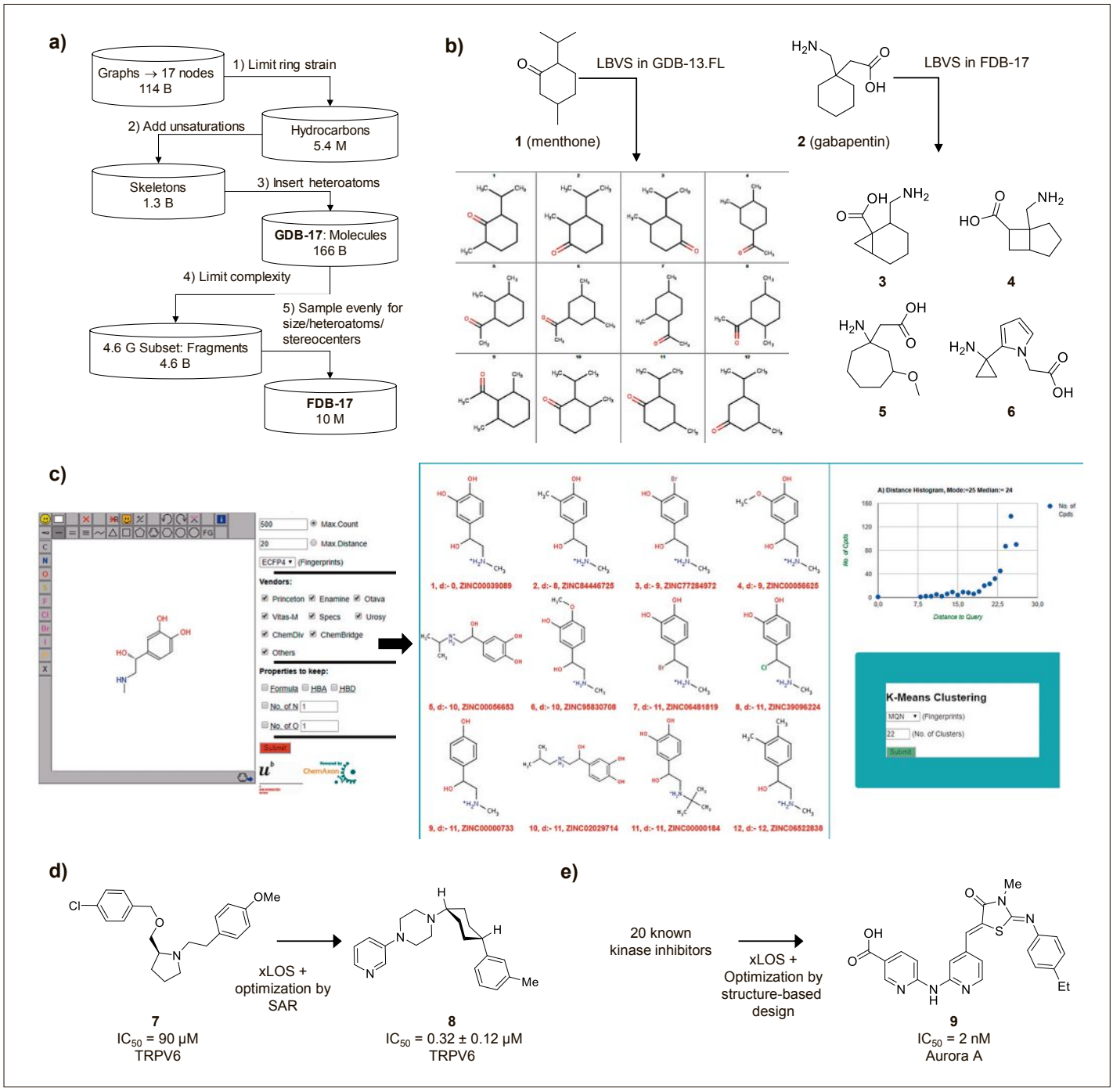

Fig. 1. a) Computation of the GDB-17 database from graphs and selection of the fragment-like subset FDB-17. b) Example of similarity searches in GDB-13.FL and in FDB-17. c) Identification of nearest neighbors of adrenaline in ZINC by substructure similarity using the multi-fingerprint browser for ZINC. The ECfp4 nearest neighbors are shown in the primary output window. They can be further clustered by K-means clustering using different fingerprints. d,e) Discovery of TRPV6 and Aurora A kinase inhibitors by LBVS using the 3D-shape and pharmacophore similarity algorithm xLOS followed by lead optimization by structure-activity relationship (SAR) or by structure-based design.

to the presence of the potentially electrophilic exocyclic double bond, however we did not detect any significant reactivity with glutathione or any hint of a non-selective activity.

Target Prediction with the Polypharmacology Browser (PPB)

Most if not all drugs interact with multiple targets, a phenomenon known as polypharmacology.[34] By performing similarity searches in databases containing information on the activity of small molecule drugs and their protein targets one can predict the possible off-targets of a hit compound or drug candidate. The challenge here is to perform a similarity search in an annotated database using a similarity method which is fast yet identifies relevant compounds in terms of their shared biological activity. We have constructed such a target prediction tool, called the polypharmacology browser (PPB), which uses data from the ChEMBL database, and predicts targets based on similarities computed from compounds with over 4000 different possible targets, using 10 different types of fingerprints.[35]

Although the computation requires a 10-fold similarity search with almost one million compounds, we developed highly efficient code, such that the search with PPB is complete in less than one minute. In general PPB delivers a broader set of results than other comparable web tools. As an application example, we have used PPB to predict the off-targets of the TRPV6 inhibitor **8** dis-

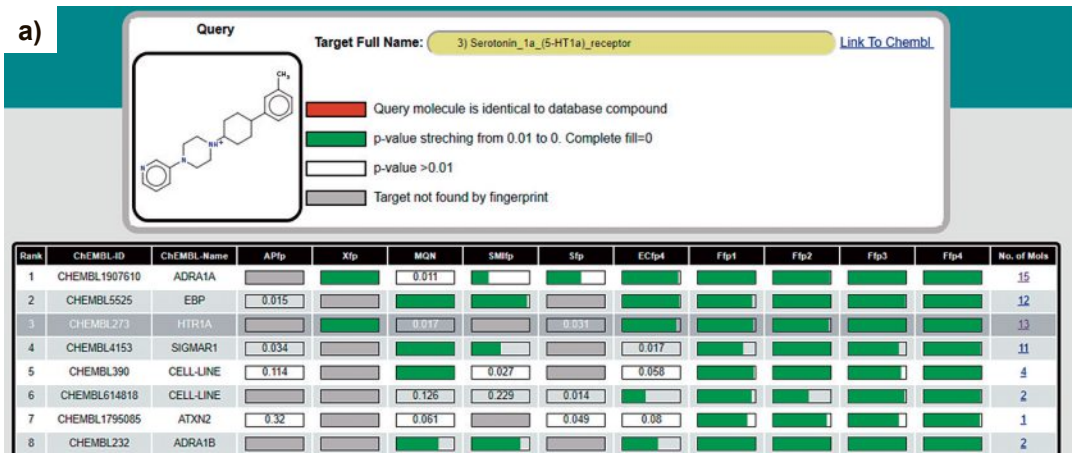

Fig. 2. Off-target predictions by the polypharmacology browser PPB. a) Prediction of off-targets for the TRPV6 inhibitor **8**. b) Experimentally confirmed targets of **8**. Green color: targets predicted by PPB. Target full names: Adrenergic α1A (ADRA1A), Cholinergic muscarinic receptor 1 (CHRM1) and 2 (CHRM2), Dopamine receptor subtypes D1-4 (DRD1-4), 5-hydroytryptamine receptor 1A (HTR1A) and 2A (HTR2A), Voltage gated potassium channel subfamily H member 2 (HERG), μ opioid receptor (OPRM) and voltage gated Na^+ channel (SCN2A).

cussed above. Actual measurement proved that the compound also binds to adrenergic, dopamine, and 5-hydroxy-tryptamine receptors, as predicted by PPB (Fig. 2).

Visualizing Chemical Space

Methods for visualizing chemical spaces address the problem of interfacing our own human brain with large compound databases to provide intuitive insights and informed user-defined choices, which are unavoidable and often critical within drug discovery projects.[36] In our approach to visualize multi-dimensional chemical spaces we use dimensionality reduction to obtain 2D or 3D maps.[37] We then bin these maps into 2D- or 3D-pixels (voxels) at a chosen resolution, and color-code each pixel according to various properties computed for the molecules in the associated bin. The resulting image is finally converted to an interactive format for on-screen display such that the 2D-structures of the molecules in each pixel are shown on-screen on mouse-over.[38] We have produced several interactive color-coded 2D-maps in the form of downloadable Java-applets called 'mapplets' for our GDB and various other databases. These maps represent chemical spaces generated from the fingerprints in Table 2 and were obtained by principal component analysis (PCA) or similarity mapping as a dimensionality reduction method.[15,17,18,38,39]

Most recently we turned our attention to producing a web-based version of these Java applets that can be used within a web-browser from any platform including not only computers but also tablets and cell phones. In this case we implemented 3D-chemical space maps because they produce a much better distribution of compounds in voxels and allow to render more complex chemical spaces. Our first implementation, called WebDrugCS, renders all molecules in DrugBank in 3D-spaces obtained by PCA of the chemicals spaces in Table 2 (Fig. 3a).[40] In a second implementation of this approach, we have produced a related online tool called WebMolCS to render 3D-chemical spaces produced by either PCA or similarity mapping for up to 5000 molecules defined by the user.[41] This tool is particularly useful to visualize the structural diversity in nearest neighbor selections from our GDBs, as illustrated for the case of 5000 MQN-nearest neighbors of nicotine selected from GDB-13 (Fig. 3b).[42]

Conclusion and Outlook

Chemical space provides an organizing concept to understand, exploit and visualize very large databases of molecules in support for drug discovery projects. The approach is particularly useful because databases in excess of hundreds of millions of molecules are becoming more and more common, including not only our GDB databases, but also the expanded ZINC database now listing over 300 million compounds that can be synthesized on demand.[12] Even much larger compound databases might require a chemical space based analysis in the future, such as compounds in DNA-encoded libraries reaching potentially trillions of molecules or more.[43] Understanding the

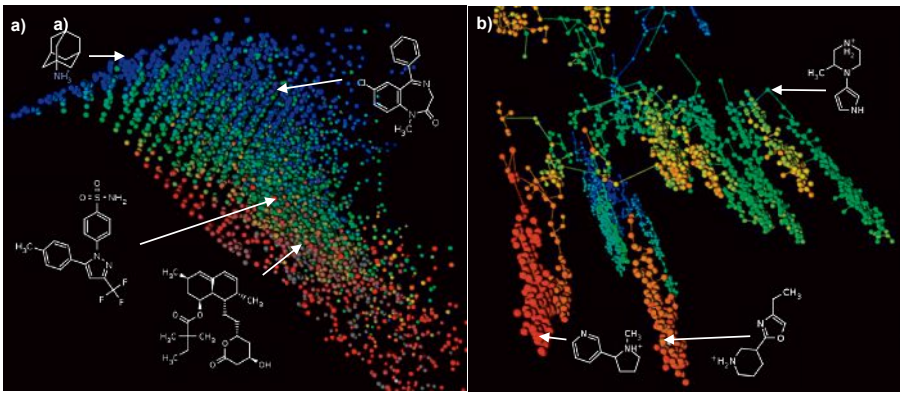

Fig. 3. Visualizing chemical space. a) WebDrugCS: visualization of DrugBank (see Table 1) in the chemical space of the APfp fingerprint (see Table 2) projected in 3D by PCA. The map is color-coded by increasing rotatable bond count (RBC) per molecule, which represents structural rigidity, from blue (RBC = 0) to magenta (max. value). Example drugs are shown with arrows pointing to the corresponding voxel: amandatine (upper left), diazepam (upper right), celecoxib (lower left) and simvastatin (lower right). b) WebMolCS: visualization of 5000 MQN-nearest neighbor of nicotine from GDB-13 in the chemical space of the Xfp fingerprint (see Table 2) projected in 3D by similarity mapping. The map is color-coded by Xfp-similarity to nicotine from blue (lowest similarity) to red (highest similarity).

structural diversity of such large compound collections represents a significant technical challenge in terms of data handling, but also an enormous opportunity to improve our understanding of chemical space and enable the discovery of new drugs.

Acknowledgements

This work was supported financially by the Swiss National Science Foundation, NCCR TransCure, and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676434, 'Big Data in Chemistry' ('BIGCHEM', *www.bigchem.eu*).

Received: July 4, 2017

[1] P. Ertl, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374.
[2] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3.
[3] a) L. M. Mayr, D. Bojanic, *Curr. Opin. Pharmacol.* **2009**, *9*, 580; b) S. Renner, M. Popov, A. Schuffenhauer, H. J. Roth, W. Breitenstein, A. Marzinzik, I. Lewis, P. Krastel, F. Nigsch, J. Jenkins, E. Jacoby, *Future Med. Chem.* **2011**, *3*, 751.
[4] a) J. L. Reymond, R. Van Deursen, L. C. Blum, L. Ruddigkeit, *MedChemComm* **2010**, *1*, 30; b) J.-L. Reymond, L. C. Blum, R. van Deursen, *Chimia* **2011**, *65*, 863; c) J. L. Reymond, L. Ruddigkeit, L. C. Blum, R. Van Deursen, *WIREs comput. Mol. Sci.* **2012**, doi: 10.1002/wcms.1104; d) J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722.
[5] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, D. S. Wishart, *Nucleic Acids Res.* **2014**, *42*, D1091.
[6] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, C. J. Mattingly, *Nucleic Acids Res.* **2009**, *37*, D786.
[7] J. H. Voigt, B. Bienfait, S. Wang, M. C. Nicklaus, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702.
[8] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, *Nucleic Acids Res.* **2016**, *44*, D1045.
[9] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100.
[10] G. Papadatos, M. Davies, N. Dedman, J. Chambers, A. Gaulton, J. Siddle, R. Koks, S. A. Irvine, J. Pettersson, N. Goncharoff, A. Hersey, J. P. Overington, *Nucleic Acids Res.* **2016**, *44*, D1220.
[11] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, S. H. Bryant, *Nucleic Acids Res.* **2016**, *44*, D1202.
[12] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324.
[13] T. Fink, H. Bruggesser, J. L. Reymond, *Angew. Chem. Int. Ed.* **2005**, *44*, 1504.
[14] L. C. Blum, J. L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*, 8732.
[15] L. Ruddigkeit, M. Awale, J. L. Reymond, *J. Cheminform.* **2014**, *6*, 27.
[16] L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*, 2864.
[17] R. Visini, M. Awale, J.-L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 700.
[18] J. Schwartz, M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 1979.
[19] K. T. Nguyen, L. C. Blum, R. van Deursen, J.-L. Reymond, *ChemMedChem* **2009**, *4*, 1803.
[20] T. R. Hagadone, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515.
[21] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742.
[22] E. Cayley, *Chem. Ber.* **1875**, *8*, 1056.
[23] T. Fink, J. L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*, 342.
[24] a) L. C. Blum, R. van Deursen, J. L. Reymond, *J. Comput.-Aided Mol. Des.* **2011**, *25*, 637; b) L. Ruddigkeit, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 56.
[25] a) K. T. Nguyen, S. Syed, S. Urwyler, S. Bertrand, D. Bertrand, J. L. Reymond, *ChemMedChem* **2008**, *3*, 1520; b) E. Luethi, K. T. Nguyen, M. Burzle, L. C. Blum, Y. Suzuki, M. Hediger, J. L. Reymond, *J. Med. Chem.* **2010**, *53*, 7236; c) N. Garcia-Delgado, S. Bertrand, K. T. Nguyen, R. van Deursen, D. Bertrand, J.-L. Reymond, *ACS Med. Chem. Lett.* **2010**, *1*, 422; d) L. Brethous, N. Garcia-Delgado, J. Schwartz, S. Bertrand, D. Bertrand, J. L. Reymond, *J. Med. Chem.* **2012**, *55*, 4605.
[26] M. Congreve, R. Carr, C. Murray, H. Jhoti, *Drug Discov. Today* **2003**, *8*, 876.
[27] T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martinez-Mayorga, T. Langer, K. Cuanalo-Contreras, D. K. Agrafiotis, *J. Chem. Inf. Model.* **2012**, *52*, 867.
[28] H. Geppert, M. Vogt, J. Bajorath, *J. Chem. Inf. Model.* **2010**, *50*, 205.
[29] M. Awale, J. L. Reymond, *Nucleic Acids Res.* **2014**, *42*, W234.
[30] C. Simonin, M. Awale, M. Brand, R. van Deursen, J. Schwartz, M. Fine, G. Kovacs, P. Häfliger, G. Gyimesi, A. Sithampari, R. P. Charles, M. Hediger, J. L. Reymond, *Angew. Chem. Int. Ed.* **2015**, *54*, 14748.
[31] F. Kilchmann, M. J. Marcaida, S. Kotak, T. Schick, S. D. Boss, M. Awale, P. Gönczy, J.-L. Reymond, *J. Med. Chem.* **2016**, *59*, 7188.
[32] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894.
[33] a) S. J. Capuzzi, E. N. Muratov, A. Tropsha, *J. Chem. Inf. Model.* **2017**, *57*, 417; b) S. Jasial, Y. Hu, J. Bajorath, *J. Med. Chem.* **2017**, *60*, 3879.
[34] a) J. P. Overington, B. Al-Lazikani, A. L. Hopkins, *Nat. Rev. Drug Discov.* **2006**, *5*, 993; b) A. Anighoro, J. Bajorath, G. Rastelli, *J. Med. Chem.* **2014**, *57*, 7874; c) A. Lavecchia, C. Cerchia, *Drug Discov. Today* **2016**, *21*, 288.
[35] M. Awale, J. L. Reymond, *J. Cheminform.* **2017**, *9*, 11.
[36] a) J. L. Medina-Franco, R. Aguayo-Ortiz, *Mol. Inform.* **2013**, *32*, 942; b) T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model.* **2015**, *55*, 460.
[37] a) M. Reutlinger, G. Schneider, *J. Mol. Graph. Model.* **2012**, *34*, 108; b) H. A. Gaspar, I. I. Baskin, G. Marcou, D. Horvath, A. Varnek, *J. Chem. Inf. Model.* **2014**, *55*, 84.
[38] M. Awale, R. van Deursen, J. L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 509.
[39] M. Awale, J. L. Reymond, *J. Chem. Inf. Model.* **2015**, *55*, 1509.
[40] M. Awale, J. L. Reymond, *J. Cheminform.* **2016**, *8*, 25.
[41] M. Awale, D. Probst, J. L. Reymond, *J. Chem. Inf. Model.* **2017**, *57*, 643.
[42] L. C. Blum, R. van Deursen, S. Bertrand, M. Mayer, J. J. Burgi, D. Bertrand, J. L. Reymond, *J. Chem. Inf. Model.* **2011**, *51*, 3105.
[43] G. Zimmermann, D. Neri, *Drug Discov. Today* **2016**, *21*, 1828.