Journal of
CHEMICAL INFORMATION
AND MODELING

Article

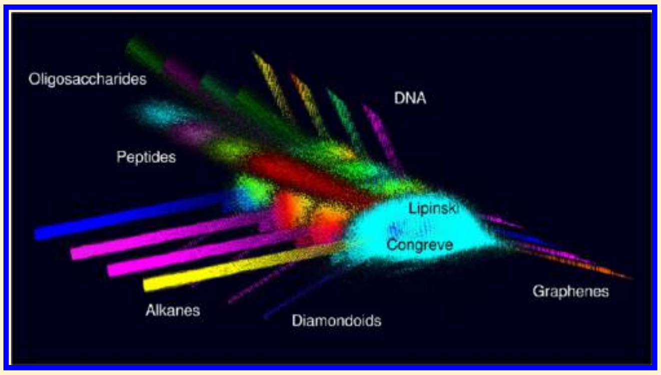
pubs.acs.org/jcim

SMIfp (SMILES fingerprint) Chemical Space for Virtual Screening and Visualization of Large Databases of Organic Molecules

Julian Schwartz, Mahendra Awale, and Jean-Louis Reymond*

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

Supporting Information



ABSTRACT: SMIfp (SMILES fingerprint) is defined here as a scalar fingerprint describing organic molecules by counting the occurrences of 34 different symbols in their SMILES strings, which creates a 34-dimensional chemical space. Ligand-based virtual screening using the city-block distance CBD_{SMIfp} as similarity measure provides good AUC values and enrichment factors for recovering series of actives from the directory of useful decoys (DUD-E) and from ZINC. DrugBank, ChEMBL, ZINC, PubChem, GDB-11, GDB-13, and GDB-17 can be searched by CBD_{SMIfp} using an online SMIfp-browser at www.gdb.unibe.ch. Visualization of the SMIfp chemical space was performed by principal component analysis and color-coded maps of the (PC1, PC2)-planes, with interactive access to the molecules enabled by the Java application SMIfp-MAPPLET available from www.gdb.unibe.ch. These maps spread molecules according to their fraction of aromatic atoms, size and polarity. SMIfp provides a new and relevant entry to explore the small molecule chemical space.

INTRODUCTION

The chemical space of organic small molecules is the ensemble of all known and theoretically possible molecules up to a certain size (usually MW < 500 Da).[1,2] This vast resource presents many opportunities for drug development because it holds one of the keys to discover new chemical entities that might lead to useful new drugs for a variety of diseases.[3] The efficient exploitation of chemical space for drug discovery critically depends on virtual screening (VS) and visualization tools to support the identification and comparison of molecules with interesting activity profiles.[4−10] One approach to this problem pioneered by Pearlman and Smith consists of organizing molecules in a property space whose dimensions correspond to numerical descriptors of the molecules and their properties.[11] The descriptor values associated with any molecule define its position in the property space in form of a vector or fingerprint.[12] Such property spaces may be used for ligand-based virtual screening (LBVS), which consists in scoring molecules by similarity to a reference molecule using a distance or similarity measure.[13] Property spaces can furthermore be visualized by principal component analysis (PCA) and representation of selected PC-planes, an approach used by various groups to compare compound collections, such as drugs, screening compounds, and natural products.[14−20]

Property spaces are very appealing because they propose a geometrical realization of the concept of "chemical space" by positioning molecules in a multidimensional euclidean space. How useful and inspiring a particular property space can be depends not only on the chemical or biological significance of the selected properties, but also on the availability of efficient tools for rapid similarity searching and visualization. Recently, we reported a property space based on 42 integer value descriptors of molecular structure counting atoms and bonds by their types, as well as polar groups and topological features, called "molecular quantum numbers" (MQN).[21,22] Efficient similarity searching in MQN-space was made possible by a web-based "MQN-browser" application capable of identifying biologically relevant drug analogs in databases up to billions of molecules within seconds.[23−25] MQN-space was also adapted to interactive visualization by PCA and representation of PC-planes via an interactive "google-maps"-like application.[22,26]

We asked the question whether a new property space related to MQN-space and the corresponding browser and visualization tools might be derived by reading information from the SMILES (Simplified Molecular Input Line Entry System),[27,28] which is the most commonly used and compact representation of molecules. SMILES have been used previously to construct so-called "lingos", which are binary fingerprints analyzing the occurrence of the most frequent strings of three characters within SMILES.[29,30] Lingos provide valuable information on the molecules and allow construction of predictive models for solubility and logP within defined molecular series, and are also useful for virtual screening.[31] Similar SMILES fragments have been used to construct QSAR models.[32] Considering that the efficiency of LBVS in MQN-space can be partly explained by the

Received: April 5, 2013
Published: July 11, 2013

Table 1. Thirty-Four Symbols Counted in the SMIfp

no.	symbol	feature counted
1	C	nonaromatic carbon atoms
2	c	aromatic carbon atoms
3	N	nonaromatic nitrogen atoms
4	n	aromatic nitrogen atoms
5	O	nonaromatic oxygen atoms
6	o	aromatic oxygen atoms
7	S	nonaromatic sulfur atoms
8	s	aromatic sulfur atoms
9	F	fluorine atoms
10	Cl	chlorine atoms
11	Br	bromine atoms
12	I	iodine atoms
13	P	nonaromatic phosphorus atoms
14	p	aromatic phosphorus atoms
15	B	boron atoms
16	"X"	any other element[a,b]
17	—	explicit single bonds[c]
18	=	double bonds
19	#	triple bonds
20	[	special features[d]
21	-	negative charges[b]
22	+	positive charges[b]
23	H	explicit hydrogen atoms[b,e]
24	(	acyclic branching points
25	1	nonfused ring systems
26	2	bicyclic systems[f]
27	3	tricyclic systems[f]
28	4	tetracyclic systems[f]
29	5	pentacyclic systems[f]
30	6	hexacyclic systems[f]
31	7	heptacyclic systems[f]
32	8	octacyclic systems[f]
33	9	nonacyclic systems[f]
34	%	higher order ring systems[f]

[a]Includes 2-letter combinations. [b]Always within [ ]. [c]Always outside [ ]. [d]Nonorganic elements, charges, isotopes, protonation states. [e]On charged atoms. [f]Counts fused ring systems only. The following symbols are not considered: /, \,:, ., @.

fact that molecules with similar biological activity often have comparable global features such as size, rigidity and polarity, as previously noted by Bender and Glen in LBVS with various simple fingerprints,[33,34] we aimed for a scalar fingerprint simply counting the occurrences of the symbols used in SMILES.

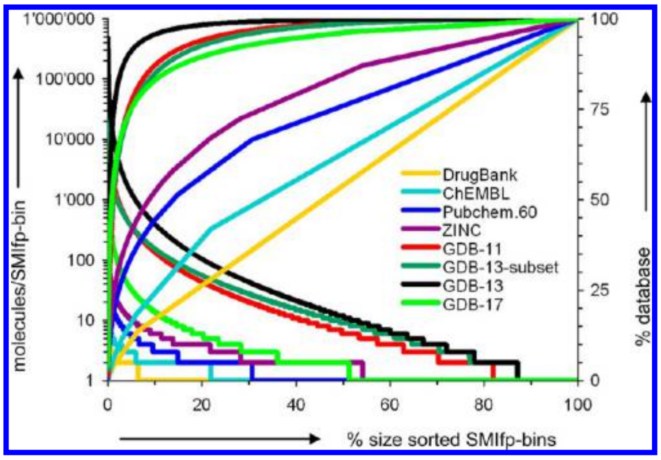


Figure 2. Distribution of molecules in the SMIfp value combinations (SMIfp bins).

Table 2. Databases of the Known and Unknown Chemical Space Analyzed by SMIfp

database[a]	no. of cmpds[b]	no. of SMIfp[c]	SMIfp occupancy av/max[d]
DrugBank	6400	5899	1.1/8
ChEMBL.50	1 094 469	813 329	1.3/56
ZINC.50	19 614 276	5 472 417	3.6/1973
PubChem.60	24 498 884	11 702 070	2.1/1010
GDB-11	26 410 137	236 057	112/39 122
GDB-13	975 821 530	1 199 023	814/1 229 042
GDB-13 subset	43 474 445	352 472	123/104 552
GDB-17	49 922 564	3 615 041	13.9/22 736

[a]The latest version of each database was downloaded in February 2013. ChEMBL.50 contains all molecules up to 50 heavy atoms (96% of ChEMBL). PubChem.60 contains all molecules up to 60 heavy atoms. The GDB-13 subset contains molecules from GDB-13 without 3- or 4-membered rings, nonaromatic C=C, N–O, or N–N bonds, and unstable functional groups (esters, aldehydes). The analysis of GDB-17 was limited to a 50 million subset. [b]Number of compounds with unique ID codes, considering all entries successfully converted to a SMIfp. Counter ions were removed to compute the SMIfp. [c]Number of occupied SMIfp bins. [d]Average and maximum number of compounds per SMIfp bin.

Herein, we describe a scalar fingerprint called SMIfp (SMILES fingerprint) describing organic molecules by counting the occurrence of 34 symbols found in SMILES, which defines a 34-dimensional property space (Table 1). LBVS by proximity in this SMIfp space using the city-block distance CBD_{SMIfp} as similarity measure gives good enrichment factors for recovery of

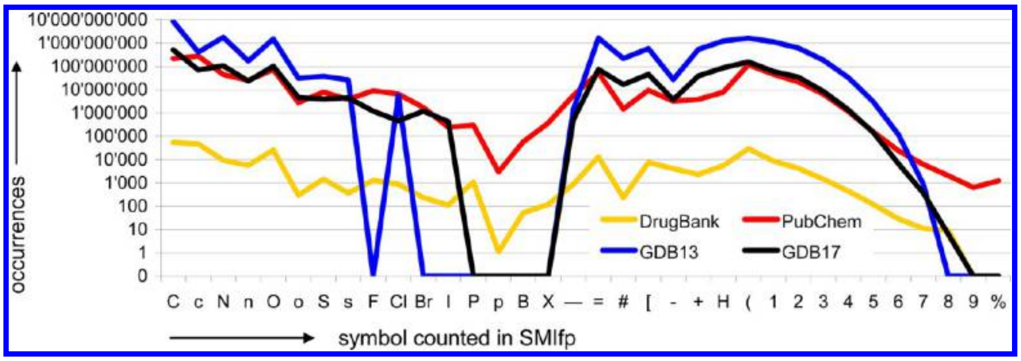
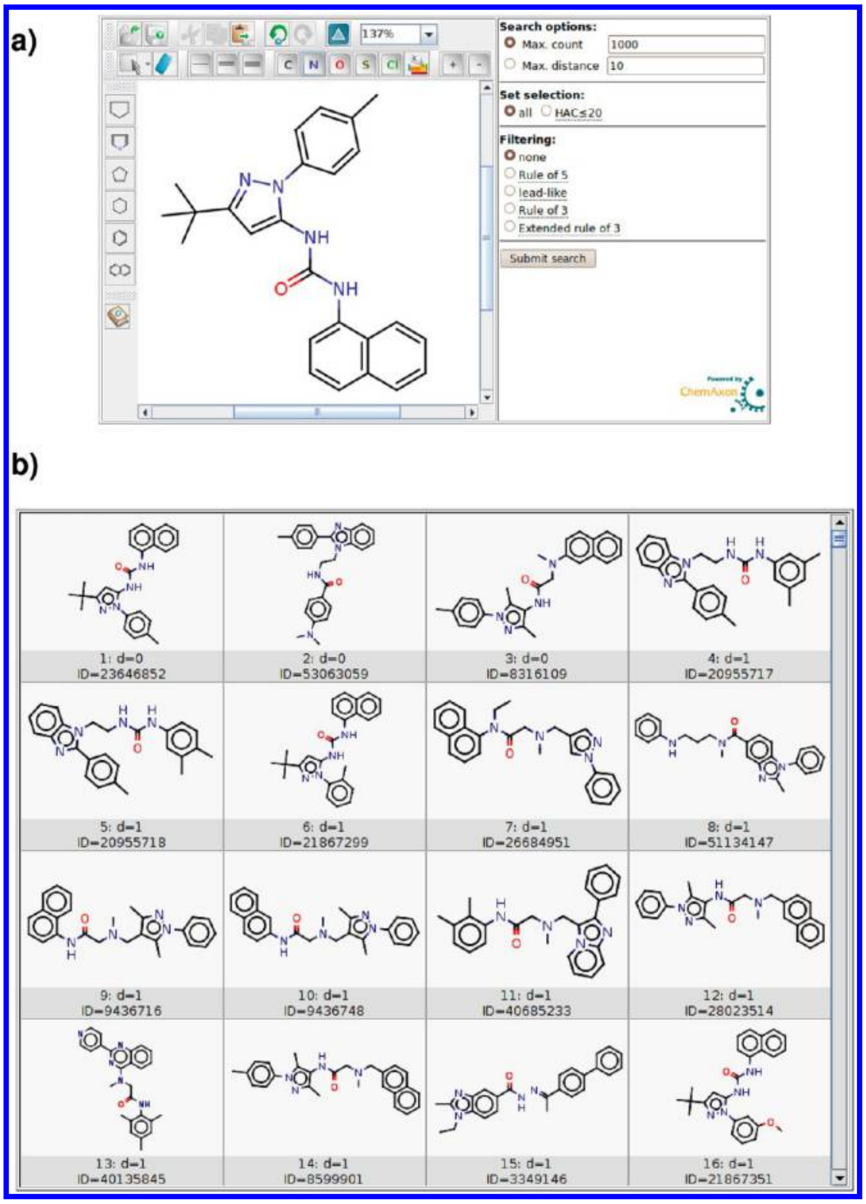


Figure 1. Number of occurrences of symbols in SMILES in various databases.



Search options:
● Max. count 1000
○ Max. distance 10

Set selection:
● all ○ HAC≤20

Filtering:
● none
○ Rule of 5
○ lead-like
○ Rule of 3
○ Extended rule of 3

Submit search

Figure 3. Example of a SMIfp-similarity search in PubChem using the SMIfp-browser. The search-time for retrieving the first 1000 $\mathrm{CBD}_{\mathrm{SMIfp}}$-neighbors was 1.3 s.

actives from the enhanced directory of useful decoys (DUD-E) and from ZINC.[35,36] Similarity searching can be performed in seconds using the SMIfp browser available at www.gdb.unibe.ch. The SMIfp space is furthermore suitable for database visualization via PCA analysis and color-coded representation of the (PC1,PC2)-plane. An interactive access to the molecules is made possible by the Java application SMIfp-MAPPLET available from www.gdb.unibe.ch. These maps spread molecules according to their size, fraction of aromatic atoms, and polarity. SMIfp provides a new and relevant entry into the small molecule chemical space.

■ RESULTS AND DISCUSSION

SMIfp. To select relevant symbols in SMILES for creating a scalar fingerprint, the number of occurrences of each symbol was counted in the public database PubChem, a public repository of over 32 million substances.[37] In total 34 symbols were selected because they occurred at least 500 times in the SMILES list of PubChem (Figure 1). Sixteen of these symbols represented elements, distinguishing aromatic and nonaromatic C, N, O, S, and P, and including the two-letters symbol elements chlorine (Cl) and bromine (Br). Nonstandard elements were counted collectively with the symbol X. Three symbols corresponded to bonds, four symbols to polar groups, and eleven symbols to topological features (Table 1). Symbols for explicit aromatic bonds (:), fragment separation (.), and stereochemistry (/, \, @) were not considered because they were too rare or not used systematically.

The SMIfp was applied to analyze the public databases of known molecules DrugBank (>6,000 experimental or approved drugs),[38] ChEMBL (>1.1 million compounds with documented bioactivity),[39] PubChem (>47 million reported molecules),[37] and ZINC (>20 million commercially available small molecules), as well as the databases of virtual molecules GDB-11 (26.4 million molecules up to 11 atoms of C, N, O, F),[40,41] GDB-13 (977 million molecules up to 13 atoms of C, N, O, S, Cl),[42] and a 50 million molecule subset of GDB-17 (166.4 billion molecules up to 17 atoms of C, N, O, halogens, S).[43,44] In each case, the

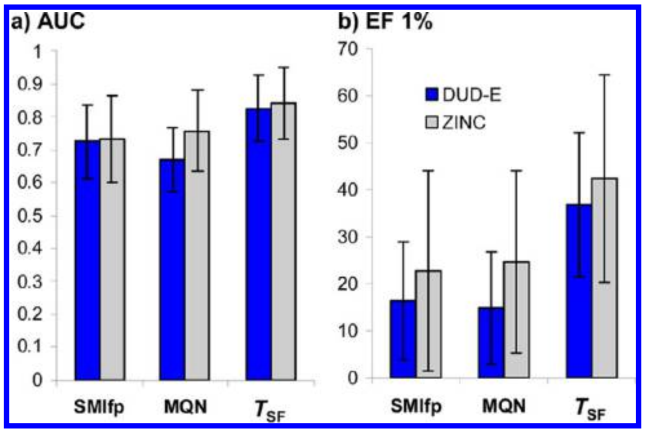


Figure 4. Average performance for recovering actives in 101 sets of the DUD-E from the corresponding decoys (blue bars) or from the entire ZINC database (gray bars) using the scoring functions CBD_{SMIfp}, CBD_{MQN}, or T_{SF}. The molecule most similar to all others in each set with respect to each fingerprint was used as reference. See also full data in Supporting Information Tables S1 and S2 and ROC curves in Supporting Information Figure S1.

distribution of molecules into SMIfp value combinations, called SMIfp-bins, followed a typical power law distribution similar to that obtained when classifying molecules with MQN (Figure 2, Table 2).[26] The most highly occupied SMIfp-bins contained up to thousands of molecules, reflecting the fact that counting symbols in SMILES does not encode the molecular structure precisely. On the other hand approximately 50% of the SMIfp-bins were occupied by only a single molecule in databases of known molecules, such as DrugBank, ChEMBL, ZINC, and PubChem. This proportion decreased to 20% in the enumerated databases GDB-11, GDB-13, and GDB-17, reflecting the exhaustive combinatorial enumeration, which generates many similar molecules.

SMIfp Browser. The SMIfp annotated databases were organized for similarity searching by CBD_{SMIfp} via a hash-table using the sum of all SMIfp values as hash function, in a manner similar to that previously described for MQN.[25,44,45] Similarity searching was enabled via a web-browser application, which allows to retrieve CBD_{SMIfp}-nearest neighbors of any molecule of choice within seconds from any of the above-mentioned databases. Visual inspection of a typical set of such nearest neighbors illustrates the structural analogies between CBD_{SMIfp}-nearest neighbors, which usually have similar overall composition in terms of rings, linkers, atom types and functional groups, but overall different relative arrangements of the rings and connecting points (Figure 3).

Ligand-Based Virtual Screening (LBVS). To test the relevance of the SMIfp classification with regards to biological activities, recovery of the active series in the "directory of useful decoys—enhanced" (DUD-E)[36] from the list of decoys and from the entire ZINC database was performed using the city-block distance CBD_{SMIfp} as similarity measure. The performance of the SMIfp-based recovery was compared to the results obtained using similarities measured by MQN using the city-block distance CBD_{MQN}, and by a 1024 bit Daylight type substructure fingerprint using the Tanimoto similarity coefficient T_{SF}.[12,13] The ligand most similar to all other ligands within each fingerprint was used as the reference ligand for each series and the ROC (receiver operator characteristic) curves were calculated (Supporting Information Figure S1). The data was analyzed by AUC (area under the curve) and EF (enrichment factors) values (Supporting Information Table S1, Figure 4). SMIfp similarity performed comparably to the MQN similarity for recovery of actives from ZINC, but was also surprisingly efficient for separating active from decoys, although it could not match SF similarity because the decoys are selected for low T_{SF} similarity to the actives. Note that the results with SMIfp and MQN were very similar when using the Tanimoto coefficient rather than the city-block distance as similarity measure.

As observed previously for MQN similarity searching,[22] SMIfp similarity allowed scaffold hopping[46] by assigning high SMIfp similarity to molecules with quite different substructures as judged by T_{SF}. Similarities by SMIfp and by MQN were however substantially different from one another, showing that these fingerprints perceive different molecular features (Figure 5). Examples of actives and their analogs from the DUD-E series recovered from decoys at different CBD_{SMIfp} distances from the reference within each set of actives are shown in Figure 6.

SMIfp-Mapplet. To gain an insight into how molecules are organized in the SMIfp chemical space, several databases were analyzed by PCA of the SMIfp data. The loading plots for the first three principal components (PC), which account for 71−92% of

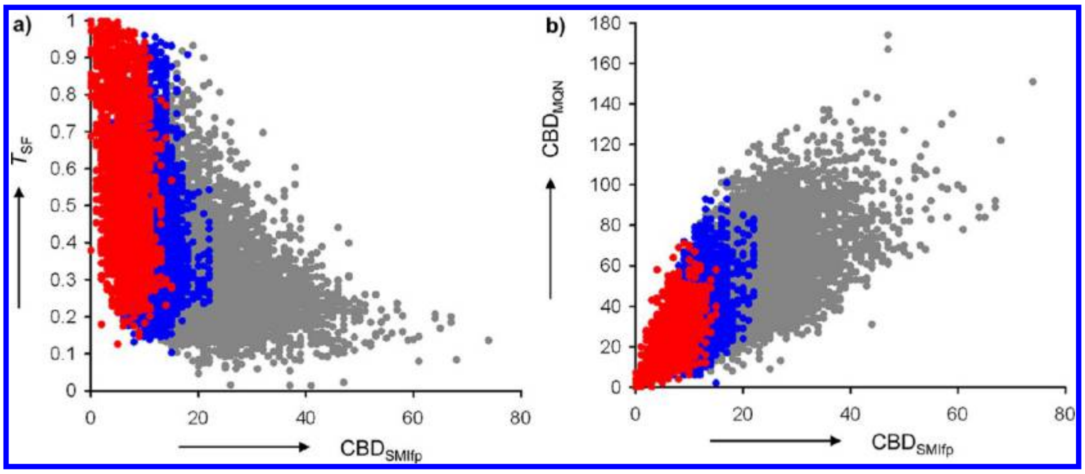


Figure 5. Scatter plots of (a) T_{SF} and (b) CBD_{MQN} similarities of actives within DUD-E as a function of CBD_{SMIfp} similarity. Similarities are calculated relative to the reference molecule within each set of actives. Points are color-coded for recovery by SMIfp similarity from the respective set of decoys within the top 1% (red), 10% (blue), or remaining 90% (gray).

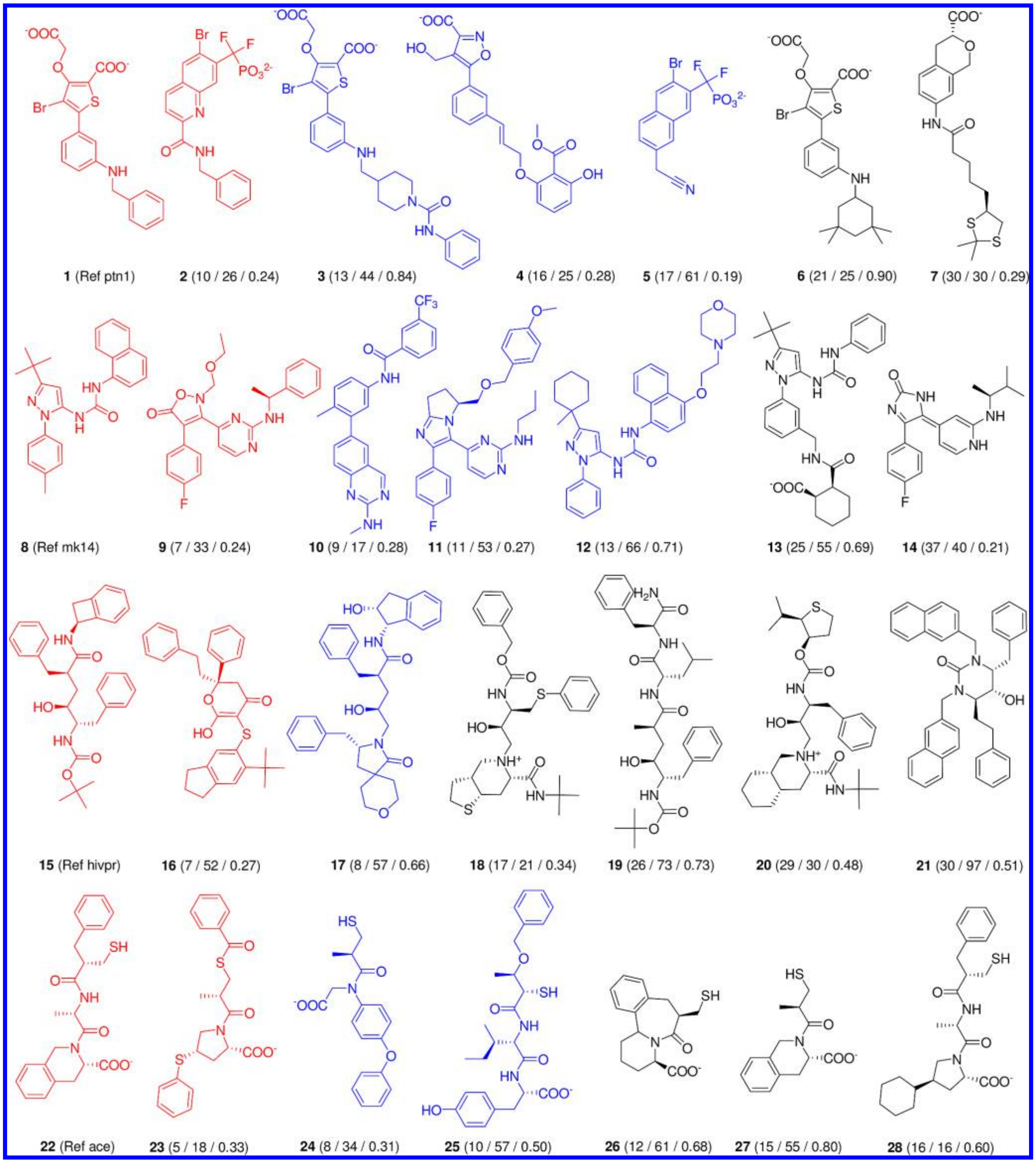


Figure 6. Examples of actives and their $\text{CBD}_{\text{SMIfp}}$-distance relative to the reference, which is the ligand closest to all other ligands (by $\text{CBD}_{\text{SMIfp}}$) in the respective set of actives. **1**: Inhibitor of ptn1 (protein-tyrosine phosphatase 1B). **8**: Inhibitor of mk14 (MAP kinase p38 alpha). **15**: Inhibitor of hivpr (human immunodeficiency virus type 1 protease). **22**: Inhibitors of ace (angiotensin-converting enzyme). The actives found are listed in order of increasing $\text{CBD}_{\text{SMIfp}}$ and color coded according to recovery from decoys by $\text{CBD}_{\text{SMIfp}}$ to the reference within the top 1% (red), top 10% (blue), or remaining 90% (black) The number in parentheses are the distances to the reference as $\text{CBD}_{\text{SMIfp}}/\text{CBD}_{\text{MQN}}/T_{\text{SF}}$.

data variability, are shown in Figure 7 and 8. The data variability coverage by the first three PCs was comparable to that obtained previously in PCA of the MQN data, however the structural features separating the molecules in each PC were quite different. The first principal component PC1, which covered 48−61% data variability in databases of known molecules and 34% - 57% of data variability in the GDB databases, separated molecules according to aromaticity. This separation reflected the distinction made between upper case letters (nonaromatic atoms) and lower case letters (aromatic atoms) in the SMIfp analysis, a feature which is completely absent in the MQN classification. For the databases of known molecules, PC2 covered 17−33% of data

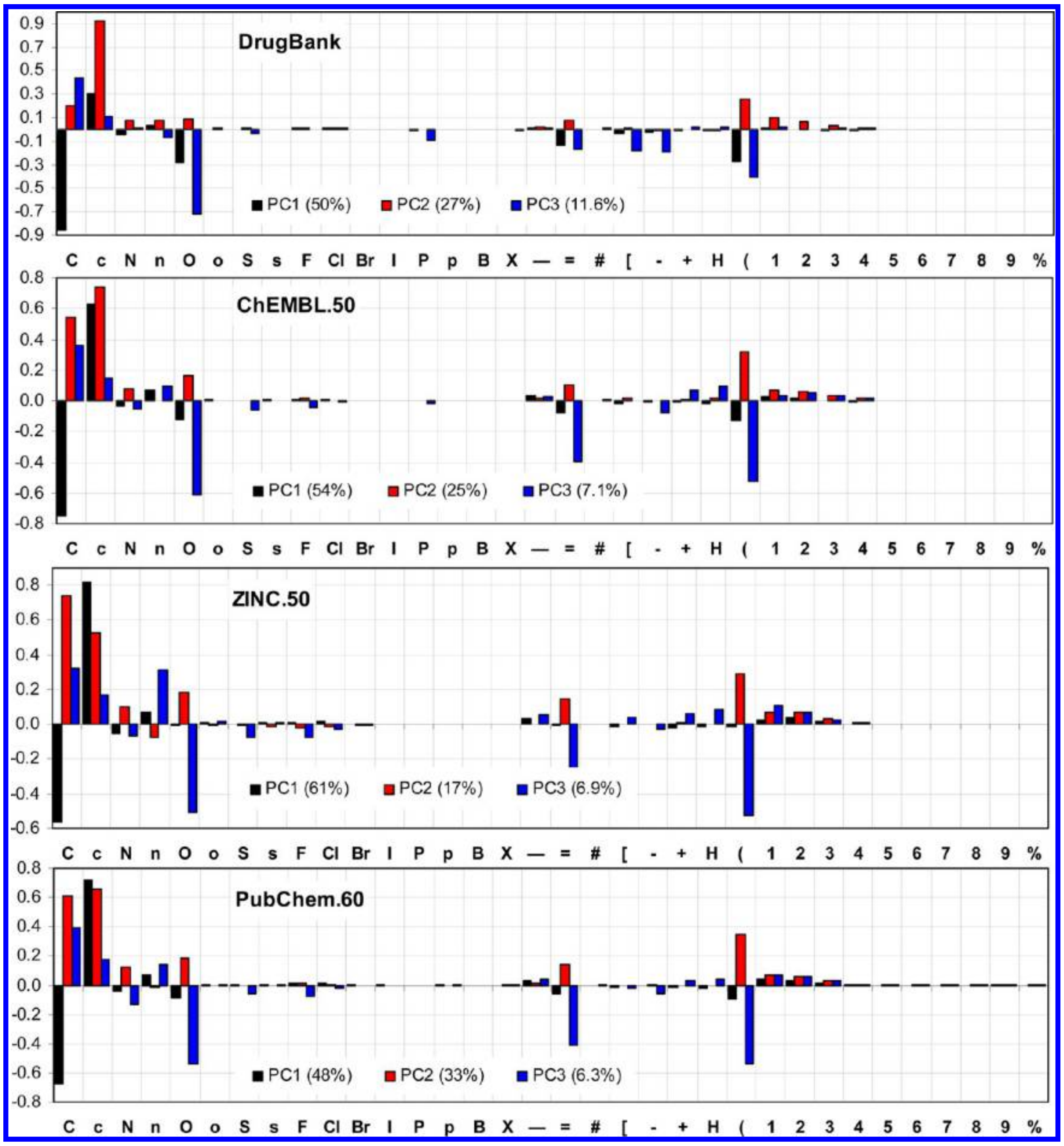


Figure 7. Loading plots for PC1, PC2, and PC3 of SMIfp analysis for DrugBank, ChEMBL.50, ZINC.50, and PubChem.60.

variability and reflected molecular size, which reproduces the separation obtained in PC1 for MQN analyses. In the case of the GDB databases, where molecular size only plays a minor role, PC2 covered 18−25% data variability and separated the structures according to explicit hydrogen count and positive charge, as well as oxygen and nitrogen atom count, reflecting the fraction of H-bond donor atoms.

To produce visually interpretable maps of the SMIfp chemical space, the (PC1,PC2)-plane was represented as color-coded map according to various descriptor values. An interactive access to the molecules within these maps was enabled in a SMIfp-MAPPLET application. This Java application is available for download as a 130 Mb .jar archive from www.gdb.unibe.ch and can be used to visualize the molecules in each pixel of the color-coded maps. The application furthermore provides links to the SMIfp browser and, if available, to the original database Web site. The graphical user interface and functions are identical to those of the related MQN-mapplet, which we have recently described in detail.[26]

For the databases of known molecules (DrugBank, ChEMBL, ZINC and PubChem), the (PC1,PC2)-maps formed a downward pointing triangle composed of parallel diagonal stripes stretching from upper left to lower right (Figure 9). These stripes contained molecules with increasing numbers of aromatic carbons (Figure 9A). Molecules with an increasing number of nonaromatic carbons spread from lower right to upper left within each stripe (Figure 9B). The diagonal stripes with increasing numbers of aromatic carbons also distributed molecules with increasing number of cycles, which is logical considering that acyclic molecules cannot have aromatic carbons (Figure 9C). Overall the molecules were distributed by size along the vertical axis, as measured for example by the heavy atom count (Figure 9D).

The SMIfp (PC1,PC2) maps of the GDB-databases appeared as two groups sharply separated vertical stripes (Figure 10). The left group of stripes contained all nonaromatic molecules, and the right group of stripes all aromatic molecules (Figure 10A). The individual stripes contained molecules with an increasing number of nonaromatic carbon atoms from right to left (Figure 10B).

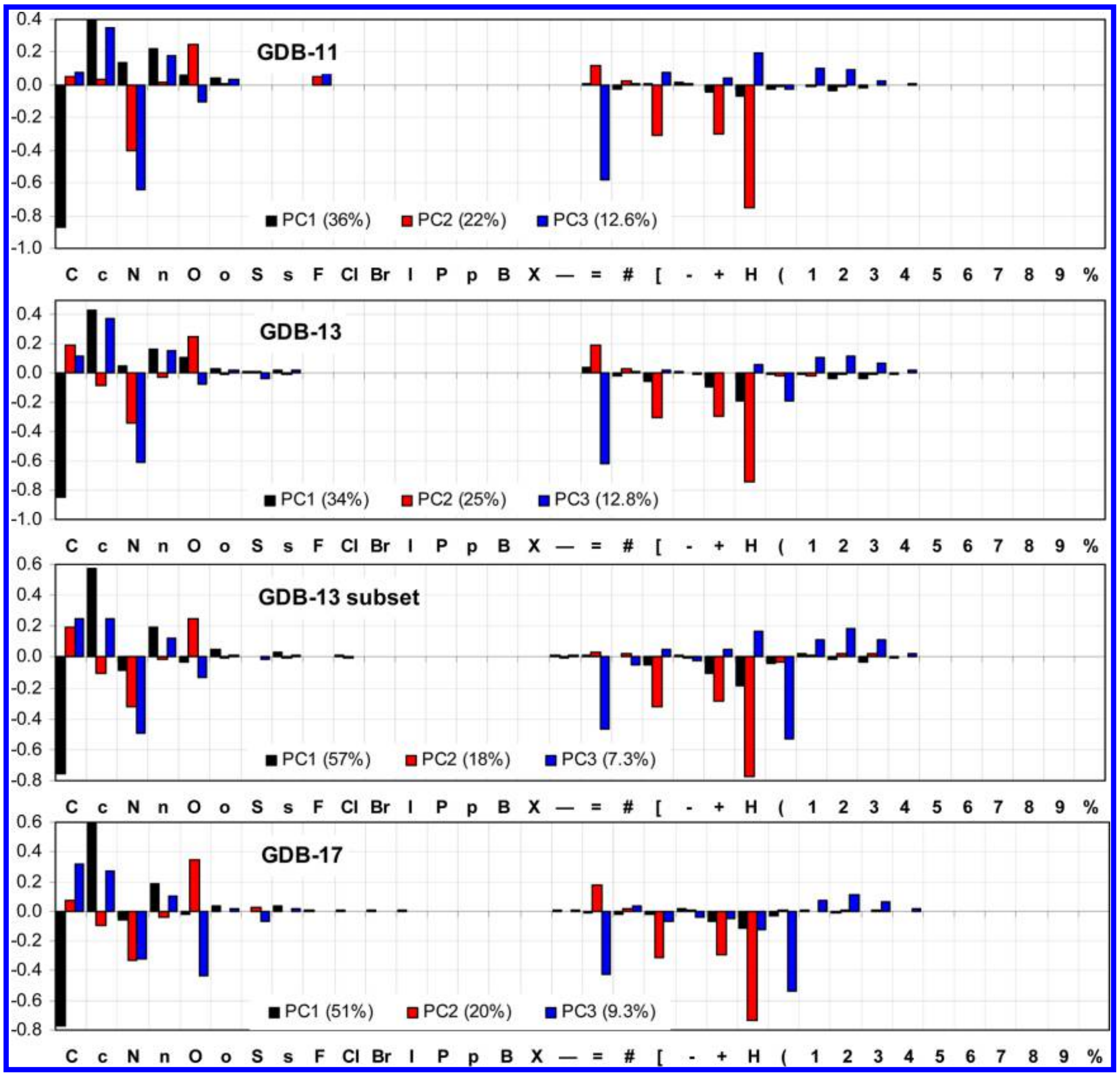


Figure 8. Loading plots for PC1, PC2, PC3 of SMIfp analysis for GDB databases.

Within each stripe, the molecules were distributed vertically according to the number of positive charges (Figure 10C), which also varied according to the number of explicit hydrogen atoms because positive charges and explicit H-atoms in the GDB appear almost exclusively on protonated ammonium groups (Figure 10D).

An intuitive understanding of the SMIfp chemical space was further provided by representation of a virtual database combining the PubChem molecules obeying Lipinski's "rule of 5"[47] and Congreve's "rule of 3"[48] with virtual oligomer categories up to 500 non-hydrogen atoms featuring linear branched alkanes,[49] peptides, oligosaccharides, oligonucleotides, graphenes,[50] and diamondoids.[51,52] This composite database was projected in the (PC1, PC3)-plane using the PC computed for PubChem, which produced a map where molecules of increasing size project radially from the center, with the various oligomer categories projecting as stripes (Figure 11). This representation was quite comparable to that previously obtained from a projection from MQN-space,[22] highlighting the similarities between the two descriptor sets. In this map, the fraction of cyclic atoms per molecule increased from left to right (Figure 11A), while the fraction of carbon atoms, reflecting polarity, increased from top to bottom (Figure 11B). The notable difference to MQN was the differentiation between aromatic and nonaromatic atoms in the SMIfp, which resulted in a separation of molecules according to their content in aromatic carbon from left to right (Figure 11C). Among the virtual oligomer categories, only the graphenes, which are highly aromatic, occupied the right portion of the map (Figure 11D). The ability of SMIfp space to separate molecules according to their content of aromatic atoms may be particularly relevant in the context of the "escape out of flatland" discussion in drug design since it also separates molecules according to their Fsp^3 content.[53,54]

■ CONCLUSION

In summary, we have described a new analysis method for large databases of organic molecules based on classifying molecules on the basis of the composition of their SMILES. The SMIfp (SMILES fingerprint) counts the occurrences of 34 different symbols in the SMILES, from which a 34-dimensional chemical space was created. LBVS using the city-block distance CBD_{SMIfp} as similarity measure gave results comparable to MQN-similarity, with good AUC values and enrichment factors for recovering series of actives from the directory of useful decoys (DUD-E) and from ZINC. Searching by CBD_{SMIfp} is publicly available by a rapid SMIfp-browser at www.gdb.unibe.ch which enables searching of DrugBank, ChEMBL, ZINC, PubChem, GDB-11, GDB-13, and GDB-17. The SMIfp also proved suitable for

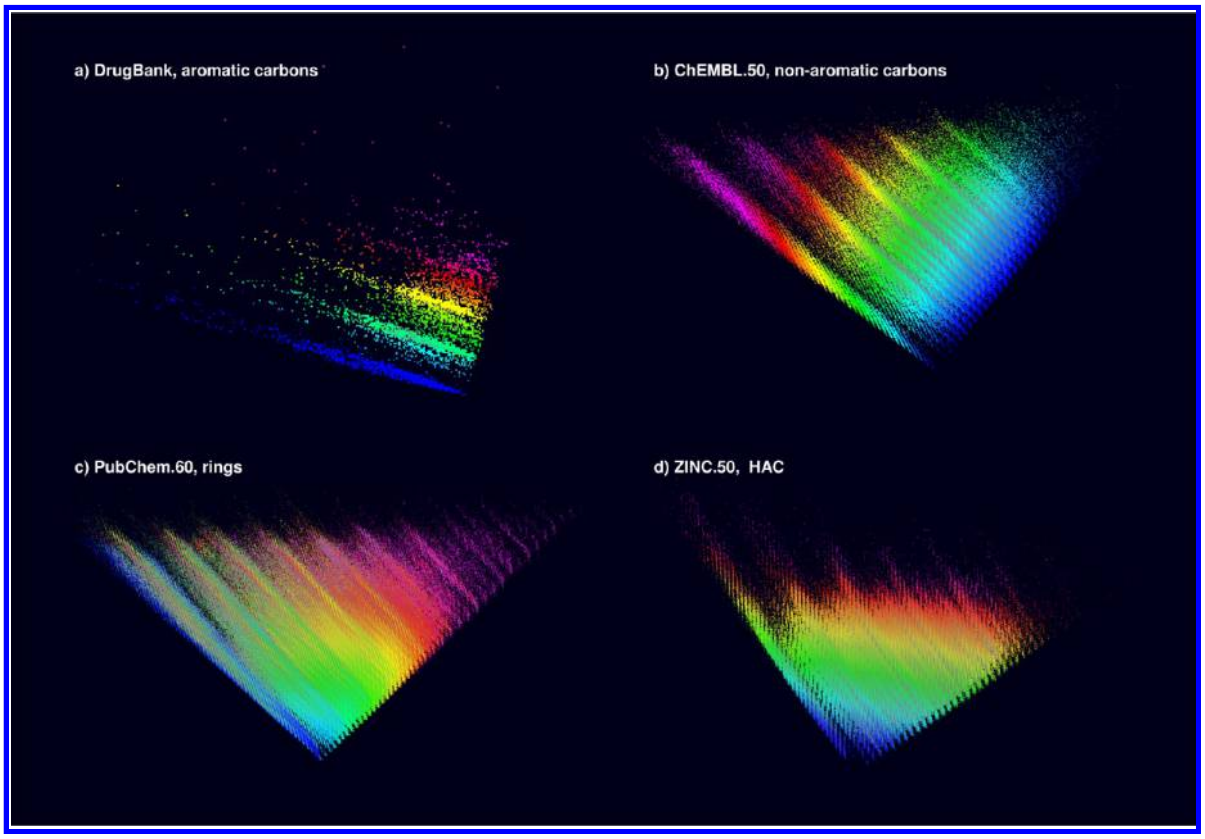


Figure 9. Color-coded maps of the (PC1,PC2)-plane of the SMIfp data set for DrugBank, ChEMBL.50, PubChem.60, and ZINC.50. The color scale spans from blue (lowest value) to magenta (highest value), with saturation to gray according to the standard deviation of the value in each pixel as described previously.[25,26]

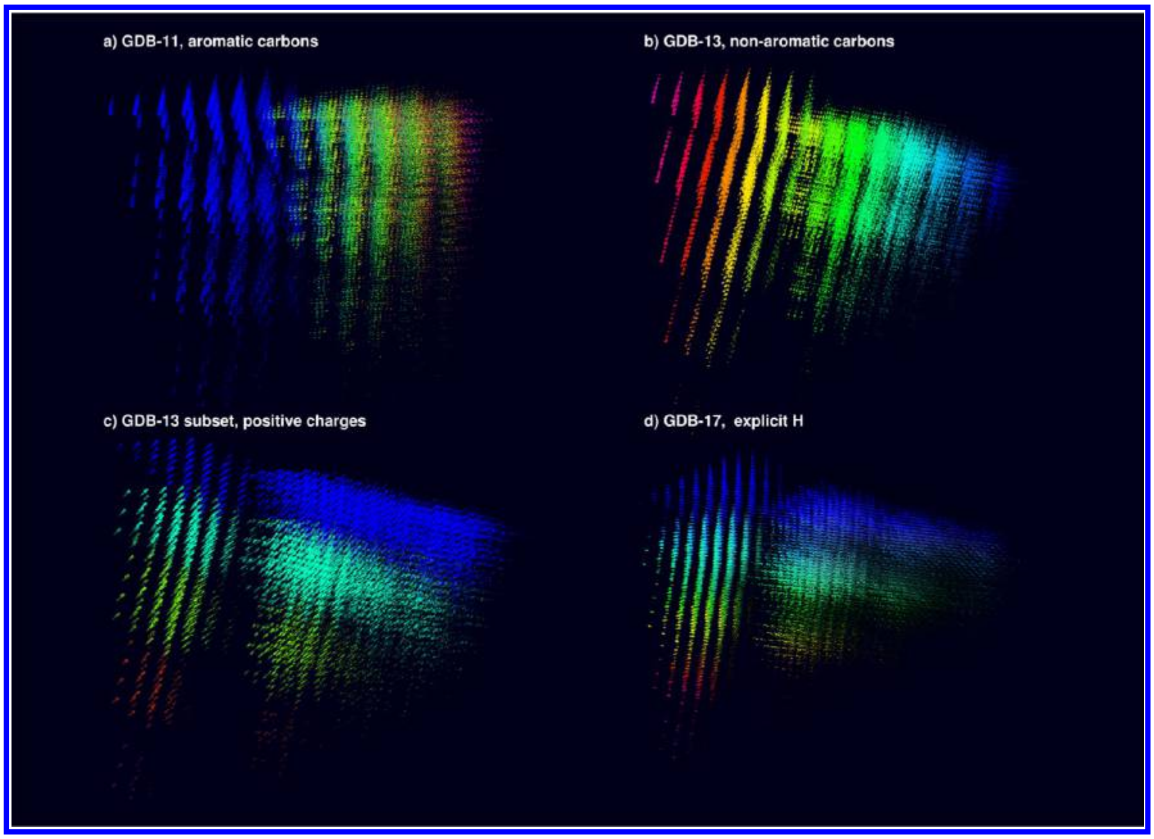


Figure 10. Color-coded maps of the (PC1,PC2)-plane of the SMIfp data set for GDB-11, GDB-13, GDB-13 subset, and GDB-17. See also legend of Figure 9.

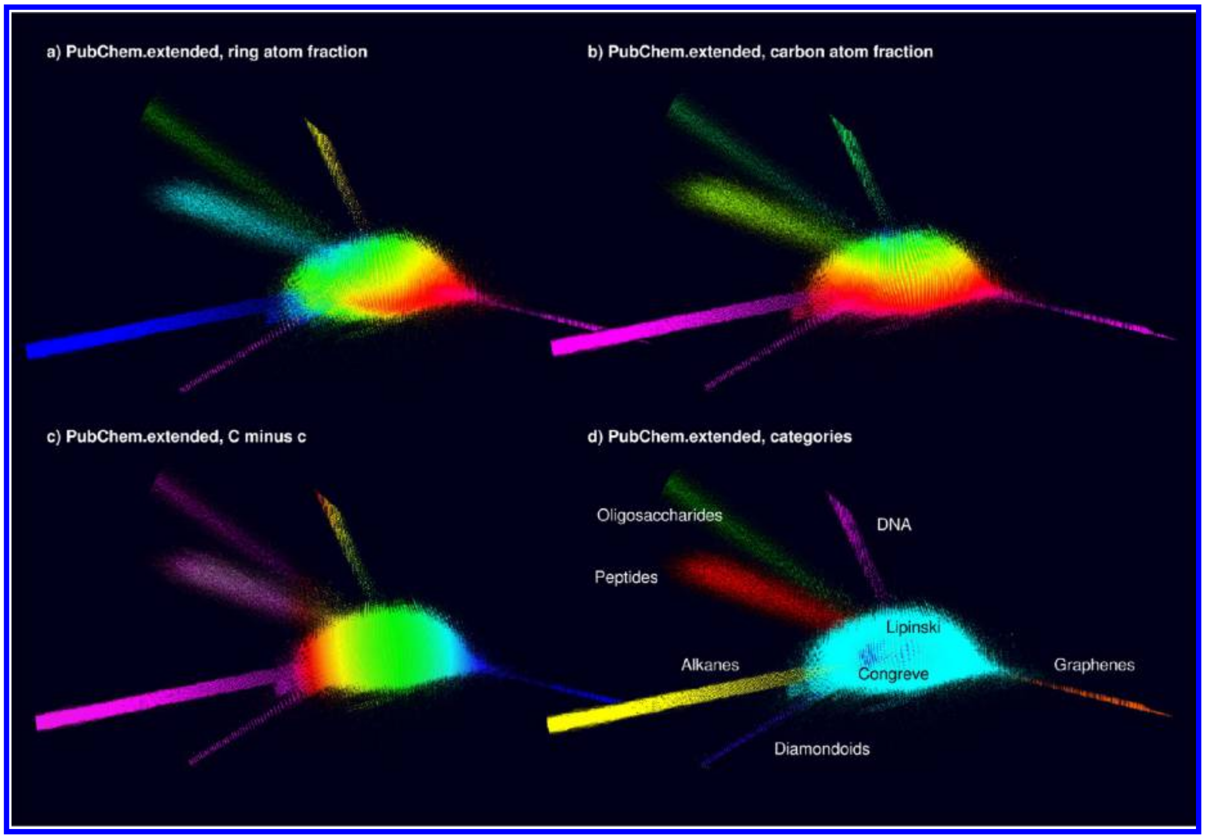


Figure 11. Color-coded maps of the (PC1,PC3)-plane of the SMIfp data set for PubChem.Extended, which contains PubChem molecules obeying Lipinski's "rule of 5" and Congreve's "rule of 3", as well as virtual categories of oligomers up to 500 non-hydrogen atoms as described previously.[22,26] See also legend of Figure 9 and Methods. In panel d, the pixel color is assigned to the majority category in that pixel.

database visualization using PCA of the SMIfp data sets and representation of color-coded maps of the (PC1, PC2)-planes for these databases. Interactive access to the molecules was made possible by the Java application SMIfp-MAPPLET that can be obtained freely from www.gdb.unibe.ch. These maps spread molecules according to their fraction of aromatic atoms, size and polarity. SMIfp provides a new and relevant entry into the small molecule chemical space. In particular, the separation between aromatic and nonaromatic molecules in SMIfp space should provide interesting insights with respect to the "escape out of flatland" discussion in drug design.[53,54] The SMIfp-MAPPLET and SMIfp-browsers should be generally useful to facilitate the exploration of the vast chemical space of known and theoretically possible molecules in search for new drugs.

■ METHODS

Database Selection. DrugBank, ChEMBL, PubChem, GDB-11, GDB-13, and GDB-17 databases were downloaded in February 2013 from their respective web sites (Table 1). ChEMBL.50 and Pubchem.60 are subsets 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 [25] for details). The Pubchem. extended database combines PubChem with six different virtual categories of molecules: peptides, DNA, graphenes, diamondoids, acyclic alkanes, and oligosaccharides (details for peptides, DNA, acyclic alkanes and oligosaccarides available in ref 22), up to a maximum heavy atom count HAC = 500. Molecules were processed into SMILES by removal of counterions.

SMILES Fingerprint Generation. SMILES finger prints (SMIfps) were calculated using an in-house developed Java program which is utilizing Java Chemistry library (JChem) from Chemaxon, Ltd. as a starting point. Before the SMIfp calculation, the ionization state of each molecule was adjusted to pH 7.4 and each molecule was aromatized. The SMIfp was generated by going through the SMILES character-wise and counting the occurrences of single characters or character combinations. Since one character can have different chemical meanings depending on their relative position within the SMILES, the SMIfp generator generally differentiates between characters within and outside of square brackets. Additionally, to differentiate between different elements, the generator has to take the next adjacent character into consideration (e.g., to differentiate between C and Cl). See Supporting Information for full source code.

Ligand-Based Virtual Screening. Enrichment studies were carried out using an in-house developed Java program to recover active ligands of the 102 target specific sets of the "enhanced directory of useful decoys" (DUD-E)[36] from the list of decoys and from the entire ZINC database. The SMILES fingerprint city-block distance ($\mathrm{CBD_{SMIfp}}$), MQN city-block distance ($\mathrm{CBD_{MQN}}$), as well as the 1024-bit daylight-type substructure fingerprint using Tanimoto coefficient (T_{SF})[12,13] were used as similarity measures. Within each set, the molecule most similar to all the other actives (the molecule with highest T_{SF}, respectively lowest CBD_{SMIfp}/CBD_{MQN} to all other actives in the set) was used as reference structure for the virtual screening in the respective similarity metric. The ROC (receiver operator characteristic) curves, as well as the AUC (area under the curve) and the EF (enrichment factors) at 0.1% and 1% were obtained by sorting the resulting data set according to the shortest CBD_{SMIfp}/CBD_{MQN} and highest T_{SF} respectively.

Principal Component Analysis. PCA of the 34-dimensional SMIfp space of all databases was carried out using an in-house developed Java program 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).

SMIfp-Map Generation and Color-Coding. Each molecule was assigned to its plane coordinates (PC1,PC2). The largest (PCmax) and smallest (PCmin) PC values appearing in the PC1 or PC2 values were used to define the value range ΔPC = PCmax − PCmin and set the binning scale as ΔPC/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 SMIfp-map. The (PC1,PC3)-plane was used for PubChem.extended, and the binning was reduced to 300 × 300 pixels for DrugBank.

SMIfp-maps were color coded in the HSL (hue, saturation, luminance) color space as described previously.[25] 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.

Nonlinear Distortion Methods for PubChem.extended. SMIfp-maps of 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 containing the hydrogen atom on the SMIfp-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:

$$\begin{aligned} \Delta PC &= \mathrm{sqrt}(\Delta PC1^2 + \Delta PC2^2) \\ \Delta PC1' &= \Delta PC1(\mathrm{sqrt}(\Delta PC + 1) - 1)/\Delta PC \\ \Delta PC2' &= \Delta PC2(\mathrm{sqrt}(\Delta PC + 1) - 1)/\Delta PC \end{aligned}$$

Molecules in the corrected PC′-plane were binned again into the 1000 × 1000 pixel grid as above.

SMIfp-Mapplet. SMIfp-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. SMIfp-Mapplet can be downloaded free of charge from www.gdb.unibe.ch, where also information is provided 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.

■ ASSOCIATED CONTENT

Supporting Information

ROC curves for recovery of actives from DUD-E (Figure S1), AUC and EF values obtained (Table S1 and S2), and Java source code for computing the SMIfp from SMILES. This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

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) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823−823.

(3) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and lead generation: Beyond high-throughput screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(4) Klebe, G. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580−94.

(5) Kolb, P.; Ferreira, R. S.; Irwin, J. J.; Shoichet, B. K. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotechnol.* **2009**, *20*, 429−36.

(6) 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.

(7) 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.

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

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

(10) Ertl, P.; Rohde, B. The Molecule Cloud—Compact visualization of large collections of molecules. *J. Cheminf. [online]* **2012**, *4*, No. Article 12, http://www.jcheminf.com/content/4/1/12.

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

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

(13) 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.

(14) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) Akella, L. B.; DeCaprio, D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr. Opin. Chem. Biol.* **2010**, *14*, 325−330.

(20) 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.

(21) 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.

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

(23) 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.

(24) Blum, L. C.; van Deursen, R.; Bertrand, S.; Mayer, M.; Burgi, J. J.; Bertrand, D.; Reymond, J. L. Discovery of alpha7-nicotinic receptor ligands by virtual screening of the chemical universe database GDB-13. *J. Chem. Inf. Model.* **2011**, *51*, 3105−3112.

(25) 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.

(26) 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.

(27) Weininger, D. Smiles, A chemical language and information-system 0.1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(28) Weininger, D.; Weininger, A.; Weininger, J. L. Smiles 0.2. Algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(29) Vidal, D.; Thormann, M.; Pons, M. LINGO, An efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(30) Kristensen, T. G.; Nielsen, J.; Pedersen, C. N. Using inverted indices for accelerating LINGO calculations. *J. Chem. Inf. Model.* **2011**, *51*, 597−600.

(31) Vidal, D.; Thormann, M.; Pons, M. A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model.* **2006**, *46*, 836−843.

(32) Karwath, A.; De Raedt, L. SMIREP: Predicting chemical activity from SMILES. *J. Chem. Inf. Model.* **2006**, *46*, 2432−2444.

(33) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(34) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.

(35) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(36) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582−6594.

(37) 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.

(38) 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.

(39) 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.

(40) Fink, T.; Bruggesser, H.; Reymond, J. L. Virtual exploration of the small-molecule chemical universe below 160 Da. *Angew. Chem., Int. Ed. Engl.* **2005**, *44*, 1504−1508.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) 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.

(46) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.

(47) 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.

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

(49) 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.

(50) Allen, M. J.; Tung, V. C.; Kaner, R. B. Honeycomb carbon: A review of graphene. *Chem. Rev.* **2009**, *110*, 132−145.

(51) Dahl, J. E.; Liu, S. G.; Carlson, R. M. Isolation and structure of higher diamondoids, nanometer-sized diamond molecules. *Science* **2003**, *299*, 96−9.

(52) Schwertfeger, H.; Fokin, A. A.; Schreiner, P. R. Diamonds are a chemist's best friend: diamondoid chemistry beyond adamantane. *Angew. Chem., Int. Ed. Engl.* **2008**, *47*, 1022−1036.

(53) 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.

(54) Ritchie, T. J.; Macdonald, S. J.; Young, R. J.; Pickett, S. D. The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. *Drug Discovery Today* **2011**, *16*, 164−171.