JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

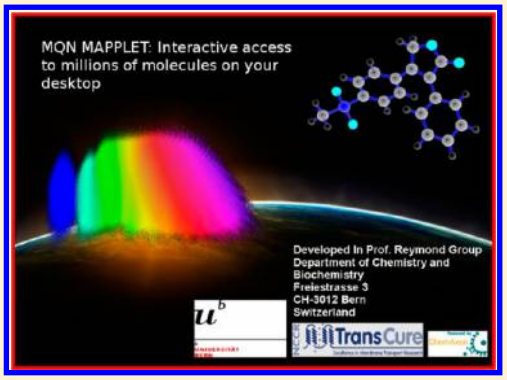
MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13

Mahendra Awale,[†] Ruud van Deursen,[‡] and Jean-Louis Reymond*[,†]

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

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

ABSTRACT: The MQN-mapplet is a Java application giving access to the structure of small molecules in large databases via color-coded maps of their chemical space. These maps are projections from a 42-dimensional property space defined by 42 integer value descriptors called molecular quantum numbers (MQN), which count different categories of atoms, bonds, polar groups, and topological features and categorize molecules by size, rigidity, and polarity. Despite its simplicity, MQN-space is relevant to biological activities. The MQN-mapplet allows localization of any molecule on the color-coded images, visualization of the molecules, and identification of analogs as neighbors on the MQN-map or in the original 42-dimensional MQN-space. No query molecule is necessary to start the exploration, which may be particularly attractive for nonchemists. To our knowledge, this type of interactive exploration tool is unprecedented for very large databases such as PubChem and GDB-13 (almost one billion molecules). The application is freely available for download at www.gdb.unibe.ch.



INTRODUCTION

Organic small molecules play an essential role in modern technology in the form of drugs for human and animal diseases, weed and pest control agents for agriculture, and building blocks and additives for polymers. One of the striking features of organic small molecules is their sheer number, which has become particularly impressive in recent years since the implementation of combinatorial and parallel synthesis to support academic and industrial drug discovery.[1−3] The combined corporate, academic, and commercial collections worldwide probably total over 100 million different small molecules.[4−7] Various open access databases collect the structures, properties, and activities of thousands to millions of molecules of medicinal interest, such as DrugBank (>6000 experimental or approved drugs),[8] ChEMBL (>1.1 million compounds with documented bioactivity),[9] PubChem, and Chemspider (>32 million reported molecules).[10] Many more molecules are theoretically possible following basic rules for covalent bonds and functional groups and have been in part enumerated in the chemical universe database GDB-11, which lists 26.4 million molecules up to 11 atoms of C, N, O, and F that are possible following simple rules of chemical stability,[11] and GDB-13, which similarly lists 977 million molecules up to 13 atoms of C, N, O, Cl, and S.[12,13] The total number of possible drug-like small molecules is probably even much larger and has been estimated to reach 10^{60} compounds.[1,7,13,14]

Visualizing the content of large databases of molecules represents a particular challenge which can be addressed by a variety of methods.[15−19] The concept of "chemical space" addresses this challenge by defining multidimensional property spaces in which dimensions are assigned to selected numerical descriptors of molecular structure, for example the Lipinski "rule of 5" parameters.[20] Molecules are placed in these property spaces at the coordinates corresponding to their property values, which generates a spatial distribution of compounds, as originally described by Pearlman and Smith.[21] One can then perform principal component analysis (PCA) to project this multidimensional property space in a lower dimensionality space, typically a 2D- or 3D-space which can be visualized. This approach has been used by various groups to compare compound collections such as drugs, screening compounds, and natural products.[22−28] If a property space is well chosen, compounds of similar bioactivity are often grouped together, and the distance between compounds can be used as sorting function for ligand-based virtual screening (LBVS).[17,29,30] Herein we report the MQN-mapplet, a Java application that allows to visualize the molecules contained in color-coded two-dimensional representations of the property spaces of large databases. The application is exemplified for PCA maps representing projections from the MQN-property space, which is constructed from 42 integer value descriptors calculated from the structural formula[31] called Molecular Quantum Numbers (MQN).[32] The MQN-mapplet provides efficient interactive access to millions of molecules, allowing one to browse through large compound collections in a rapid and intuitive manner. The application is freely available for download at www.gdb.unibe.ch. The paper reviews key features

Received: October 24, 2012
Published: January 8, 2013

   dx.doi.org/10.1021/ci300513m | *J. Chem. Inf. Model.* 2013, 53, 509−518

of the MQN descriptors and their PCA maps and describes the basic functionalities of the MQN-mapplet allowing visualization of large databases. Current limitations are also discussed. A more detailed description is available in the HELP page of the application. While similar interactive visualization tools have been reported previously, they lack the use of the intuitive color-coded maps and are limited to displaying datasets of hundreds to a few thousands of molecules only.[33,34]

RESULTS AND DISCUSSION

MQN-Space. The 42 MQNs count different categories of atoms, bonds, polar groups, and topological features of molecules and categorize molecules not by substructure,[35] but by their size, rigidity, and polarity (Table 1). The spatial relationship between compounds in MQN-space as measured by the city-block distance CBD_{MQN} allows extremely rapid recovery of nearest neighbors even for very large databases such as PubChem and GDB-13 using a freely available search tool accessible at www.gdb.unibe.ch.[36] Despite its simplicity, CBD_{MQN} is relevant to biological activities, for instance excellent enrichment factors are obtained for recovering from PubChem various active series listed in DUD[37] or ChEMBL.[38−41] CBD_{MQN}-similarity searching reveals scaffold-hopping relationships between actives that are not seen by substructure similarity searching. Nearest neighbors in MQN-space are often similar in molecular shape as measured by the OpenEye scoring function ROCS (Rapid Overlay of Chemical Structures), a well validated tool for LBVS.[42,43] MQN-similarity has been successfully applied in prospective drug discovery by the identification of new nicotine analogs from GDB-13.[44]

Table 1. Forty-Two Molecular Quantum Numbers

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

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

MQN-Maps. Similarly to other property spaces, MQN-space can be projected into principal component planes for visualization. The MQN-map of the (PC1, PC2)-plane usually covers over 60% of a database variability and thus provides a relevant overview of the database contents. Molecules are usually distributed according to size for PC1 and rigidity/cycles for PC2, while polarity appears strongly in PC3. The MQN-determination and PCA was carried out with seven different large databases representing the publicly available chemical space. MQN-maps were realized for the (PC1, PC2)-plane, with PC1 as the horizontal axis and PC2 as the vertical axis. For the "PubChem.extended" database, the (PC2,PC3)-plane was used because it allowed a more insightful visualization than the (PC1, PC2)-plane by spreading molecules by rigidity/cycles (PC2) and polarity (PC3), while molecules of increasing size (PC1) appear in concentric circles around the center of the map.[38] In the case of the PubChem.60 and PubChem.extended, a nonlinear distortion of the selected PC-plane was carried out to spread the densely occupied central part of the map and shrink the sparsely occupied outer areas, such as to better spread the molecules over the available pixels.

The database size and pixel-occupancy statistics of the various MQN-maps are listed in Table 2. The loadings of the first three principal components are shown in Figure 1, and a selection of color-coded MQN-maps used in the MQN-mapplet are shown in Figures 2−4. Color-coding uses the HSL-system with the hue (H) coding for the average property value from blue (lowest value) to purple or red (highest value) via the sequence cyan−green−yellow−orange−red (intermediate values).[40] The saturation to gray (S) is used to represent the standard deviation of a property for each pixel, such that no color appears if the value is not well-defined. In the case of the MQN-maps of the GDB databases, the lightness value (L; fading to background color) was used to code the pixel occupancy to de-emphasize the image edges which contain low occupancy pixels that are almost insignificant compared to the more densely populated areas of the GDB maps, a problem that does not occur in the smaller databases. The average value and standard deviation corresponding to each color-coding for each pixel are displayed interactively when using the MQN-mapplet.

MQN-Mapplet Functions and Uses. The MQN-mapplet graphical user interface is written in Java. Molecule display and MQN-calculations are enabled by the JChem Java library from ChemAxon Ltd. The MQN-mapplet is primarily a visualization tool allowing to look at the structural formula of molecules and familiarize oneself with the content of large databases in an efficient manner. Inspecting molecules by their structural formula is meaningful because the structural formula is the most broadly used graphical representation of molecules and the basic tool for teaching and working in organic chemistry; in many contexts, the structural formula "is" the molecule. MQN-maps spread molecules by global features such as the molecule size, the number of cycles, and the number of polar atoms. Similarly to the "Google maps" system for geography, the MQN-mapplet does not require a specific query to initiate the exploration, and is suitable for nonchemists. The various

Table 2. Contents of Databases and Their MQN-Map as Used in the MQN-Mapplet

database[a]	no. of cpds[b]	PC1[c]	PC2[c]	PC3[c]	no. of occupied pixels[d]	pixel occupancy aver/max
DrugBank	6404	63%	18%	8.6%	2991	2.1/21
ChEMBL.50[e]	1094469	55%	22%	7.1%	160054	6.8/104
PubChem.60[f]	24498884	61%	20%	5.4%	502406	49/2524
PubChem.extended[g]	25601906	70%	16%	4.3%	302885	85/2786
GDB-11	26434540	50%	14%	9.4%	355421	74/2586
GDB-13 subset[h]	43729742	55%	13%	7.6%	262715	170/4283
GDB-13	977460392	51%	12%	9.0%	387708	2500/22335

[a]Databases downloaded in June 2012 from the following Web sites: http://www.drugbank.ca; https://www.ebi.ac.uk/chembldb; http://pubchem.ncbi.nlm.nih.gov; www.gdb.unibe.ch. [b]The entries were converted to SMILES without counterions and duplicates were removed. [c]Variance of the first three PC for PCA analysis of the 42-dimensional MQN data set; see Figure 1 for PC loadings. [d]The MQN-maps are 1 megapixel except for DrugBank which is 90 000 pixels. [e]ChEMBL.50 contains all molecules up to 50 heavy atoms (96% of ChEMBL). [f]PubChem.60 contains all molecules up to 60 heavy atoms (99% of PubChem). [g]PubChem.extended contains the entire PubChem (24.7 million cpds) plus virtual categories: peptides, oligonucleotides, oligosaccharides, diamondoids, graphenes, and acyclic alkanes up to Hac = 500 (0.9 million cpds); see ref 38. The MQN-map of PubChem.extended shows the (PC2,PC3)-plane. [h]The GDB-13 subset lists 43.7 million structures of GDB-13 without 3- or 4-membered rings and free of synthetically and/or metabolically problematic functional groups such as aldehydes, esters, etc.; see ref 40.

functions and their possible uses for visualization and LBVS are outlined in the following paragraphs.

Browsing MQN-Maps. A view of the mapplet as displayed on screen is shown in Figure 5A, corresponding to the entry display with the MQN-map of DrugBank color-coded by the number of rings per molecule. One first selects one of the seven databases listed in Table 2 using the "DB" pull-down list. For each database various color-coded MQN-maps can be selected from a second pull-down list under "map". When the pointer passes over a pixel of the selected MQN-map, the "Average Molecule" window shows the molecule closest to the average MQN-values in that pixel (the message "molecule is too large to display" appears for molecules larger than 60 heavy atoms to avoid slowing down the browsing). Simultaneously, the average value and standard deviation of the property corresponding to the selected color-code in that pixel are displayed in the "Average" and "Deviation" text fields. One can zoom in/out of the MQN-image using the "zoom+" and "zoom−" buttons or using the mouse wheel, which allows one to see the details of the MQN-map and precisely position the pointer on a pixel of choice. To maintain the overview while zooming one can optionally open the "Map Trace" window, which displays the miniature map with a square indicating the magnified region.

For example, inspecting with the mapplet the MQN-map of DrugBank color-coded by the number of rings (Figure 2 upper right pannel), one can see that the smaller acyclic molecules in DrugBank (central part of the bottom diagonal blue stripe) are mostly amino acids and carboxylic acids, while many lipids appear in the larger acyclic molecules (right portion of the blue stripe). Switching to the "ring atom" color-coding offers further perspectives, for example a few cyan pixels stand out just above the acyclic blue stripe containing molecules with three-membered rings (cyclopropanes, epoxides), which would otherwise be difficult to find. Further switching to the rotatable bond count (rbc) color coding shows that a significant number of molecules in DrugBank are entirely rigid (blue pixels in the rbc map). This map also shows that many molecules in DrugBank have ten or more rotatable bonds (red pixels in rbc map), and the H-bond acceptor atoms (hba) color-coding mode shows that many of these flexible molecules have more than 10 hba (red pixels in hba map). Looking at the structural formula of all these rather "non drug-like" molecules in DrugBank shows that many of them are oligosaccharides and peptides.

Browsing the ChEMBL.50 and PubChem.60 maps with the mapplet shows that the layout and molecule types in these databases are comparable to DrugBank. In PubChem.60, color-coding by pixel occupancy (Figure 3, upper left pannel) reveals that molecules are particularly abundant in the fourth diagonal stripe from bottom (red area). Adjusting full zoom to this region and switching between the various color-coded modes shows that this highly populated region of the PubChem.60 map contains molecules with 3 cycles, 16 ring atoms, 6 rotatable bonds, 20 carbons atoms, 4 H-bond acceptor atoms, and 26 heavy atoms. The MQN-mapplet shows that these molecules usually contain two benzene rings, one 5- or 6-membered heterocycle, and one or two carbonyl/sulfonyl groups often part of amides, sulfonamides, or ureas.

PubChem.extended contains the entire PubChem (24.7 million cpds) plus virtual categories: peptides, oligonucleotides, oligosaccharides, diamondoids, graphenes, and acyclic alkanes up to HAC = 500 (0.9 million cpds) as described in an earlier MQN-analysis of PubChem.[38] The map shown in the MQN-mapplet is the (PC2, PC3)-plane which offers an alternative view of PubChem spreading molecules by rigidity (horizontal axis) and polarity (vertical axis). Similarly to PubChem.60 however, browsing over PubChem.extended shows that the majority of PubChem molecules contain two aromatic rings and one heterocycle connected by flexible single-bond linkers often including amides and ethers, highlighting a limited structural diversity which is often perceived as a limitation of the known chemical space.

The GDB-11, GDB-13, and GDB-13 subset images are organized similarly to PubChem.extended by rigidity (PC1, horizontal axis) and polarity (PC2, vertical axis) because molecular size does not play a role in these databases. Browsing the GDBs with the MQN-mapplet reveals the absence of the aromatic systems characteristic of PubChem, and an overwhelming abundance of fused heterocyclic amines and ethers including many small ring compounds, which form the bulk of the enumerated chemical space. The direct visualization of these rather unusual molecules will be most inspiring for synthetic chemists in search for new challenges.

Accessing the Entire Database Contents. When the pointer is positioned over a pixel, one can view the entire list of molecules in that pixel using the "Show Bin" option in the "right-click" menu. This function requires an active Internet connection to enable downloading the SMILES list for pixel-contents from the application server. The molecules are sorted

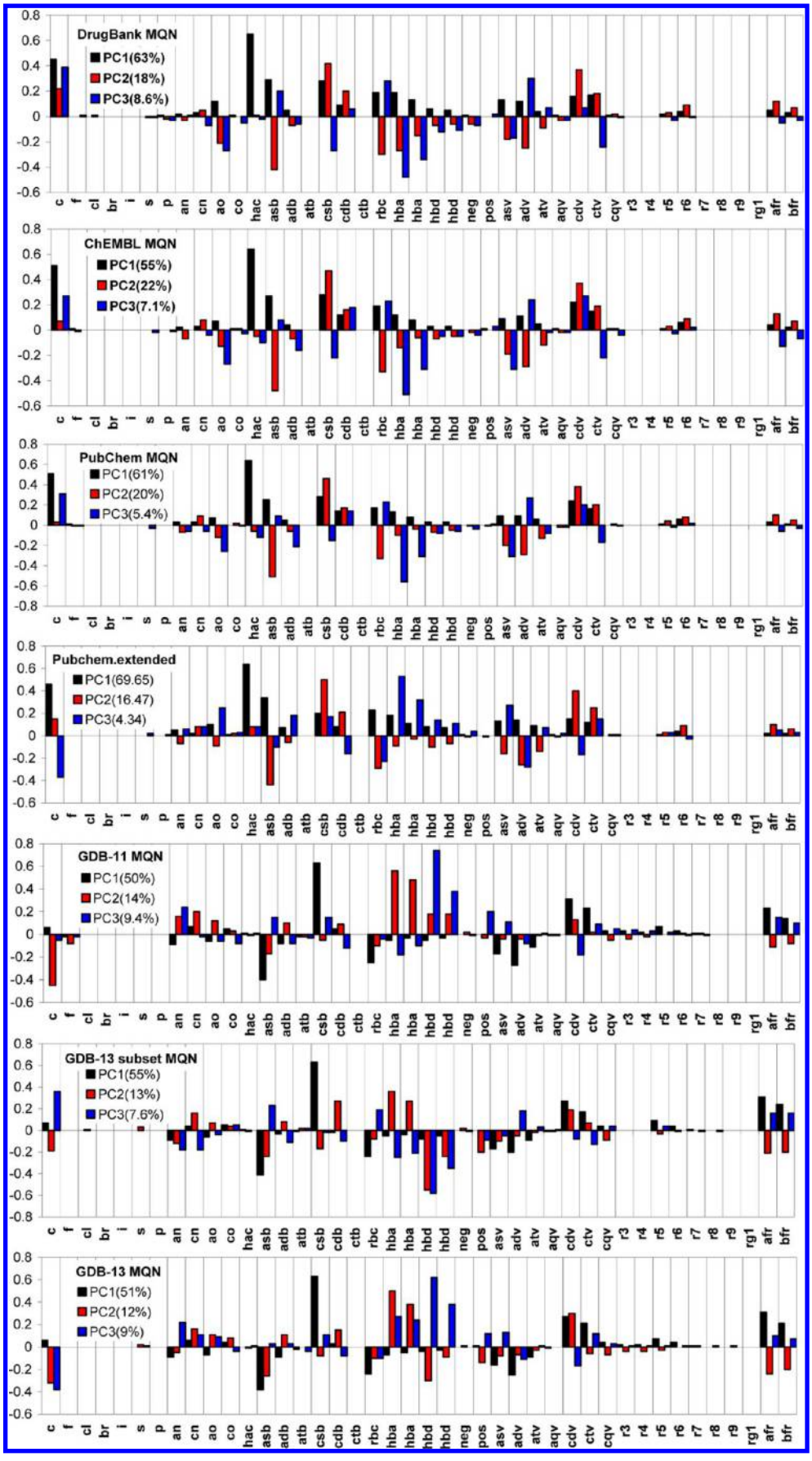


Figure 1. PC loadings of the first three PCs for PCA analysis of MQN data set of various databases.

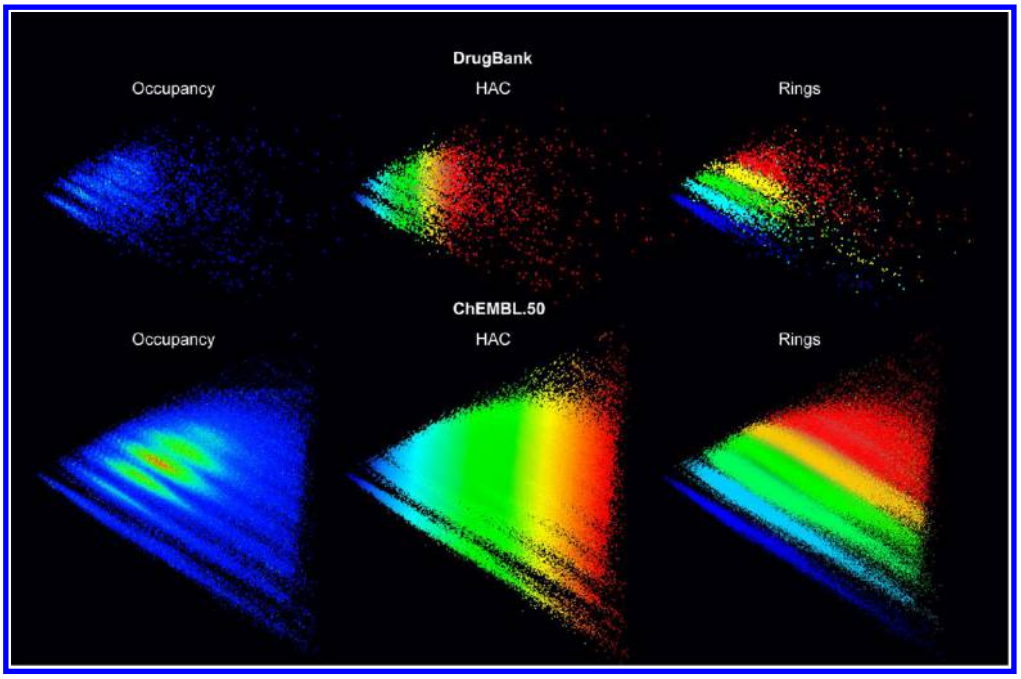


Figure 2. MQN-maps of (PC1,PC2)-plane of DrugBank and ChEMBL.50 (molecules in ChEMBL up to heavy atom count (HAC) = 50). Color-coding from blue (lower values) to red (highest value). In these maps, PC1 shows molecule size and PC2 shows the rigidity.

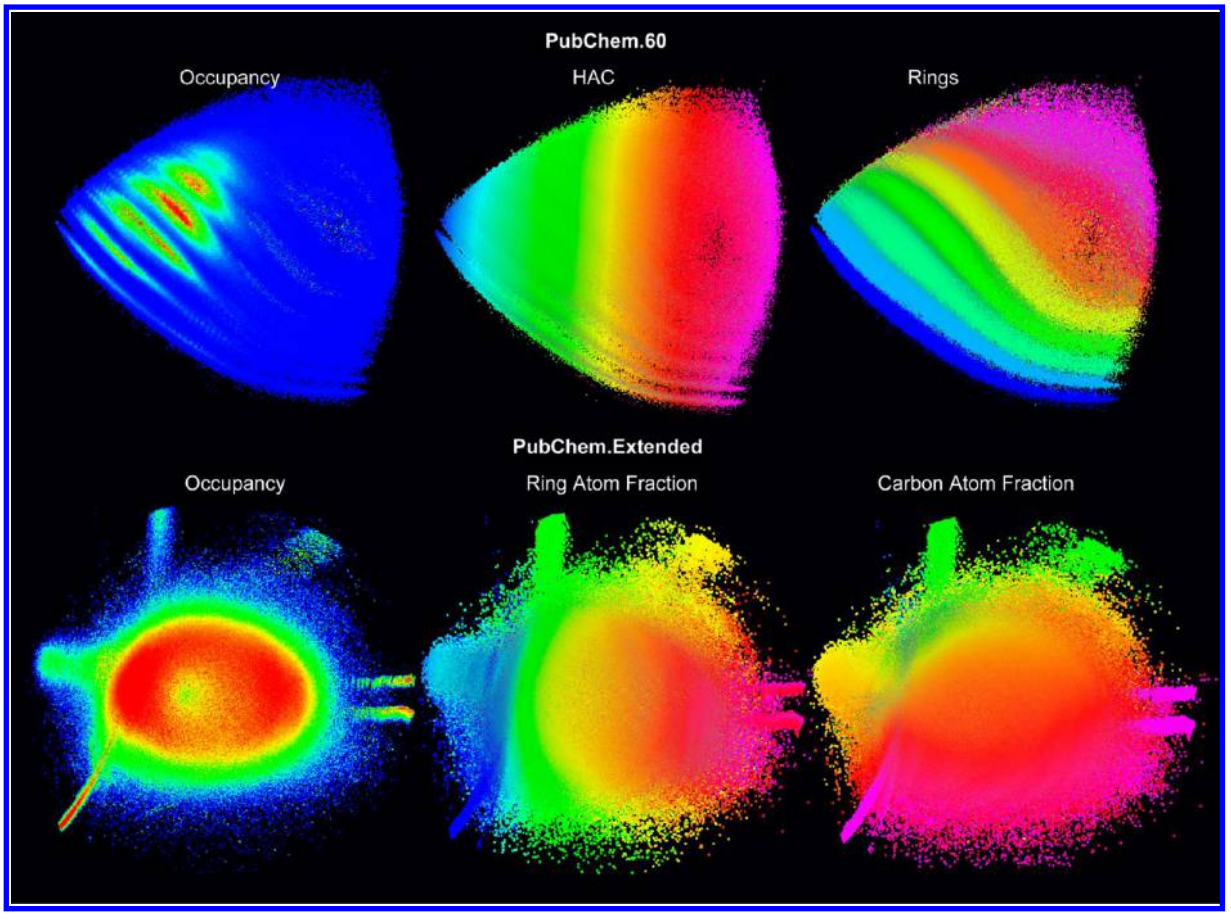


Figure 3. MQN-maps of PubChem.60 (molecules in PubChem up to HAC = 60) in the (PC1, PC2)-plane (size, rigidity) and PubChem.extended (PubChem together with enumerated oligomers; see legend of Table 2 and ref 14) in the (PC2, PC3)-plane (rigidity, polarity). The planes have been distorted for better spreading of molecules on the available pixels (see Methods). Color-coding from blue (lower values) to magenta (highest value).

dx.doi.org/10.1021/ci300513m | *J. Chem. Inf. Model.* 2013, 53, 509−518

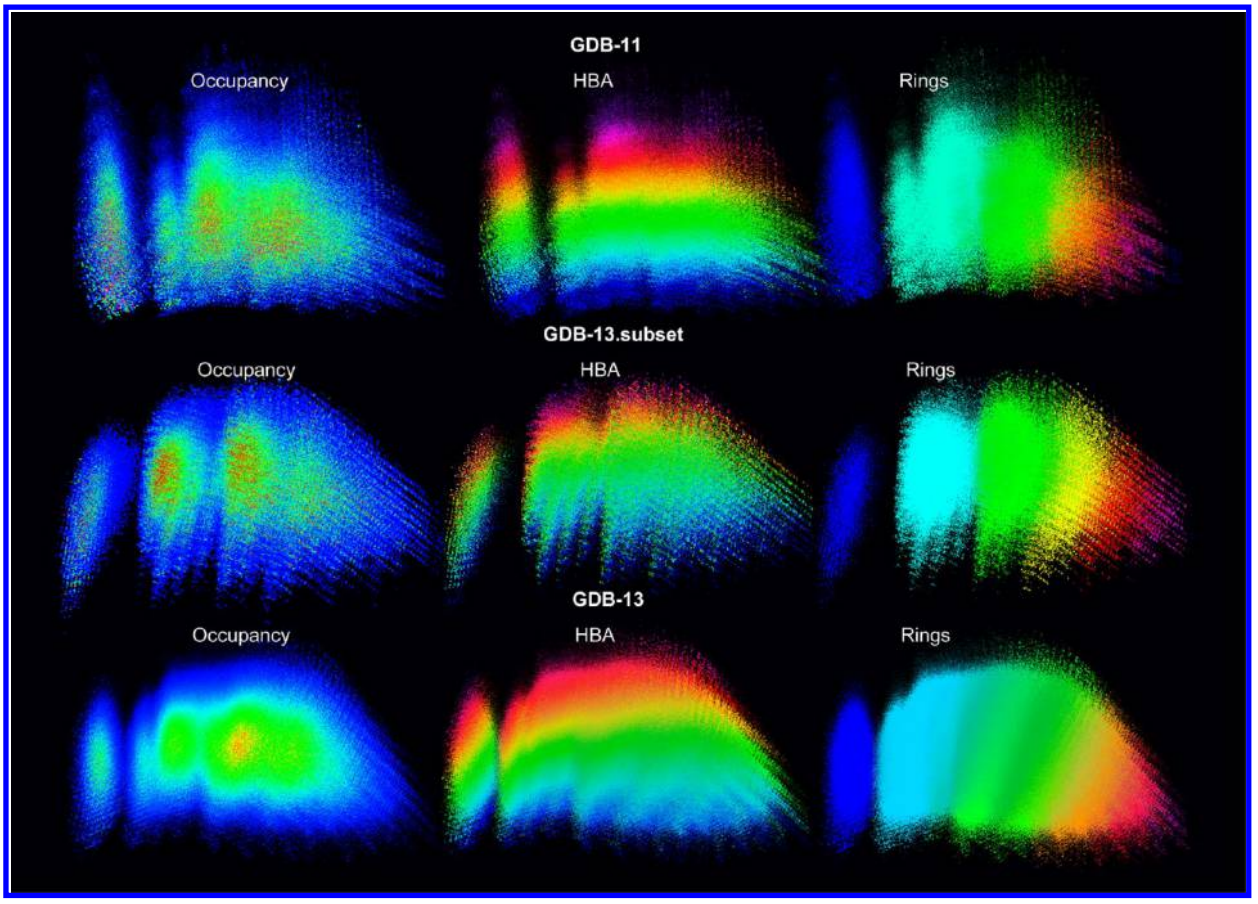


Figure 4. MQN-maps of GDB databases in the (PC1, PC2)-plane showing rigidity (PC1) and polarity (PC2). See Table 2 for database contents. Color-coding from blue (lower values) to purple (highest value).

by decreasing PC3-value in the case of (PC1,PC2)-maps, and the PC3 value is shown under each molecule. If the pixel contains more than 1000 molecules, a regular sampling of 1000 molecules along the PC3 value is shown. The entire content of each pixel can be downloaded as a list of SMILES using the "save bin" option and viewed separately if desired. This option is necessary for large molecules, in particular in PubChem.extended, because the viewer in the application is unable to display molecules with HAC > 60. An example of accessing the entire contents of a pixel is shown in Figure 5B for the ChEMBL.50 map. In the case of DrugBank, ChEMBL, and PubChem, one can further link from any molecule displayed in the full list to the entry in the original database Web site using the "look in database" when an active Internet connection is available (Figure 5D).

Locating a Molecule on MQN-Maps. Within each of the databases, the "locate molecule" function can be used to find the location of any molecule of choice on the MQN-map. The "locateMol" button in the main menu opens a molecule drawing window in which a molecule of choice can be either drawn or pasted as SMILES. Activating the "submit" function, which requires an active Internet connection, locates the pixel on the currently displayed map in which the molecule is located. This pixel appears blinking, and the full zoom mode is active. Note that the query molecule may or may not be present in the indicated pixel depending on whether the query is actually listed in the database. As an example, pixel (553, 609) in the ChEMBL.50 map was identified by copying the molecule from DrugBank in Figure 5A to the "locateMol" window, searching in ChEMBL.50 (Figure 5B). In this case the query macrocyclic lactone is not the average molecule shown for the pixel of the ChEMBL.50 map, but it can be found upon opening the pixel content as molecule no. 21 when the pixel contents are opened, together with another four diastereoisomers that are not present in DrugBank (Figure 5C/D).

Ligand Based Virtual Screening (LBVS). The MQN-maps are projections from the 42-dimensional MQN-space. The PC-planes cover 60−85% of the data variance, resulting in a good correlation between MQN-distances in the PC-plane and in the original 42-dimensional MQN-space (Figure 6A/B). Nevertheless, the projection into PC-planes removes many of the fine details of the MQN-classification, and reduces the efficiency of LBVS compared to nearest-neighbor searching in the original 42-dimensional MQN-space.[38] The MQN-mapplet therefore enables LBVS in the original 42-dimensional space by linking to the MQN-browser, a previously reported application.[36] The MQN-browser is accessible by clicking on any molecule of a displayed pixel content list, and choosing the "MQN search in database" button. As an example, one can search for MQN-nearest neighbors of the macrocyclic lactone selected from DrugBank in Figure 5A in the Pubchem.60 database, which reveals further analogs (Figure 5E/F).

Limitations of the MQN-Mapplet. The MQN-mapplet functions are constrained by the data-intensive nature of the application. To limit the overall size of the application for download (<100 MB), only seven different color-coded images are presented for each of the seven databases. Indeed each color-coded image is approximately 500 kilobytes, totaling 31

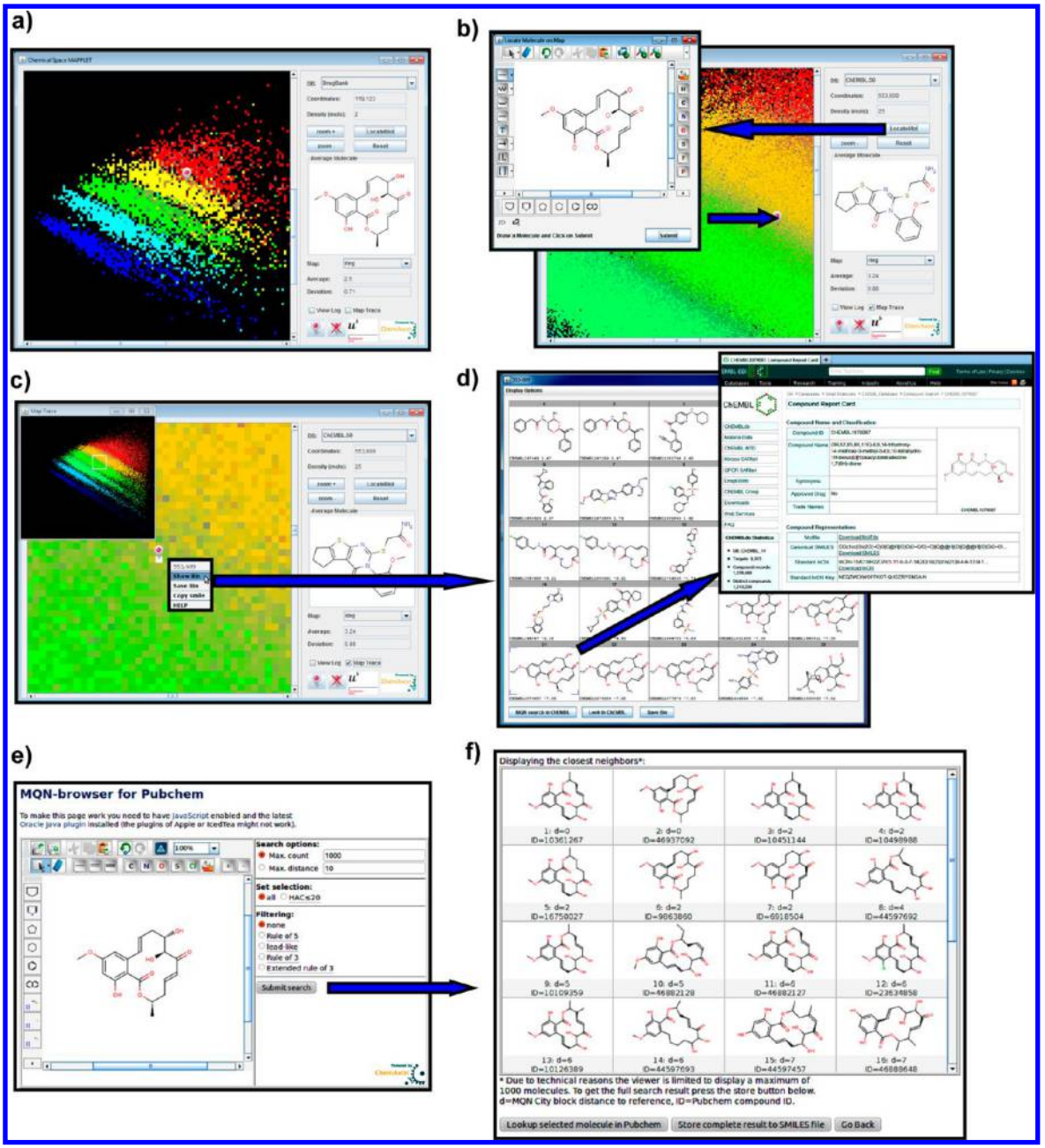


Figure 5. View of the MQN-mapplet as appearing on screen. (A) Main window as visible for DrugBank. (B) Molecule shown as example in DrugBank is copied into the "locateMol" window for ChEMBL.50 and the function locates the corresponding pixel. The molecule displayed in the "Average Molecule" window is the average molecule in ChEMBL.50 for that pixel. (C) Full zoom effect on the ChEMBL.50 map shown with the active "Map Trace" for the overview. (D) "mview" window accessed from the ChEMBL.50 pixel using the "Show Bin" function. The webpage of ChEMBL accessed for molecule no. 21 using the "look in ChEMBL" function is overlaid at right. (E) MQN-browser window for PubChem with the same example molecule no. 21 as query. (F) Results window displaying the MQN-nearest neighbors of the query macrolactone in PubChem.

MB for the 49 images available. For the same reason only one possible PCA view was selected for each database. The application furthermore only contains the structure of the average molecule in each pixel. An active Internet connection is required to access the rest of the databases located on the application server, which amount to a total of 16 GB of compressed archive files. The Internet connection is also necessary to activate the "locate molecule" function because this function uses the ChemAxon weblicense that is only available from the application server.

Despite of the analogy of the MQN-mapplet with Google maps, the resolution of the images does not increase upon zooming. While this choice is dictated by image data size, increasing the pixel resolution in full zoom mode is not meaningful. Indeed variations in other PCs are more relevant to classify the molecules within a pixel than a further spreading in the main PC-plane. Upon opening each pixel of the MQN-mapplet the molecules are displayed in the order of decreasing PC3 value, which appears as a very logical order (e.g., Figure 5D).

One further limitation of the MQN-mapplet concerns the MQN-system itself, which is a rather coarse classification method where molecules often share the same MQN-value combination and therefore occupy the same MQN-bin. With

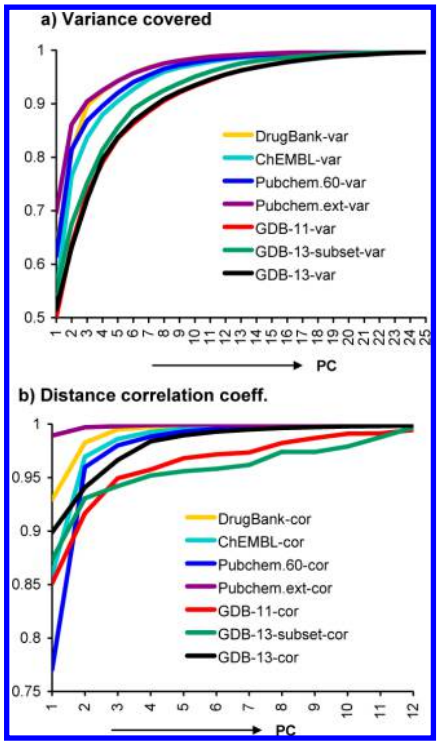


Figure 6. Variance and distance correlation in PCA analysis of MQN-space of the databases shown in the MQN-mapplet. (A) Percentage of MQN-data variance covered by increasing numbers of PC. (B) Pearson's correlation coefficient between pairwise euclidean distances in the *n*-dimensional PC-subspace and the original 42-dimensional MQN-space, calculated from analyzing 10 000 randomly selected molecule pairs in each database.

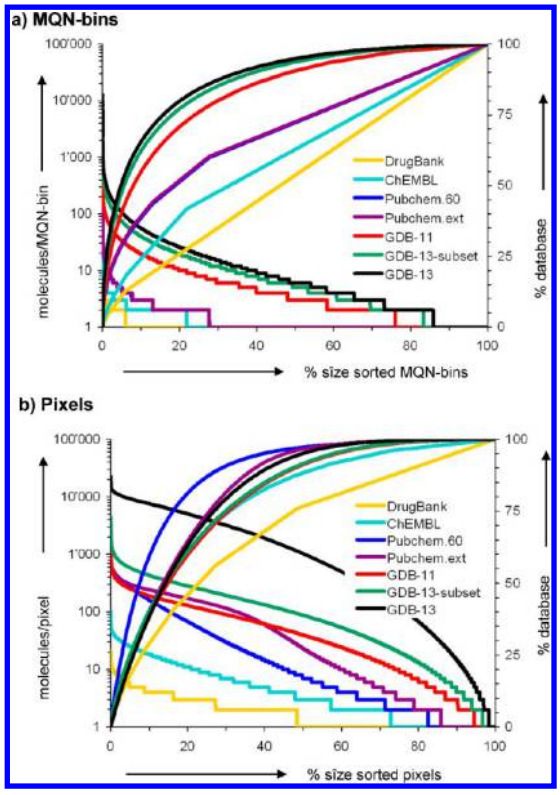


Figure 7. Distribution of molecules in (A) the MQN-value combination (MQN-bins) of the 42-dimensional MQN-space and (B) the pixels of the MQN-maps.

the exception of the very small database DrugBank, this results in a power-law distribution of molecules in MQN-bins (Figure 7A), which induces a similar power-law distribution of molecules in the pixels of the MQN-maps, such that 50% of each database is typically contained in only 10−20% of the MQN-map pixels (Figure 7B). It might be possible to obtain maps with a more even distribution of molecules using a finer, higher-dimensionality classification system such as a substructure fingerprint, and producing 2-dimensional representations by applying nonlinear projection methods tailored for the visualization of multidimensional spaces as 2D-maps.[45−50] However these methods would require to compute the complete distance matrix between all molecules, which is not possible for the databases of millions of molecules presented here with currently available computational resources.

CONCLUSION

MQN-mapplet represents a rapid visualization tool for very large databases (for technical reasons a corresponding 3D-map viewers appears at present far too data-intensive for interactive viewing). In contrast to other database search tools, MQN-mapplet does not require the definition of a query molecule to start the exploration, which may be particularly attractive for nonchemists. To our knowledge, this type of interactive exploration tool is unprecedented for very large databases such as PubChem and GDB-13 (almost one billion molecules). It can be used as an entry portal to search for interesting molecules. While the current mapplet shows a PCA projection from MQN-space, images suitable for browsing with the mapplet might be generated from other property spaces such as those described earlier by Pearlman, Oprea, Medina-Franco, and others,[21−23,25−27] which could offer alternative entries into large databases.

METHODS

Database Selection and Determination of MQN Values. DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13 databases were downloaded in June 2012 from their respective Web sites (Table 2). ChEMBL.50 and Pubchem.60 are subset containing all molecules with up to 50 (ChEMBL.50) and 60 (Pubchem.60) heavy atoms, respectively. A subset of GDB-13 was created by excluding molecules with (a) nonaromatic cyclic NN and NO bonds, (b) acyclic NN and NO bonds, mostly from oximes and hydrazones (c) aldehydes, esters, carbonates, sulfates, epoxides, aziridines, (d) nonaromatic CC double and triple bonds inside cycles, (e) acyclic CC double and triple bonds, and (f) three- and four-membered rings (see ref 40 for details). Pubchem.extended database combines PubChem with six different virtual categories of molecules: peptides, DNA, graphenes, diamondoids, acyclic alkanes, and oligosaccharides (details available in ref 38), up to a maximum heavy atom count HAC = 500. Molecules were processed into SMILES removing counterions, and any duplicates were removed. Molecular quantum numbers (MQNs) were calculated using an in-house developed Java program which is utilizing Java Chemistry library (JChem) from Chemaxon, Ltd. as a starting point.[32] During MQN calculation, the ionization state of each molecule was adjusted to pH 7.4.

Principal Component Analysis. PCA of the 42-dimensional MQN space of all databases was carried out using an in-house developed Java program,[38] utilizing some of the available

mathematical functions from JSci (A science API for Java: http://jsci.sourceforge.net/) library. The Java source code is based on the tutorial of Lindsay I. Smith (http://www.cs.otago.ac.nz/cosc453/student_tutorials/princ ipal_components.pdf).

MQN-Map Generation and Color-Coding. Each molecule was assigned to its plane coordinates (PC1,PC2). The largest (PC_{max}) and smallest (PC_{min}) PC values appearing in the PC1 or PC2 values were used to define the value range $\Delta PC = PC_{max} - PC_{min}$ and set the binning scale as $\Delta PC/1000$. The (PC1,PC2)-plane was binned in a 1000 × 1000 grid using the same absolute bin size on the PC1 and PC2 axis. Each molecule was assigned to a bin on the PC-plane, each defining a pixel of a 1000 × 1000 pixel MQN-map. The (PC2,PC3)-plane was used for PubChem.extended, and the binning was reduced to 300 × 300 pixels for DrugBank.

MQN-maps were color coded in the HSL (hue, saturation, luminance) color space as described previously.[40] H (color) codes for the average property value in a pixel, from blue (lowest value) to magenta (highest value) via cyan−green−yellow−orange−red (intermediate values) in a continuous manner. S (gradual color fading to gray) codes for the standard deviation of the property value in the pixel, and L (brightness, fading to black) codes for the pixel occupancy (used only for the GDB maps).

Nonlinear Distortion Methods for PubChem.60 and PubChem.extended. MQN-maps of PubChem.60 and PubChem.extended were distorted to spread the area with the highest density of molecules per pixel and concentrated the sparsely populated outer regions of the MQN-maps as follows: The pixel of highest density on the MQN-map was set as (ΔPC1,ΔPC2) = (0,0). For each molecule (ΔPC1,ΔPC2) coordinates were calculated relative to this point of highest density, and new coordinates (ΔPC1′,ΔPC2′) were computed as follows:

$$\Delta PC = \mathrm{sqrt}(\Delta PC1^2 + \Delta PC2^2)$$

$$\Delta PC1' = \Delta PC1(\mathrm{sqrt}(\Delta PC + 1) - 1)/\Delta PC$$

$$\Delta PC2' = \Delta PC2(\mathrm{sqrt}(\Delta PC + 1) - 1)/\Delta PC$$

Molecules in the corrected PC′-plane were binned again into the 1000 × 1000 pixel grid as above. The procedure was repeated until a satisfactory density distribution was obtained. For comparison, noncorrected representations of the (PC2,PC3)-plane of the PubChem.extended MQN-map are shown in ref 13 and the corresponding issue cover page.

MQN-Mapplet. MQN-Mapplet is a desktop application, partially utilizing a web interface for the interactive visualization of chemical space. The application is written in Java programming language and utilizes the JChem library from Chemaxon Ltd. (http://www.chemaxon.com/). Some functionalities require an active Internet connection, e.g. for visualizing the content of a pixel and for locating a molecule on the map. MQN-Mapplet can be downloaded free of charge from www.gdb.unibe.ch, which also provide information on how to set up and use the application along with some of the Internet security issues (for e.g. virtual private network (VPN) wall) which one needs to consider.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

REFERENCES

(1) 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.

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

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

(4) 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.

(5) Thomas, G. L.; Wyatt, E. E.; Spring, D. R. Enriching chemical space with diversity-oriented synthesis. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 700−712.

(6) Williams, A. J. Public chemical compound databases. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 393−404.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) 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.

(12) 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.

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

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

(15) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(16) Ivanenkov, Y. A.; Savchuk, N. P.; Ekins, S.; Balakin, K. V. Computational mapping tools for drug discovery. *Drug Discovery Today* **2009**, *14*, 767−775.

(17) 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.

(18) Ertl, P.; Schuffenhauer, A.; Renner, S. The scaffold tree: an efficient navigation in the scaffold universe. *Methods Mol. Biol.* **2011**, *672*, 245−260.

(19) Ertl, P.; Rohde, B. The Molecule Cloud - compact visualization of large collections of molecules. *J. Cheminf.* **2012**, DOI: 10.1186/1758-2946-4-12.
(20) 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.
(21) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Persp. Drug Discovery Des.* **1998**, *9*−*11*, 339−353.
(22) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.
(23) 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.
(24) 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.
(25) 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.
(26) 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.
(27) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325−330.
(28) Le Guilloux, V.; Colliandre, L.; Bourg, S. p.; Guénegou, G.; Dubois-Chevalier, J.; Morin-Allory, L. Visual Characterization and Diversity Quantification of Chemical Libraries: 1. Creation of Delimited Reference Chemical Subspaces. *J. Chem. Inf. Model.* **2011**, *51*, 1762−1774.
(29) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−594.
(30) Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **2009**, *20*, 429−436.
(31) Mason, J. S.; Beno, B. R. Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. *J. Mol. Graphics Modell.* **2000**, *18*, 438-451, 538.
(32) 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.
(33) Takahashi, Y.; Konji, M.; Fujishima, S. MolSpace: a computer desktop tool for visualization of massive molecular data. *J. Mol. Graphics Modell.* **2003**, *21*, 333−339.
(34) Gütlein, M.; Karwath, A.; Kramer, S. J. CheS-Mapper - Chemical Space Mapping and Visualization in 3D. *J. Cheminf.* **2012**, DOI: 10.1186/1758-2946-4-7.
(35) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.
(36) 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.
(37) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.
(38) van Deursen, R.; Blum, L. C.; Reymond, J. L. A searchable map of PubChem. *J. Chem. Inf. Model.* **2010**, *50*, 1924−1934.
(39) 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.
(40) 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.
(41) Awale, M.; Reymond, J. L. Cluster analysis of the DrugBank chemical space using molecular quantum numbers. *Bioorg. Med. Chem.* **2012**, *20*, 5372−5378.
(42) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(43) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53*, 3862−3886.
(44) Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. Discovery of alpha7-Nicotinic Receptor Ligands by Virtual Screening of the Chemical Universe Database GDB-13. *J. Chem. Inf. Model.* **2011**, *51*, 3105−3112.
(45) Agrafiotis, D. K.; Lobanov, V. S. Nonlinear Mapping Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1356−1362.
(46) Xie, D.; Tropsha, A.; Schlick, T. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-newton minimization. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 167−177.
(47) Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. Data Visualization during the Early Stages of Drug Discovery. *J. Chem. Inf. Model.* **2006**, *46*, 1806−1818.
(48) von Korff, M.; Hilpert, K. Assessing the Predictive Power of Unsupervised Visualization Techniques to Improve the Identification of GPCR-Focused Compound Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 1580−1587.
(49) Macchiarulo, A.; Thornton, J. M.; Nobeli, I. Mapping Human Metabolic Pathways in the Small Molecule Chemical Space. *J. Chem. Inf. Model.* **2009**, *49*, 2272−2289.
(50) Owen, J. R.; Nabney, I. T.; Medina-Franco, J. L.; López-Vallejo, F. Visualization of Molecular Fingerprints. *J. Chem. Inf. Model.* **2011**, *51*, 1552−1563.