A Searchable Map of PubChem

Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond*

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

Received June 21, 2010

The database PubChem was classified using 42 integer value descriptors of molecular structure, here called molecular quantum numbers (MQNs), which count atoms and bond types, polar groups, and topological features. Principal component analysis of the MQN data set shows that PubChem compounds occupy a partially filled elliptical cone in the (PC1,PC2,PC3) space whose axis is the first principal component PC1 (65% variability) representing molecular size, and the ellipse axes are PC2 (18% variability, representing structural flexibility) and PC3 (7% variability, representing polarity). A visual overview of PubChem is provided by color-coded representations of the (PC2,PC3) plane. The MQNs form a scalar fingerprint which can be used to measure the similarity between pairs of molecules and enable ligand-based virtual screening, as illustrated for the enrichment of bioactives from the DUD data set from PubChem. An MQN-annotated version of PubChem with an MQN-similarity search tool is available at www.gdb.unibe.ch.

INTRODUCTION

The development of synthetic chemistry and of cheminformatics has led to an explosion of compound databases available in the public domain. These databases comprise known molecules, including ZINC,[1] BindingDB,[2] PubChem,[3] or Chembl,[4] or theoretically possible molecules, including GDB,[5−7] and together form the chemical space, an unavoidable and fascinating data reservoir to explore molecular diversity in general and drug discovery in particular. It would be highly desirable to have a generally valid classification and visualization system for all of these molecules to facilitate the understanding of molecular diversity.

The words “chemical space”, used to describe the ensemble of all molecules, suggest a spatial classification for molecules. The similarity between molecules including their structural,[8] pharmacophore,[9] or shape similarity[10] or their scaffold relatedness,[11,12] can be used to construct chemical spaces in which pairwise distances are defined on the basis of these similarity measures.[13] Alternatively, one may use sets of physicochemical descriptors to define the chemical space as a property space, taking properties such as polarity and permeability as well as structural parameters such as molecular size and rigidity into account. In this case, maps of the chemical space are produced by principal component analysis and mapping of the (PC1,PC2) plane, such as with the ChemGPS system.[14,15] One can also produce maps of chemical space from sets of descriptors in the form of self-organizing maps (SOMs), in which compounds are distributed in a grid of neurons arranged by similarity, each neuron corresponding to a different descriptor value combination.[16−18] While these classification methods are in principle general, they all rely on measures that are only accessible through complex computations.

We recently proposed a simple classification system for organic molecules based on 42 integer value descriptors of molecular structure, which we call molecular quantum numbers (MQNs).[19] MQNs count evident structural features of molecules such as atom and bond types, polar groups, and topology, which can all be determined “by hand” through visual inspection of the structural formula. The classification of molecules by MQNs is similar to the periodic classification of the elements by their atomic number and principal quantum number.[20] In our preliminary report on the MQN system, we visualized the chemical space of the small molecule databases ZINC[1] and GDB[6] by principal component analysis of the 42-dimensional MQN-space and representation of the plane of the first two principal components. The resulting maps organized molecules by size, number of cycles and rotatable bonds, and polarity. The study also indicated that MQN city block distances (the sum of the absolute differences between MQN values of two molecules) produced meaningful enrichments of bioactivity classes when retrieving compounds from ZINC.

Herein, we report the extension of the MQN system for the analysis and visualization of the PubChem database. PubChem is currently the largest publicly available molecular collection spanning from small organic molecules to relatively large biomolecules such as natural products, peptides, and oligonucleotides.[3] A representation of PubChem as an MQN map is presented providing an unprecedented and readily accessible overview of the PubChem chemical space (Figure 5). Ranking PubChem by MQN similarity to a reference ligand is shown to provide meaningful enrichments of related bioactive compounds, as illustrated with the DUD data set.[21] A MQN-annotated version of PubChem is presented together with a search tool for fast retrieval of nearest neighbors of any molecule in MQN space.

RESULTS AND DISCUSSION

Molecular Quantum Numbers. Forty-two MQNs were chosen as integer value descriptors of molecular structure, as shown in Table 1. MQNs count evident features that can be determined “by hand” through visual inspection of the

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

10.1021/ci100237q © 2010 American Chemical Society
Published on Web 10/14/2010

Table 1. Definition of the 42 Molecular Quantum Numbers (MQNs)

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

[a] hbam counts lone pairs on H-bond acceptor atoms, and hbdm counts H atoms on H-bond-donating atoms for the protonation state predicted at pH = 7.4. [b] afr and bfr count atoms and bonds, respectively, shared by at least two rings in the smallest set of smallest rings.

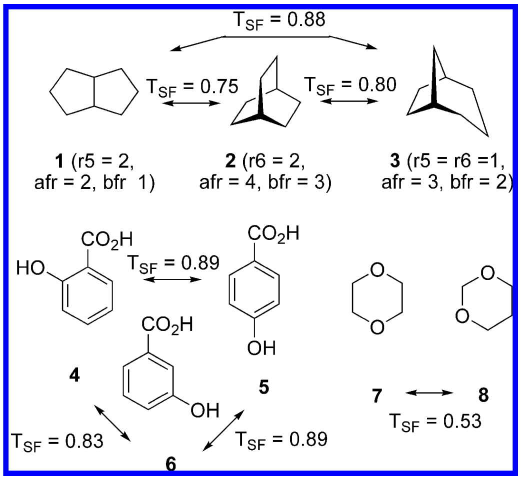


Figure 1. Examples of organic molecules. Hydroxybenzoic acids **4**, **5**, and **6** are MQN isomers, as are the dioxanes **7** and **8**. The Tanimoto similarity coefficient for a 1024-bit Daylight-type structural fingerprint is given as a standard measure of structural similarity (T_{SF}).

structural formula of a molecule (a computational determination is however the only option for analyzing large databases).

Atom counts consider the elements C, N, O, S, P, and halogens, which are most frequently found in organic molecules. Oxygen and nitrogen are subdivided into cyclic and acyclic atoms (co and cn; ao and an) due to the frequent occurrence and prominent role of these elements in determining molecular properties. The heavy atom count (hac) is used to account for the overall molecule size including elements not directly accounted for by a specific MQN such as silicon, boron, or metals.

Bonds are classified as either cyclic or acyclic, with subcategories of single, double, and triple bonds. Bonds are not assigned to elements, with the consequence that functional groups are not counted directly. For example, formaldehyde, $H_2C{=}O$, and ethylene, $H_2C{=}CH_2$, both have an acyclic double bond count adb = 1. The combination with the element count (c = 1 and ao = 1; c = 2) implicitly distinguishes the two molecules without explicitely counting a carbonyl group or an olefin. Similarly, aromatic bonds are not singled out, which circumvents the difficulty of assigning aromatic bonds automatically. Nevertheless, aromatic rings have an implicit MQN signature, for example, a benzene ring adds 1 to the number of six-membered rings (r6) and 3 to the cyclic double bond count (cdb).

Polarity elements, which are strong determinants of physicochemical properties, are counted considering the protonation state predicted for physiological pH. These include H-bond donor and H-bond acceptor counts both per atom (e.g., H_2O: hba = 1, hbd = 1) and including multiplicity (e.g., H_2O: hbam = 2, hbdm = 2). The latter distinguish, for example, ethylamine (hbdm = 3 at pH = 7.4) from dimethylamine (hbdm = 2 at pH = 7.4).

A set of topological descriptors is used to perceive molecular shape, including the counting of cycles of different sizes. Cycles larger than nine (rg10), which are rather rare, are counted together as macrocycles. Counting atoms and bonds present simultaneously in two rings (afr and bfr, considering the smallest set of smallest rings) enhances the differentiation between related polycyclic structures with nonplanar shapes, such as bicyclo[3.3.0]octane (**1**), bicyclo[2.2.2]octane (**2**), and bicyclo[3.2.1]octane (**3**) (Figure 1). The perception of shape is not well encoded in standard 2D-structural fingerprints, as illustrated by the relatively high Tanimoto similarity coefficients between **1**, **2**, and **3**.

MQNs are not unambiguous descriptors of molecular structure, nor are they thought to be so. Although many cases exist in which a given MQN combination only corresponds to a single compound, in particular for small molecules such as ethylene and formaldehyde or the bicyclic alkanes **2** and **3**, different molecules may share the same MQN combination. Such MQN isomers in the MQN system of molecules correspond to isotopes in the periodic system of the elements. They include all stereoisomers of a molecule since stereochemical descriptors are not considered in the MQNs. For example the bicyclo[3.3.0]octane **1** has one *cis*-fused and one *trans*-fused stereoisomer with the same MQN combination. MQN isomers also comprise regioisomers that chemists classically assign to the same families (e.g., *ortho*, *meta*, and *para* isomers of disubstituted benzene rings such as **4**−**6**, Figure 1), and atom permutation isomers such as 1,4-dioxane (**7**) and 1,3-dioxane (**8**). For larger molecules, all building block permutation isomers are MQN isomers, such as peptides of identical amino acid composition and length but different sequence.

Many MQNs are strongly correlated, as will be evident from the principal component analysis below. In particular, and as for most structural descriptors, the MQN values

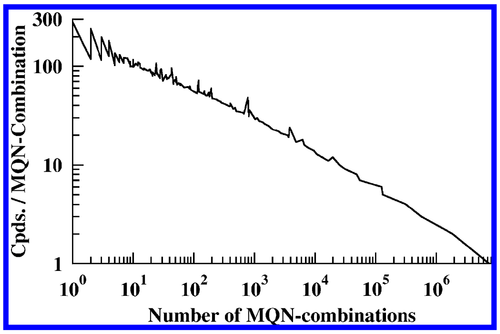


Figure 2. Distribution of the 19.2 million PubChem compounds to the 10.7 million MQN combinations.

generally increase with the molecular size, which is a primary determinant of molecular properties.

Visualizing the MQN Space of PubChem. PubChem is a publicly available database that archives the molecular structures and bioassay data within the National Institute of Health (NIH) Roadmap for Medial Research Initiative. With 44 560 000 entries (as of November 2009), PubChem is currently the largest publicly available molecular database. PubChem contains not only drug-sized compounds such as those included in ZINC but also larger molecules such as natural products, peptides, oligosaccharides, and oligonucleotides. The largest molecule registered in PubChem is an oligonucleotide of 28 bases (hac = 595, MW = 9000.36). For our MQN analysis, we considered the subset of molecules with elements C, N, O, S, P, and halogens and eliminated redundant structures such as various salts and isotopic compositions of the same compounds, which left 19 186 705 SMILES (43% of the raw database) for consideration. These 19 186 705 SMILES gave 10 720 901 different MQN combinations. Approximately 39% of the SMILES corresponded to a unique MQN combination, while the most occupied MQN combination featured 282 different SMILES (Figure 2).

Principal component analysis (PCA) was performed to gain an insight into the data structure and variability. The first three principal components (PC) accounted for 90% of the variability (Figure 3). PC1 accounted for 65% of the variability and covered molecular size, with positive loadings only and highest values in the heavy atom count (hac) and the carbon count (c). PC2 (18% of variability) corresponded to molecular rigidity, with strong PC2-positive loadings for cyclic single bonds (csb), cyclic divalent nodes (cdv), and cyclic trivalent nodes (ctv) and strong PC2-negative loadings for acyclic single bond (asb), rotatable bond count (rbc), acyclic divalent nodes (adv), and acyclic monovalent nodes (asv). PC3 (7% of variability) described polarity, with the strongest contributions from H-bond acceptor counts (hba and hbam) and the number of acyclic oxygen atoms (ao).

Visualization of the occupancy map of the (PC1,PC2) plane covering 83% of the variability showed that the PubChem compounds occupied a roughly triangular surface extending to positive PC1 values and limited by maximal and minimal boundaries in the PC2/PC1 ratio (Figure 4A). Due to the prevalence of small molecules in the database, the occupancy was particularly strong in the lower PC1

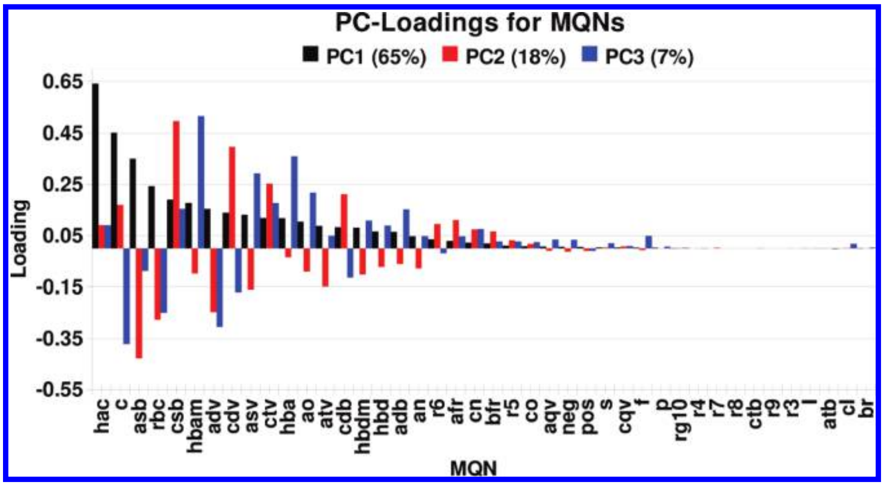


Figure 3. Loading of the first three principal components for the MQN analysis of PubChem. MQNs are ordered by decreasing PC1 value.

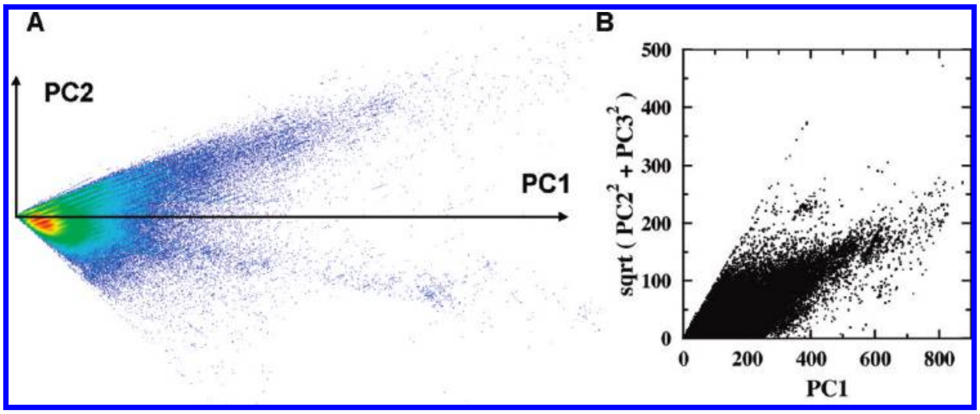


Figure 4. (A) The MQN space of PubChem viewed in the (PC1,PC2) plane. Color-coding following occupancy on the logarithmic scale from blue (one cpd) to red (100 000 cpds). (B) Distance from the PC1 axis in the (PC2,PC3) plane as a function of PC1.

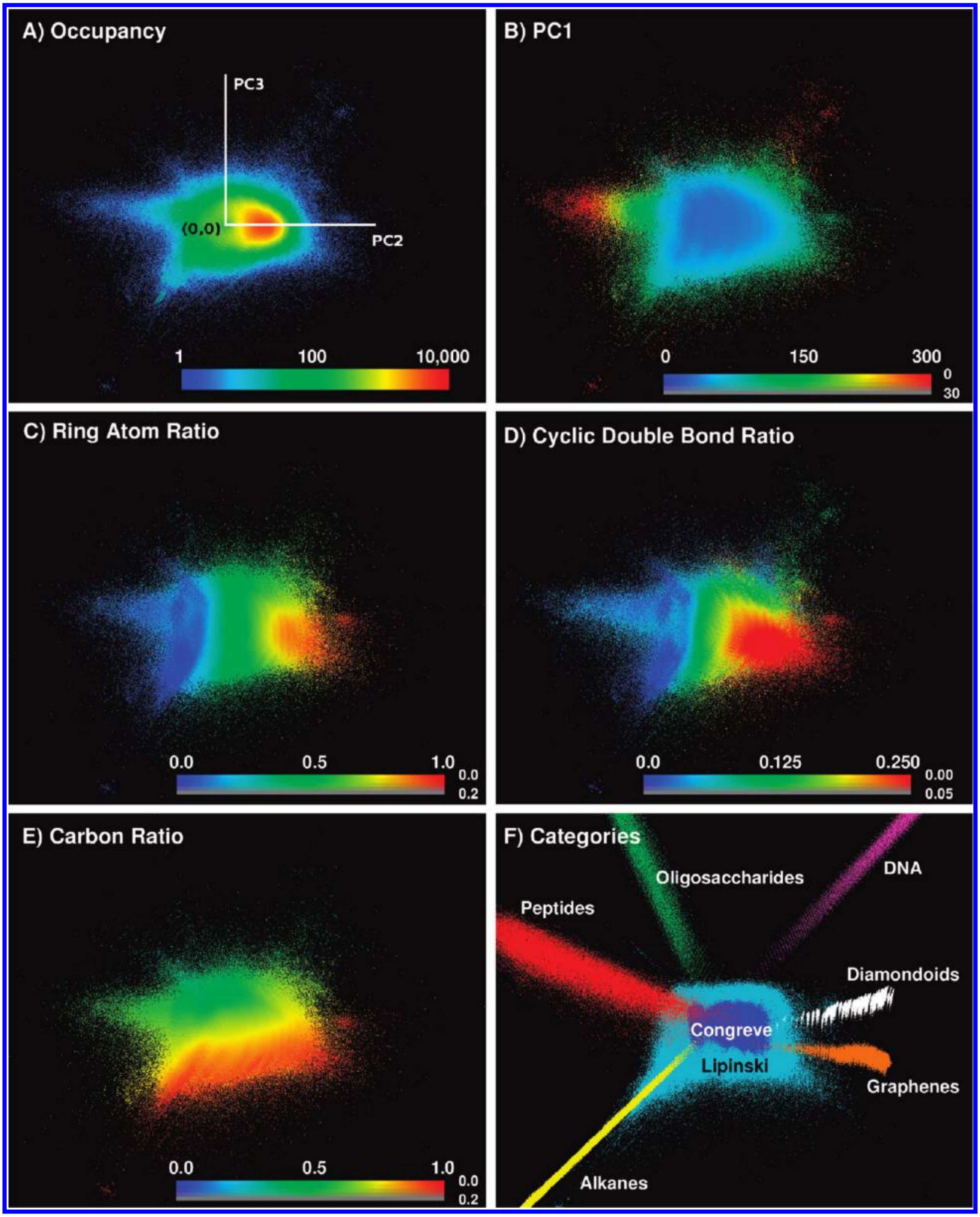


Figure 5. MQN map of PubChem as the (PC2,PC3) plane. The PC2 and PC3 values were corrected as follows: PC2′ = sqrt(1 + PC2) − 1, PC3′ = sqrt(1 + PC3) − 1. The surface was hashed in a 960 × 793 grid, and each pixel was colored with the color (hue in HSL scale) following the average descriptor value and the saturation to gray (saturation in HSL scale) following the standard deviation, as indicated in the reference scales in each image, with an in-house developed source code. Color codes according to (a) occupancy; (b) PC1, which primarily codes for molecular size: MW = (9.8 ± 1.1) × PC1, hac = (0.67 ± 0.02) × PC1; (c) the ratio of cyclic atoms to all atoms in the molecule; (d) the ratio of cyclic double bonds to all bonds in the molecule; (e) the ratio of carbon atoms to all atoms in the molecules, indicative of polarity; (f) compound categories. Lipinski's "rule of 5" cpds:15 889 684 compounds, 87.4% of PubChem. Congreve "rule of 3" cpds: 880 480 cpds, 4.6% of PubChem. In silico generated virtual libraries: peptides are linear peptides of up to 80 proteinogenic amino acids with random sequence (100 000 cpds). Graphenes are aromatic sheets with methyl groups added on up to 20% of the available CH bonds, with a size up to 176 carbon atoms (146 759 cpds). Diamondoids are extended adamantanes containing methyl groups added randomly on up to 20% of the available CH bonds with a size up to 151 carbon atoms (106 637 cpds). Alkanes are purely acyclic aklanes including branched alkanes up to 500 atoms (500 000 cpds). Oligosaccharides are 1,4-oligomers of up to 50 units of glucose, N-acetyl glucosamine, or 6-sulfoglucose (31 044 cpds). DNA are oligomers of up to 40 DNA bases (ATGC) with random sequences (40 000 cpds).

values corresponding to MW < 500, while the higher PC1 values were sparsely populated. The PubChem compounds were distributed in the (PC1,PC2,PC3) space as a partially filled cone whose axis is the PC1 axis, as indicated by the distribution of distances from the PC1 axis as a function of PC1 value (Figure 4B).

An informative map of the MQN space of PubChem was obtained by representing the (PC2,PC3) plane, corresponding to viewing the partially filled cone along its axis. In this manner, molecular flexibility and polarity were readily visible. A square root correction of the PC2 and PC3 values was introduced to facilitate the representation of both small and large molecules. The map was then color-coded according to various structural features (Figure 5). The occupancy map showed the highest compound density at the center corresponding to molecules with MW < 500 fulfilling Lipinski's criteria (87.4% of PubChem; Figure 5A). Although PC1 representing molecular size was not directly encoded in PC2 and PC3, compounds of increasing PC1 values were located at increasing ditances from the (0,0) coordinate where hydrogen is located (Figure 5B). In agreement with the PC2 loadings, flexible and acyclic molecules appeared at negative PC2 values and cyclic, rigid molecules at positive PC2 values. This distribution is illustrated by the color-coded representation of the ratio of cyclic atoms to all atoms per molecule (Figure 5C). PC2 also partially followed the ratio of cyclic double bonds to all bonds per molecule (Figure 5D), a value which also influenced PC3. PC3 however mostly represented polarity as illustrated by color-coding according to the ratio of carbon atoms to all atoms per molecule, which indicates polarity since heteroatoms are mostly polar atoms such as N and O (Figure 5E).

Compound classes were distributed in different regions of the map according to their structural characteristics (Figure 5F). Compounds following lead-like (Congreve's rule of 3)[22] and bioavailability criteria (Lipinski's rule of 5)[23] occupied the center of the map, which corresponds to the low molecular weight region. Virtual libraries of in silico generated biopolymers such as peptides, carbohydrates, or oligonucleotides and compound families such as hydrocarbons, graphenes,[24] and diamondoids[25,26] stretched out from the center to the outside, corresponding to an increase in molecule size, at angles specific for the polarity and rigidity of their constituent building blocks. Apolar molecules (e.g., alkanes, diamondoids, graphenes) occupied the southern portion of the map, while polar molecules (carbohydrates, DNA, peptides) were found in the northern half of the map. On the other hand, rigid polycyclic molecules (DNA, diamondoids, graphenes) were found at right and flexible acyclic molecules (alkanes, peptides, oligosaccharides) at left.

Molecular Similarity by MQNs versus Structural Fingerprints. While MQNs define the position of a molecule in a formally 42-dimensional chemical space, they also compose a scalar fingerprint description of molecules. Conversely, a typical binary structural fingerprint (SF), in which bit values are switched from 0 to 1 whenever a particular substructure is present,[27] defines a multidimensional binary chemical space. Both the MQN and the SF data can be used for similarity comparisons between pairs of molecules. The chemical space concept suggests the use of a distance measure, most simply the "city-block distance" (CBD) between two molecules obtained by summing the absolute values of the differences between the MQN or SF bit values of each molecule. On the other hand, the fingerprint concept calls for calculating the well-known Tanimoto similarity coefficient (T), which is the ratio of shared values to cumulated values for the considered pair of SF or MQN data.[27,28]

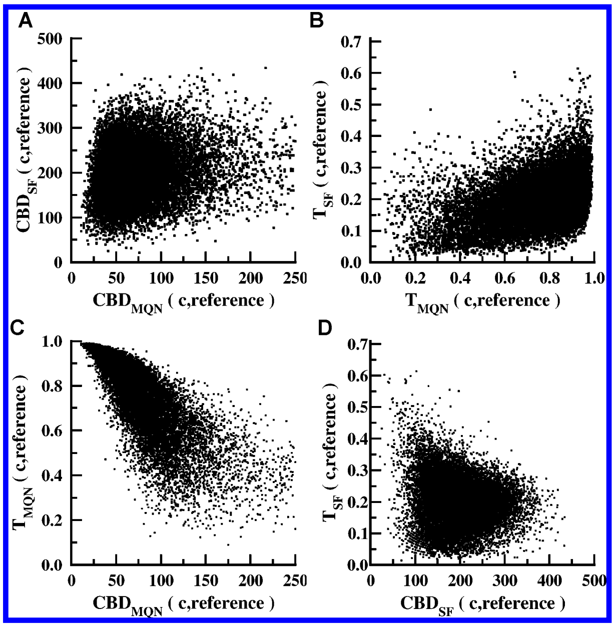
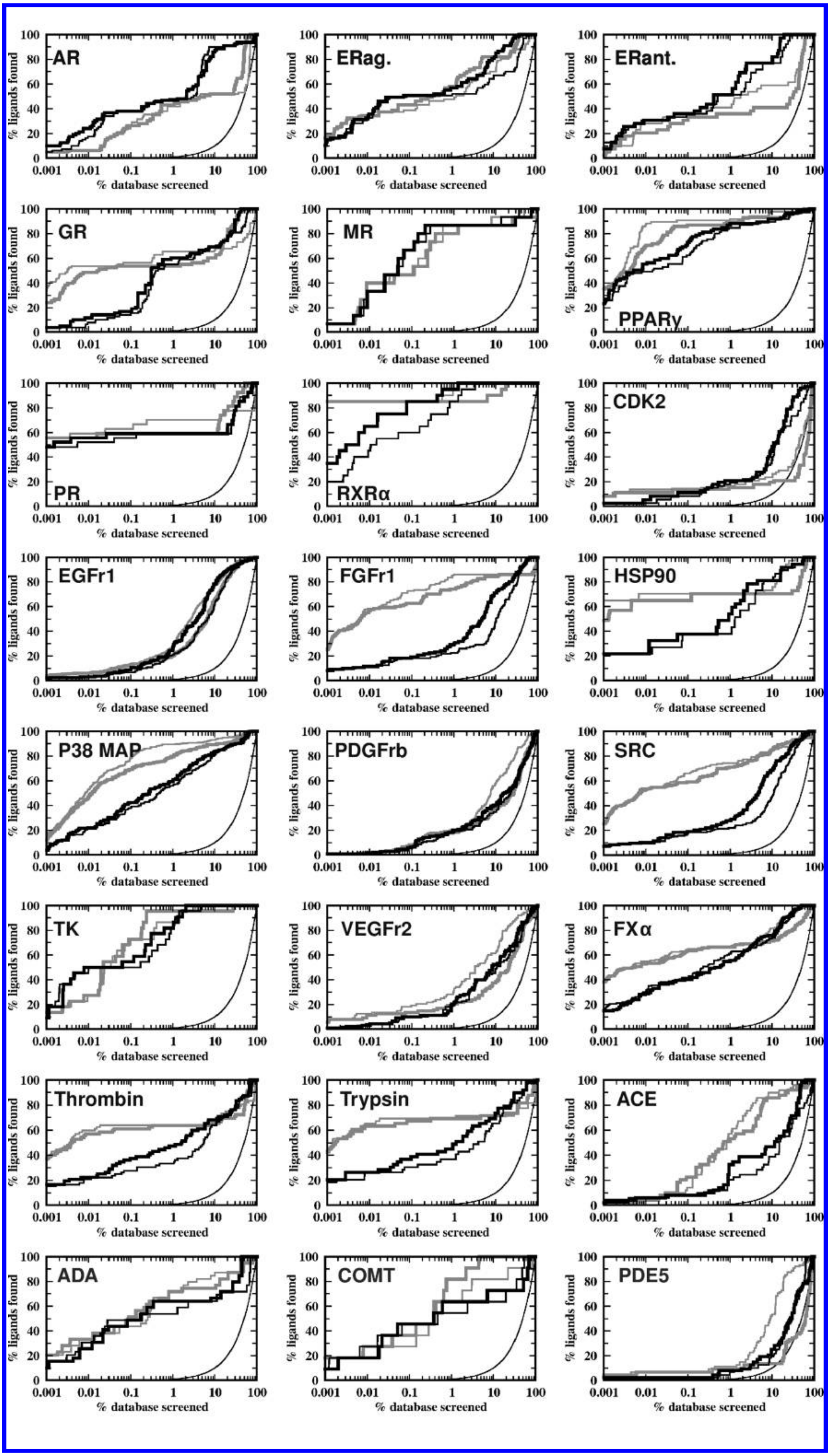


Figure 6. Comparison of MQNs with structural fingerprints for 20 000 randomly selected compound pairs from PubChem. (A) CBD_{SF} vs CBD_{MQN}, (B) T_{SF} vs T_{MQN}, (C) T_{MQN} vs CBD_{MQN}, (D) T_{SF} vs CBD_{SF}. The SF is a Daylight-type 1024 bit fingerprint.

Comparing 20 000 random pairs of molecules from PubChem showed that the MQN and SF descriptions are only weakly correlated with one another, independent of whether CBD or T is used to compare the data (Figure 6A,B). The difference between MQN and SF similarity reflects the very different approaches taken for structural description in the two systems. Indeed, the MQNs count structural features not considering any functional groups, while SF lists the presence (but not the count) of well-defined substructures including functional groups. It should be noted that the similarity measure (CBD or T) also influences the relationship between compounds within each representation method (MQN or SF; Figure 6C,D).

Virtual Screening of PubChem in MQN Space. In view of the separation of compound classes in the MQN map (Figure 5F), one can anticipate that the MQN system might generally group compounds that are structurally related and perhaps even similar in their biological properties, suggesting its use for virtual screening. In other words, the similarity measures CBD_{MQN} and T_{MQN} described above might be used for scoring PubChem relative to a reference molecule to search for other compounds with similar properties. The feasability of this approach was tested using the 40 classes of bioactive compounds reported in the DUD data set.[21] Retrieval of the active compounds from PubChem was tested using the compound nearest to the MQN average across all actives as the reference structure, considering CBD_{MQN}, T_{MQN}, CBD_{SF}, or T_{SF} as the similarity measure. Retrieval of actives was estimated from the enrichment plots (Figure 7) and quantified by the enrichment factors (EF) at 0.1% and 1% coverage of PubChem, which measure the ratio


AR
ERag.
ERant.
GR
MR
PPARγ
PR
RXRα
CDK2
EGFr1
FGFr1
HSP90
P38 MAP
PDGFrb
SRC
TK
VEGFr2
FXα
Thrombin
Trypsin
ACE
ADA
COMT
PDE5
% ligands found
% database screened
0
20
40
60
80
100
0.001
0.01
0.1
1
10
100

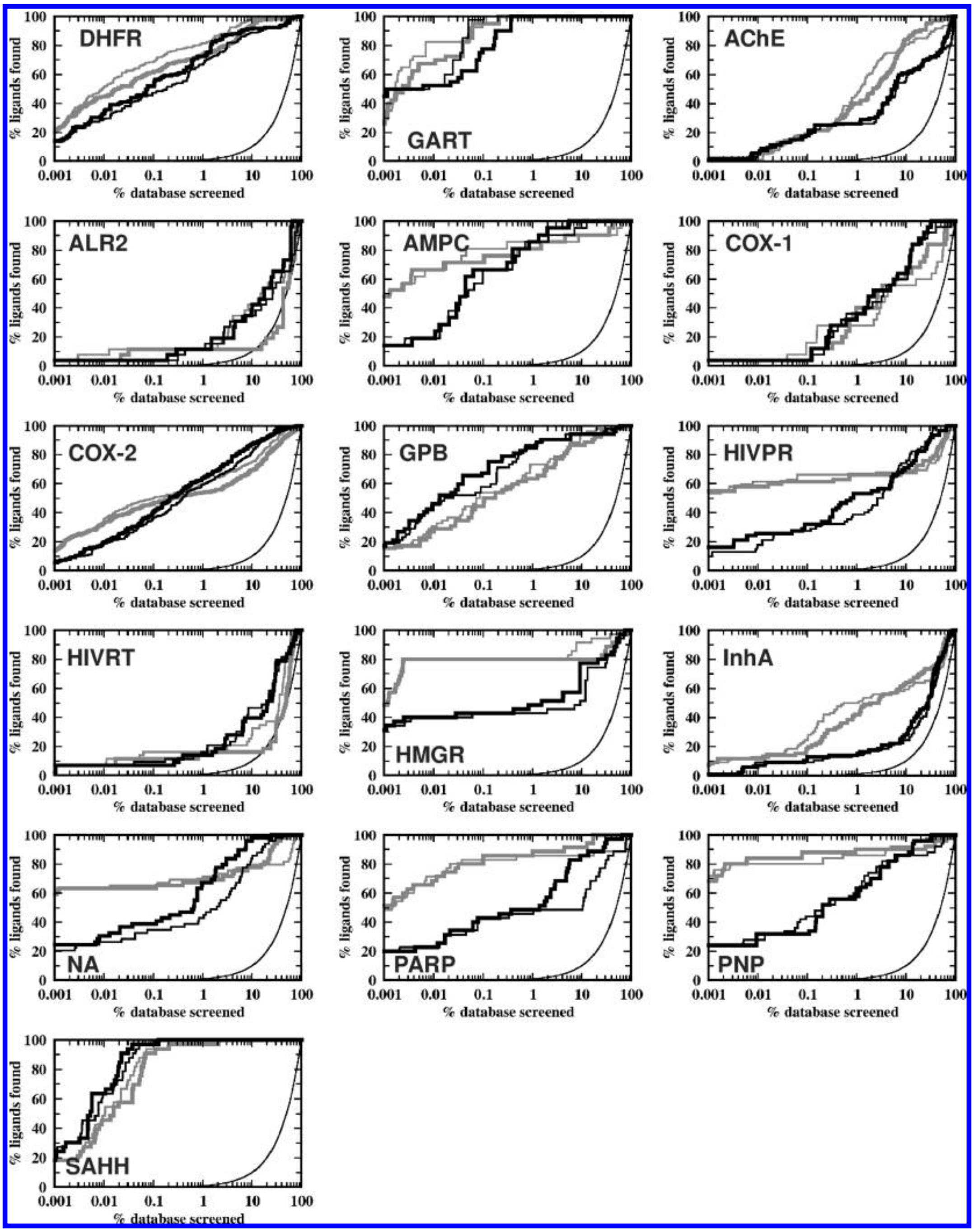


Figure 7. Virtual screening of the entire PubChem database (19.2 million cpds) for each of the 40 bioactivity classes in the DUD data set (in the same order). The compound nearest to the MQN center of gravity was selected as the lead structure for similarity calculation according to $\mathrm{CBD_{MQN}}$ (fat black line), T_{MQN} (thin black line), $\mathrm{CBD_{SF}}$ (fat gray line), and T_{SF} (thin gray line). The expected enrichment for random screening is shown as a continuous curved thin black line.

of found versus randomly expected actives in the first 19 187 and 191 867 compounds (Table 2), respectively.

On average, the best enrichments were observed using structural fingerprints compared by Tanimoto similarity (T_{SF}: $\mathrm{EF}_{0.1} = 512 \pm 279$, $\mathrm{EF}_1 = 62 \pm 26$) or by city-block distance ($\mathrm{CBD_{SF}}$: $\mathrm{EF}_{0.1} = 491 \pm 273$, $\mathrm{EF}_1 = 60 \pm 26$), which probably reflects the fact that many actives in the DUD data set are from series developed by analog synthesis on the basis of common substructures. Enrichments by MQN were slightly lower ($\mathrm{CBD_{MQN}}$: $\mathrm{EF}_{0.1} = 378 \pm 254$, $\mathrm{EF}_1 = 52 \pm 26$; T_{MQN}: $\mathrm{EF}_{0.1} = 349 \pm 231$, $\mathrm{EF}_1 = 48 \pm 26$), but the differences between the enrichments obtained by MQN and by SF were not significant (Tanimoto: $\Delta(\mathrm{EF}_{0.1}) = -163 \pm 190$, $\Delta(\mathrm{EF}_1) = -14 \pm 18$; CBD: $\Delta(\mathrm{EF}_{0.1}) = -113 \pm 180$, $\Delta(\mathrm{EF}_1) = -8 \pm 16$).

Choosing the most MQN-average compound as a reference for the virtual screening exercise above introduced a bias favoring the MQN-similarity enrichment. Nevertheless, the performances obtained with MQN similarity are quite remarkable considering the simplicity of the MQNs compared to SF. This indicates that classes of compounds with similar bioactivity often form groups in MQN space. Most interestingly, high scoring compounds in MQN similarity were often very different from high scoring compounds by SF similarity, as could be expected from the low correlation between the two measures (Figures 6, 8A,B). High-scoring

Table 2. Enrichment Factors for Virtual Screening of the Entire PubChem Database (19.2 million cpds) Using City-Block Distances (CBD) and Tanimoto Similarity Coefficient (T) between MQNs and Structural Fingerprints (SF)[a]

		$EF_{0.1}$				EF_1			
	no. of actives[b]	CBD_{MQN}	T_{MQN}	CBD_{SF}	T_{SF}	CBD_{MQN}	T_{MQN}	CBD_{SF}	T_{SF}
nuclear hormone receptors									
AR	79	379.5	379.5	265.6	265.6	46.8	48.1	43.0	41.8
ER agonist	67	507.1	507.1	432.5	387.8	56.7	50.7	58.2	47.8
ER antagonist	39	358.7	358.7	333.1	333.1	51.3	51.3	35.9	41.0
GR	78	166.6	140.9	538.1	563.7	60.2	55.1	53.8	65.4
MR	15	666.2	666.2	466.3	466.3	86.7	86.7	80.0	86.7
PPARg	85	728.9	623.1	870.0	905.3	87.0	84.7	89.4	91.7
PR	27	592.2	555.2	592.2	629.2	59.2	59.2	59.2	70.4
RXRa	20	849.4	599.6	849.4	849.4	95.0	85.0	85.0	100.0
kinases									
CDK2	72	111.0	83.3	138.8	138.8	20.8	18.1	15.3	16.7
EGFR	475	90.5	67.3	126.2	132.5	25.5	20.8	20.2	27.2
FGFr1	120	191.5	183.2	624.6	724.5	29.2	22.5	74.2	85.0
HSP90	37	378.1	378.1	648.2	702.2	54.0	37.8	70.3	70.3
P38 MAP	454	424.8	380.8	691.2	783.6	59.5	55.9	79.5	89.4
PDGFrb	170	64.7	52.9	82.3	82.3	20.0	18.2	18.2	20.0
SRC	159	188.5	182.3	590.8	659.9	27.7	22.0	71.1	74.2
TK	22	545.1	499.7	726.8	726.8	81.8	81.8	95.4	86.3
VEGFr2	88	102.2	102.2	136.3	193.0	20.5	19.3	20.5	30.7
serine proteases									
FXa	146	403.8	417.5	636.5	650.2	55.5	60.9	66.4	66.4
thrombin	72	374.7	277.6	610.7	638.4	47.2	33.3	63.9	63.9
trypsin	49	387.5	305.9	673.0	693.4	49.0	36.7	69.4	69.4
metalloenzymes									
ACE	49	81.6	81.6	224.3	142.8	32.6	20.4	53.1	57.1
ADA	39	486.8	486.8	512.5	435.6	64.1	53.8	71.8	71.8
COMT	11	454.2	454.2	363.4	272.5	63.6	54.5	81.8	63.6
PDE5	88	22.7	22.7	68.1	68.1	8.0	6.8	9.1	11.4
folate enzymes									
DHFR	410	519.2	475.3	616.7	684.9	72.4	66.6	73.6	81.5
GART	40	974.3	774.5	949.3	949.3	100.0	100.0	100.0	100.0
other enzymes									
AchE	107	177.4	177.4	186.8	196.1	26.2	26.2	40.2	45.8
ALR2	26	38.4	38.4	115.3	115.3	11.5	11.5	11.5	11.5
AmpC	21	666.2	666.2	713.8	809.0	85.7	85.7	80.9	85.7
COX-1	25	40.0	40.0	40.0	159.9	36.0	36.0	40.0	28.0
COX-2	426	401.1	361.3	459.8	497.3[a]	63.8	59.4	53.7	59.4
GPB	52	672.6	576.5	442.0	538.1	84.6	86.5	63.5	73.1
HIVPR	62	322.4	274.0	612.5	660.8	53.2	38.7	66.1	66.1
HIVRT	43	93.0	69.7	116.2	162.7	16.3	14.0	14.0	16.3
HMGR	35	428.3	428.3	799.4	799.4	48.6	42.8	80.0	80.0
InhA	86	127.8	93.0	185.9	244.0	14.0	14.0	41.9	50.0
NA	49	387.5	346.7	632.2	652.6	67.3	42.8	69.4	69.4
PARP	35	428.3	428.3	799.4	828.0	48.6	45.7	85.7	85.7
PNP	50	319.8	439.7	879.4	799.4	58.0	64.0	90.0	86.0
SAHH	33	969.0	969.0	908.5	938.7	100.0	100.0	97.0	100.0

[a] The denomination and order of the bioactivity classes is taken from ref 21. The maximum possible $EF_{0.1}$ is 1000; the maximum possible EF_1 is 100, corresponding to the possible increase in recovery compared to random screening. [b] Number of sdf-file entries for active ligands in the DUD data set as downloaded from dud.docking.org.

known actives selected by MQN similarity to the reference compounds comprised many molecules with low SF similarity and containing different scaffolds. This suggests that MQN similarity allows lead-hopping relationships between compounds which are typically not possible when selecting for SF similarity (Figure 8C,D).

On the other hand, CBD_{MQN} or T_{MQN} were not able to distinguish actives from the decoys proposed within the DUD data set. This is not surprising considering that decoys were designed as molecules with low T_{SF} but similar size, composition, and physicochemical properties as the actives,[21] which corresponds to compounds having similar MQN values as the actives. This shows that an MQN-based enrichment can only serve as a fast sorting function that must be complemented with other, more refined virtual screening tools validated for the distinction of ligands from decoys, such as docking or shape-based scoring functions.[29,30]

In view of the enrichment examples above, retrieving nearest neighbors in MQN space should be an attractive and very fast method for preselection toward new and potentially bioactive analogs of any lead structure; in particular, considering that MQN similarity seems to allow lead-hopping. Sorting by CBD_{MQN} was programmed as a fast retrieval tool containing the annotated PubChem database organized by MQN groups. Nearest neighbor searches were performed using this tool for selected drugs and natural

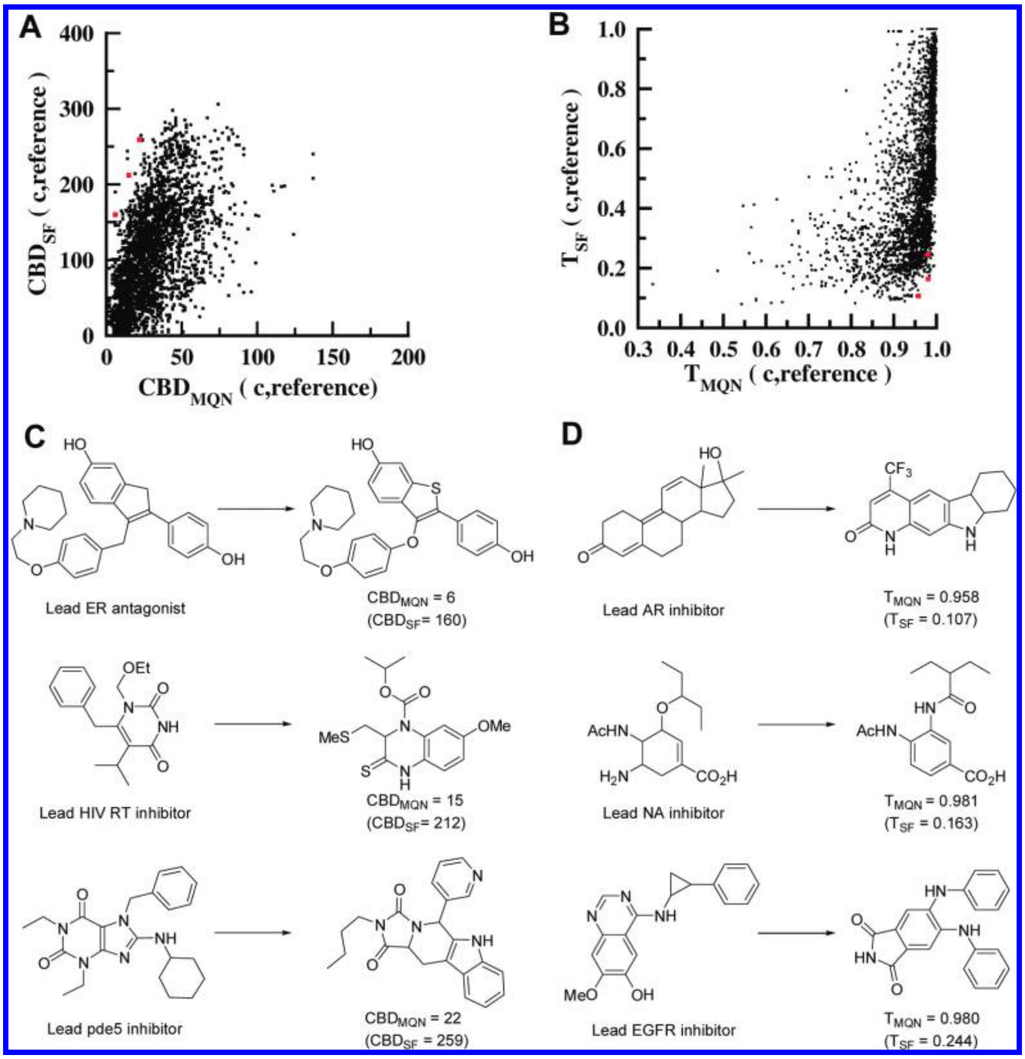


Figure 8. Lead-hopping characteristics of MQN vs SF. (A) City-block distances (CBD) and (B) Taninomoto similarity coefficients (T) in MQN (x axis) versus SF (y axis) between all actives from the DUD data set (40 targets) and their corresponding lead reference used for virtual screening. (C) Examples of lead-hopping virtual hits identified within the known active by CBD_{MQN} relative to the lead reference. D. Examples of lead-hopping virtual hits identified within the known active by T_{MQN} relative to the lead reference. The lead reference is in each case the structure closest to the MQN average across the actives. The position of the examples shown is marked with red dots in the scatter plot.

Table 3. City-Block Distance Nearest Neighbors of Selected Drugs in the MQN Space of PubChem

	CBD_{MQN} (nearest 1000 cpds)[a]			CBD_{MQN} (all PubChem)[b]	
drug	min	max	avg ± stdev	max	avg ± stdev
Keppra (**9**)	1	10	8.1 ± 1.5	3327	87.5 ± 58.2
Morphine (**10**)	0	19	15.5 ± 3.8	3269	88.6 ± 52.1
Vancomycin (**11**)	0	101	69.2 ± 27.6	2841	446.4 ± 43.7
Salmeterol (**12**)	0	18	15.4 ± 3.1	3231	88.3 ± 44.0
Penicillin G (**13**)	3	14	12.0 ± 2.1	3276	63.1 ± 51.5
Tamiflu (**14**)	2	14	12.6 ± 1.6	3280	75.6 ± 50.7

[a] City-block distance calculated for the 1000 nearest neighbors on the 42-dimensional MQN vector. [b] City-block distance measured for entire PubChem.

products in PubChem by ranking the database by CBD_{MQN} relative to each query molecule (Table 3, Figure 9A). The distance distribution of PubChem compounds relative to each query molecule showed approximately Gaussian distributions, with the nearest 1000 compounds corresponding to the tailing end of each curve at the shortest CBD_{MQN} distances (Figure 9B). Inspection of the structures showed that the 1000 nearest neighbors were indeed in each case structurally closely related to the query molecule (see the Supporting Information smi files). As expected, each population of nearest neighbors formed well-separated groups in the PC2/PC3MQN map of PubChem (Figure 9C).

CONCLUSION

An overview of PubChem was constructed on the basis of 42 integer value descriptors of molecular structure, called MQNs, allowing an understanding of the diversity and a useful visualization of the PubChem database spanning from small drug-like fragments to large natural products, polypeptides, and oligonucleotides. The color-coded maps of the (PC2,PC3) plane facilitate the perception of the available structural diversity. Virtual screening performed by MQN similarity shows that compounds with similar bioactivities tend to cluster in MQN space. An MQN-annotated version of PubChem together with an application to automatically retrieve MQN nearest neighbors of any query molecule is available at www.gdb.unibe.ch. With this application, nearest neighbor searches can be completed within seconds on a

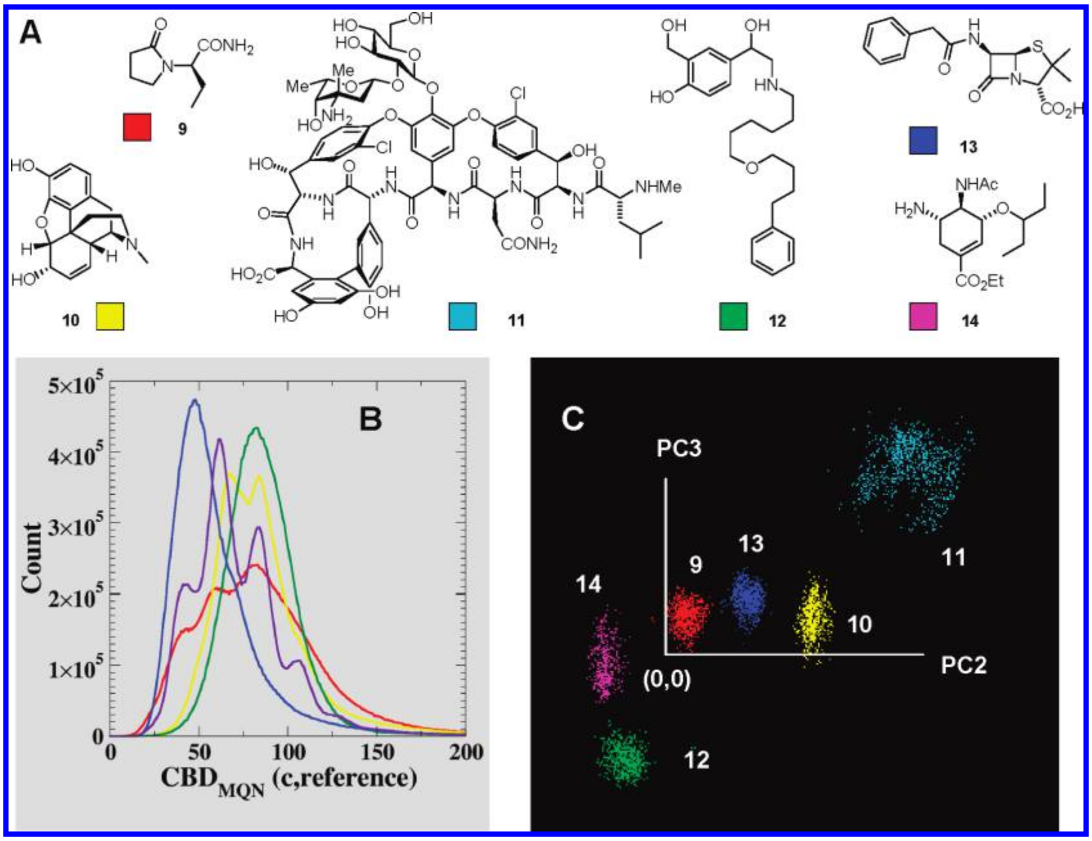


Figure 9. Nearest neighbor searches in PubChem by CBD_{MQN}. (A) Structural formulas of the drugs keppra (**9**), morphine (**10**), vancomycin (**11**), salmeterol (**12**), penicillin G (**13**), and tamiflu (**14**). (B) CBD_{MQN} histogram relative to the same drugs except vancomycin (see also data in Table 3). (C) Representation of the 1000 CBD_{MQN} nearest neighbors of the same drugs in the (PC2,PC3) plane. The nearest-neighbor search tool is available at www.gdb.unibe.ch.

single computer. The approach should greatly facilitate the use of chemical space as an inspiring resource for drug discovery.[31−35]

METHODS

Molecular Quantum Numbers (MQNs). MQNs were calculated using the previously reported source code (Supporting Information in ref 19). The source code was written in Java using JChem from Chemaxon, Ltd. as a starting library.

Principal Component Analysis (PCA). Principal component analysis is an in-house built source code written in Java using JSci (http://jsci.sourceforge.net/) as a starting library. The source code is based on the tutorial of Lindsay I. Smith (http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf).

Peptide Virtual Library. Each of the 20 amino acids was defined with SMILES of type C(=O)C(R)N, where R are SMILES for the different side chains, e.g., R = C for alanine and R = CCC(O)=O for glutamate. The generation was initialized by creating the SMILES for water ('O') and extending by adding the SMILES for a randomly chosen amino acid in a series of up to 80 selections, enumerating up to 100 000 different virtual peptides.

DNA Virtual Library. The four bases A, T, C, and G were defined using SMILES of type OP(=O)(O)OCC1OC(R)CC1O, where R is generally used for the different nucleobase, e.g., R = n2cnc3c(N)ncnc23 for the nucleobase adenine. The generation was initialized by creating the SMILES for water ('O') and randomly adding one of the four bases in 40 sequential steps, to obtain 40 000 different virtual oligonucleotides.

Graphene Virtual Library. The graphenes were created by successively removing one ring from the starting entry downloaded from PubChem, CID 16152993. The library was diversified by adding up to 20% methyl groups on the ring system, to obtain 146 759 different virtual compounds.

Diamondoid Virtual Library. A diamondoid was created with molecular formula $C_{126}H_{92}$. An initial library was generated by removing successively one ring from the ring system. The library was diversified by adding up to 20% methyl groups on the ring system, to obtain 106 637 different virtual compounds.

Alkanes Virtual Library. Alkanes were made starting from methane by adding one carbon atom at a random position on the existing alkane up to a size of 500 carbon atoms, enumerating up to 500 000 different virtual compounds.

Oligosaccharides Virtual Library. Using *β*-D-fructofuranose and α-D-glucopyranose, chains of up to 50 units were generated by creating 1,4-glycosidic linkages and then diversified by adding sulfate groups on position 6 and randomly substituting α-D-glucopyranose to *N*-acetyl-α-D-glucosamine, to obtain 31 044 different virtual compounds.

ACKNOWLEDGMENT

This work was supported financially by the University of Berne, the Swiss National Science Foundation, the Office Fédéral Suisse de l'Education et de la Science, and the COST program Angiokem. The authors thank Prof. Peter Willett for critical insights and helpful suggestions on using the MQNs as a similarity measure.

Supporting Information Available: SMILES of the 1000 MQN neighbors of compounds **9**−**14**. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) 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.

(2) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. Binding DB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.

(3) 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*, W623–633.

(4) Warr, W. A. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195–198.

(5) 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.

(6) 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.

(7) 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.

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

(9) 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.

(10) Rush, T. S., III; 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.

(11) 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.

(12) Wetzel, S.; Klein, K.; Renner, S.; Rauh, D.; Oprea, T. I.; Mutzel, P.; Waldmann, H. Interactive exploration of chemical space with Scaffold Hunter. *Nat. Chem. Biol.* **2009**, *5*, 581–583.

(13) 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.

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

(15) 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.

(16) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213.

(17) Schmuker, M.; Schneider, G. Processing and classification of chemical data inspired by insect olfaction. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20285–20289.

(18) Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P. Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol.* **2009**, *27*, 18–26.

(19) 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.

(20) Wang, S. G.; Schwarz, W. H. Icon of chemistry: the periodic system of chemical elements in the new century. *Angew. Chem., Int. Ed. Engl.* **2009**, *48*, 3404–3415.

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

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

(23) 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.

(24) Allen, M. J.; Tung, V. C.; Kaner, R. B. Honeycomb Carbon: A Review of Graphene. *Chem. Rev.* **2009**, *110*, 132–145.

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

(26) 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.

(27) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(28) 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.

(29) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.

(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*, 429–36.

(31) Reymond, J. L.; Van Deursen, R.; Blum, L. C.; Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **2010**, *1*, 30–38.

(32) Schneider, G.; Hartenfeller, M.; Reutlinger, M.; Tanrikulu, Y.; Proschak, E.; Schneider, P. Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol* **2009**, *27*, 18–26.

(33) Garcia-Delgado, N.; Bertrand, S.; Nguyen, K. T.; van Deursen, R.; Bertrand, D.; Reymond, J.-L. Exploring α7-Nicotinic Receptor Ligand Diversity by Scaffold Enumeration from the Chemical Universe Database GDB. *ACS Med. Chem. Lett.* **2010** [Online].

(34) Nguyen, K. T.; Syed, S.; Urwyler, S.; Bertrand, S.; Bertrand, D.; Reymond, J. L. Discovery of NMDA glycine site inhibitors from the chemical universe database GDB. *ChemMedChem* **2008**, *3*, 1520–1524.

(35) Luethi, E.; Nguyen, K. T.; Burzle, M.; Blum, L. C.; Suzuki, Y.; Hediger, M.; Reymond, J. L. Identification of Selective Norbornane-Type Aspartate Analogue Inhibitors of the Glutamate Transporter 1 (GLT-1) from the Chemical Universe Generated Database (GDB). *J. Med. Chem.* **2010**, *53*, 7236−7250.

CI100237Q