J Comput Aided Mol Des (2011) 25:649–662
DOI 10.1007/s10822-011-9437-x

Visualisation of the chemical space of fragments, lead-like and drug-like molecules in PubChem

Ruud van Deursen · Lorenz C. Blum · Jean-Louis Reymond

Received: 22 February 2011 / Accepted: 13 May 2011 / Published online: 27 May 2011

Abstract The 4.5 million organic molecules with up to 20 non-hydrogen atoms in PubChem were analyzed using the MQN-system, which consists in 42 integer value descriptors of molecular structure. The 42-dimensional MQN-space was visualised by principal component analysis and representation of the (PC1, PC2), (PC1, PC3) and (PC2, PC3) planes. The molecules were organized according to ring count (PC1, 38% of variance), the molecular size (PC2, 25% of variance), and the H-bond acceptor count (PC3, 12% of variance). Compounds following Lipinski's bioavailability, Oprea's lead-likeness and Congreve's fragment-likeness criteria formed separated groups in MQN-space visible in the (PC2, PC3) plane. MQN-similarity searches of the 4.5 million molecules (see the browser available at www.gdb.unibe.ch) gave significant enrichment factors for recovering groups of fragment-sized bioactive compounds related to ten different biological targets taken from Chembl, allowing lead-hopping relationships not seen with substructure fingerprint similarity searches. The diversity of different compound series was analyzed by MQN-distance histograms.

Keywords PubChem · Fragments · Chemical space · Virtual screening

R. van Deursen · L. C. Blum · J.-L. Reymond (✉)
Department of Chemistry and Biochemistry, Swiss National Center of Competence in Research, NCCR-TransCure, University of Berne, Freiestrasse 3, 3012 Berne, Switzerland
e-mail: jean-louis.reymond@ioc.unibe.ch

Introduction

In fragment-based drug discovery one starts by screening a few thousand fragment-sized molecules for weak bioactivity [1–8]. The fragment selection is critical since it should cover a relevant structural diversity. This selection would be facilitated if a simple overview of the known chemical space was available [9]. Herein we report an analysis and visualisation of the 4.5 million fragment-sized molecules currently reported in the public access database PubChem [10] using the MQN-system.

The MQN-system classifies organic molecules using 42 integer value descriptors of molecular structure including classical topological indexes such as atom and ring counts, and a few additional counts such as cyclic and acyclic unsaturations, atoms and bonds in fused rings, and electrostatic charges predicted for neutral pH (Table 1) [11]. We call these descriptors molecular quantum numbers (MQNs) in analogy to the atomic and principal quantum numbers classifying the elements in the periodic system [12]. The 42 MQNs form a 42-dimensional chemical space. Since the MQNs only have integer values, MQN-space is composed of "MQN-bins" to which molecules are assigned if they share the same MQN values. Such MQN-isomers may be compared to isotopes sharing the same atomic and principal quantum number in the periodic system of the elements.

A chemical space is created whenever molecules are described by a series of values collected in a vector or fingerprint, each value defining a separate dimension [13]. Among the various approaches reported sofar to represent chemical space [14–16], MQN-space has the advantage of simplicity. For instance chemical spaces have been constructed by assigning dimensions to descriptors such as eigenvalues of matrices constructed from primary

Table 1 The 42 Molecular quantum numbers (MQNs)

Atom counts (12)		*Topology counts*[b] *(17)*	
c	Carbon	asv	Acyclic monovalent nodes
f	Fluorine	adv	Acyclic divalent nodes
cl	Chlorine	atv	Acyclic trivalent nodes
br	Bromine	aqv	Acyclic tetravalent nodes
i	Iodine	cdv	Cyclic divalent nodes
s	Sulphur	ctv	Cyclic trivalent nodes
p	Phosphorous	cqv	Cyclic tetravalent nodes
an	Acyclic nitrogen	r3	3-Membered rings
cn	Cyclic nitrogen	r4	4-Membered rings
ao	Acyclic oxygen	r5	5-Membered rings
co	Cyclic oxygen	r6	6-Membered rings
hac	Heavy atom count	r7	7-Membered rings
		r8	8-Membered rings
Polarity counts[a] *(6)*		r9	9-Membered rings
hbam	H-bond acceptor sites	rg10	≥10 Membered rings
hba	H-bond acceptor atoms	afr	Atoms shared by fused rings
hbdm	H-bond donor sites	bfr	Bonds shared by fused rings
hbd	H-bond donor atoms		
neg	Negative charges		
pos	Positive charges		
Bond counts (7)			
asb	Acyclic single bonds		
adb	Acyclic double bonds		
atb	Acyclic triple bonds		
csb	Cyclic single bonds		
cdb	Cyclic double bonds		
ctb	Cyclic triple bonds		
rbc	Rotatable bond count		

[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

descriptors [13, 17, 18], or more directly to descriptors of molecular structure and calculated physicochemical properties [19, 20], or by using various types of fingerprints [21]. These descriptor sets are obtained from the molecular structures through complex and partly machine-controlled calculations, which renders the immediate understanding of the resulting chemical space inaccessible to non-specialists. By contrast the MQN-dimensions correspond to values which can be counted in a structural formula by anyone having basic training in organic chemistry. In this manner, the link between the strutural formula of a compound and its position in MQN-space is self-explanatory. Furthermore, MQN-calculations are extremely fast and therefore particularly well suited to classify very large databases such as PubChem, which contains over 20 million structures, or the chemical universe database GDB-13 which contains 977 million structures as shown in the following paper in this issue [22–25]. The MQN-system is also relevant for medicinal chemistry, as illustrated by the fact that similarity searches in MQN-space enrich bioactives in the DUD database [26] from the entire PubChem with efficiencies comparable to that of substructure fingerprints, with the added benefit that lead-hopping relationships between actives are allowed [27].

The following paper details the MQN-analysis of the fragment subset of PubChem comprising all molecules with up to 20 non-hydrogen atoms (hac $\leq$ 20), which corresponds to the broadest definition of fragments [28]. MQN-maps produced by principal component analysis (PCA) of the MQN data and representation of the (PC1, PC2), (PC1, PC3) and (PC2, PC3) planes illustrate the structural diversity of PubChem fragments, and how the subsets of molecules fulfilling the increasingly restrictive criteria of Lipinski's "rule of five", Oprea's "lead-likeness" and Congreve's"rule of three" are distributed [29]. We show that MQN-similarity searches allow significant enrichments for recovering groups of fragment-sized bioactive compounds related to ten different biological targets taken from Chembl, while allowing lead-hopping relationships not seen with substructure fingerprint similarity

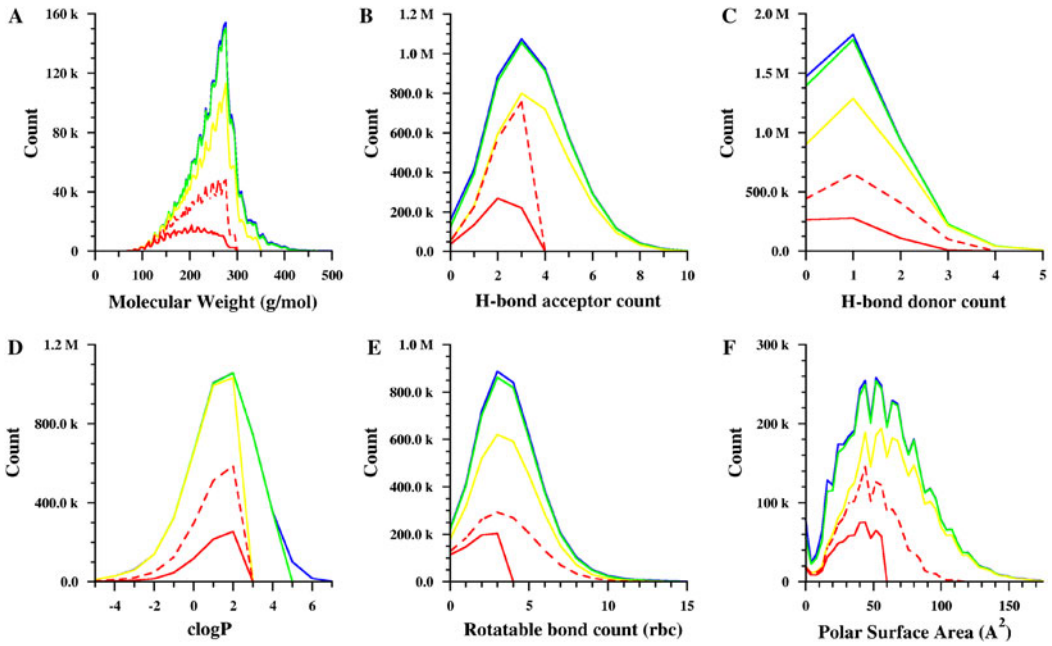


Fig. 1 Property histograms of Pubchem subsets: hac $\leq$ 20 (*blue line*), Ro5 (*green line*), LL (*yellow line*), Ro3 (*dashed red line*) and Ro3+ (*red line*). **a** Molecular weight. **b** H-bond acceptor atom count. **c**. H-bond donor atom count. **d** Calculated water-octanol partition coefficient. **e** Rotatable bond count. **f** Polar surface area

searches. Finally, the MQN-distance histograms of compound series in MQN-space are proposed as a measure for their structural diversity.

Results and discussion

The fragment subset of PubChem

The public database PubChem archives the molecular structures and bioassay data within the National Institute of Health (NIH) Roadmap for Medical Research Initiative. Among the 49,792,078 entries (December 20, 2010), 4,512,612 compounds had up to 20 non-hydrogen atoms (hac $\leq$ 20), which corresponds to the broadest definition of fragments [28]. From these compounds, 4,383,087 (97%) also fulfilled Lipinski's "rule of five" (Ro5) criteria for bioavailability [30], 3,243,980 (72%) fulfilled Oprea's criteria for lead-likeness (LL) [31], 1,604,411 (36%) fulfilled Congreve's "rule of 3" for fragment-likeness (Ro3) [32], and 616,677 (14%) had further restrictions in rotatable bonds (rbc $\leq$ 3) and polar surface area (PSA $\leq$ 60 Å^2) (Ro3+). The distribution of the compounds according to the molecular properties is shown as histograms in Fig. 1. MQNs were computed for these compounds as previously described, [11, 27] giving a total of 2,357,293 unique MQN-combinations, or "MQN-bins", of which 1,602,337 are occupied by a single molecule. The most occupied MQN-bin contains 270 compounds (Table 2; Fig. 2).

Visualisation

PCA of the MQN data showed that the first three principal components covered 86% of the variance (Fig. 3). The first principal component PC1 (38% of variance) corresponded mostly to the fraction of cyclic atoms per molecule, with strong positive loadings in cyclic single bonds (csb) and cyclic divalent atoms (cdv), and strong negative loadings in acyclic single bonds (asb). PC2 (25% of variance) represented molecular size and flexibility, with strong positive loadings in heavy atom count (hac), carbon count (c), acyclic single bonds (asb) and rotatable bond count (rbc). Finally PC3 (12% of variance) represented polarity, with strong positive loadings in hydrogen bond acceptors (hbam and hba) and negative loadings in carbon count (c).

Color-coded maps were constructed from the MQN data in form of the projections in the (PC1, PC2), (PC1, PC3) and (PC2, PC3) planes (Fig. 4). The occupancy maps showed that the fragments were distributed in a series of parallel islands in the (PC1, PC2) and (PC1, PC3) planes, corresponding to an increasing number of rings per molecule in accordance with the loadings of PC1. Each of these islands was occupied by smaller, rigid molecules at low PC2 values, and larger, more flexible molecules at high PC2 values. The highest occupancy corresponded to

Table 2 MQN-statistics for PubChem fragments and subsets

Subset[a]	cpds	MQN-bins[b]	MQN-binswith 1 cpd	MQN-bin occupancy	
				Max.	Average
All	4,512,570	2,357,293	1,602,337	270	1.91 ± 2.77
Ro5	4,383,087	2,296,292	1,561,555	270	1.96 ± 2.77
LL	3,243,980	1,797,684	1,246,420	269	1.91 ± 2.75
Ro3	1,604,441	819,422	541,952	269	1.80 ± 2.40
Ro3+	661,677	362,670	245,247	112	1.82 ± 2.15

[a] All: PubChem hac≤ 20, Ro5: Lipinski's bioavailable, LL: Oprea's lead-like, Ro3: Congreve's fragment-like, Ro3+: Congreve's fragment-like including rbc ≤ 3 and PSA ≤ 60 Å^2

[b] Number of unique combinations in the 42-dimensional MQN-system. See also legend of Fig. 3 for exhaustive listing of the MQNs

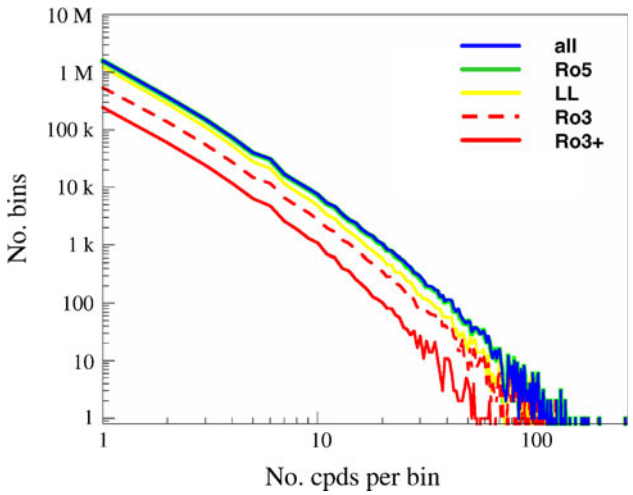


Fig. 2 Distribution of MQN-bins as a function of bin-occupancy for the PubChem fragments. See Table 1 for subset definition

Fig. 4 Projections of the MQN-space of PubChem fragments in the (PC1, PC2) plane (*left*), (PC1, PC3) plane (*center*), and (PC2, PC3) plane (*right*). The maps are color-coded for the indicated values and subsets. The projections were obtained by hasing the surface on 1,000 × 800 pixels ((PC1,PC2) and (PC1,PC3)) and 800 × 800 pixels (PC2,PC3). **a** Each pixel is colored using the hue as indication for the occupancy in a continuous range from *blue* (1) to *red* (200). **b**–**d** Pixels, representing a set of molecules, are colored using the hue for the average value and saturation to grey for standard deviation (50% saturation = 10% standard deviation) for the corresponding value. **e** The pixels are colored using *red*, *yellow*, *green* and *violet* for each of the fragment types. Saturation to grey was used to indicate the purity for the dominant fragment type

monocyclic and bicyclic molecules with 12–18 heavy atoms and 1–3 H-bond donor atoms in a region mostly occupied by the lead-like compounds. The Ro5, lead-like and Ro3 compounds were most clearly separated in the (PC2, PC3) projection spreading compounds according to

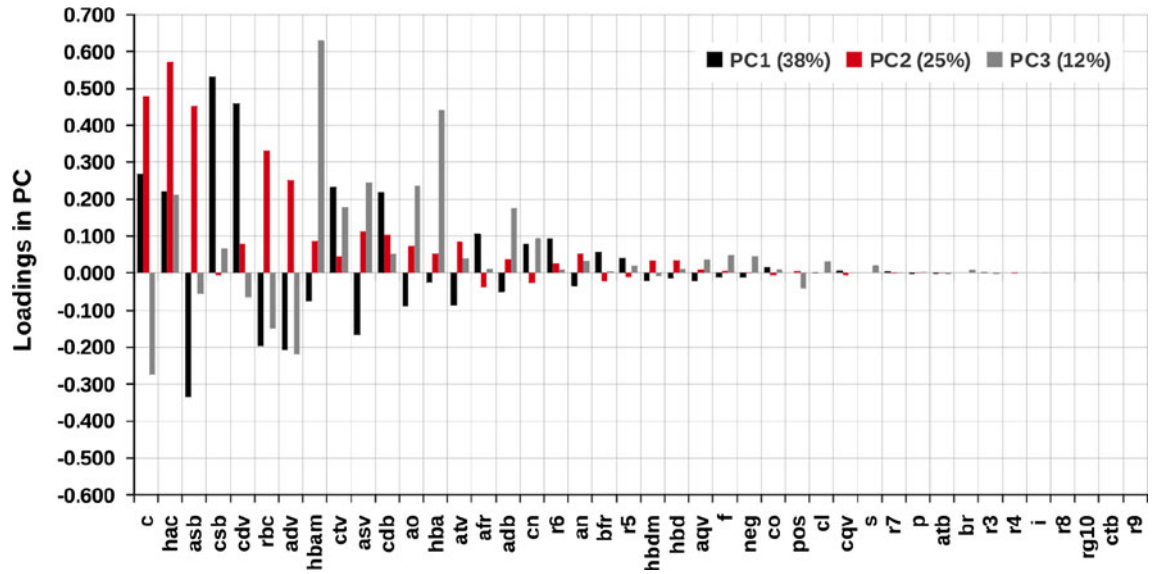


Fig. 3 Loading plots in the PCA of the MQN data of PubChem fragments. The 42 MQNs are defined in Table 1

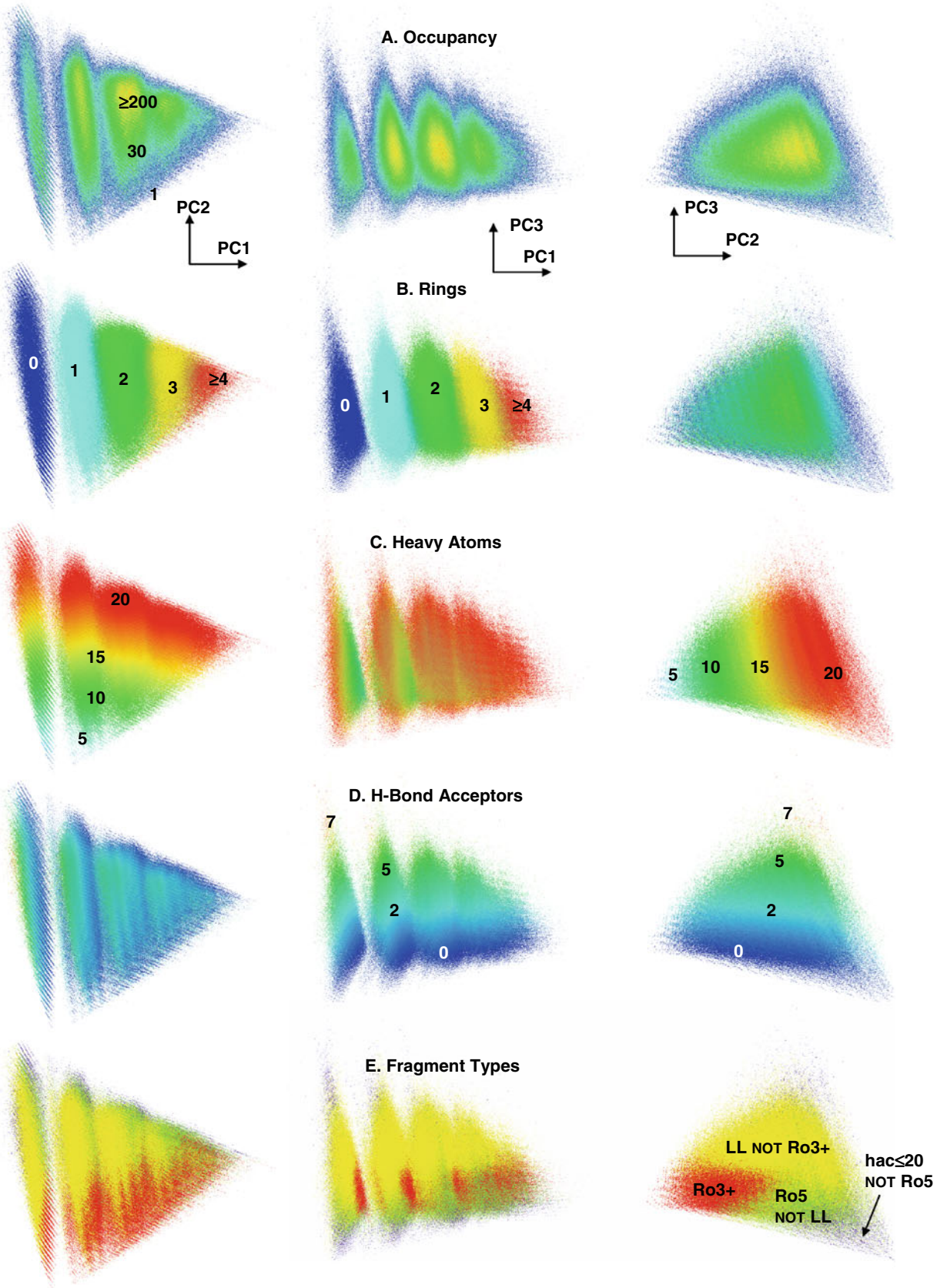

A. Occupancy
≥200
30
1
PC2
PC1
PC3
PC1
PC3
PC2
B. Rings
0
1
2
3
≥4
0
1
2
3
≥4
C. Heavy Atoms
20
15
10
5
5
10
15
20
D. H-Bond Acceptors
7
5
2
0
7
5
2
0
E. Fragment Types
LL NOT Ro3+
Ro3+
Ro5
NOT LL
hac≤20
NOT Ro5

Table 3 Drug targets from Chembl used for virtual screening

Target	Code[a]	Bioactives[b]	Drug	Code[a]	Hac
Cyclooxygenase-2 (COX2)	230	351	Diclofenac	139	19
Voltage-Gated Ca^{2+} channel $\alpha 2/\delta$ subunit 1	1919	43	Gabapentin	940	12
Beta-1 adrenergic receptor	213	84	Metoprolol	13	19
Na+ channel protein type II αsubunit	3399	25	Oxcarbazepine	1068	19
Dopamine D1 Receptor	2056	48	Ropinirole	589	19
Na-dependent serotonin transporter	228	562	Sertraline	809	20
Na-dependent dopamine transporter	338	392	Sertraline	809	20
Tyrosine-protein kinase ABL1	1862	35	Temozolomide	810	14
β-glucocerebrosidase	2179	25	Voglibose	476960	18
Farnesyl diphosphate synthase	1782	91	Zolderonic acid	924	16

[a] Chembl code listed in the Chembl database (www.ebi.ac.uk/chembl/)

[b] Compounds listed with IC_{50} or $EC_{50} \leq 10$ μM for the indicated targets were considered as bioactives. The bioactives include compounds with polypharmacology

size and polarity, in agreement with the criteria employed in these increasingly restrictive subsets.

Virtual screening in MQN-space

If a chemical space is relevant to medicinal chemistry, nearest neighbours in that chemical space should share similar bioactivities. This hypothesis can be tested by measuring the recovery from a database of a family of molecules active on the same target by sorting the database by increasing distance to a reference compound for that target. The recovery of x% of the actives after searching y% of the distance-sorted database should be multiplied by an enrichment factor $EF_y = x\%/y\%$ significantly larger than one, with maximum possible values $EF_{0.1} = 1{,}000$ and $EF_1 = 100$. The distance measure between molecules may be the euclidean distance satisfying the triangular inequality, [13] or a variety of other similarity measures [33]. We showed previously that high EFs are obtained for recovering bioactives of the DUD database [26] from the entire PubChem database using either the city-block distance CBD_{MQN} or the Tanimoto similarity coefficient T_{MQN} as distance measure. The EFs derived the MQN-values were comparable to those obtained using the corresponding similarity measures CBD_{SF} and T_{SF} derived from a daylight-type 1,024-bit substructure fingerprint (SF), however the MQN-similarity searches allowed for interesting lead-hopping enrichments not seen using SF-similarity [27].

To test if similar EFs might be possible within the more restricted fragment subset of PubChem, EFs were computed for ten groups of fragment-sized molecules (hac ≤ 20) collected from the Chembl database with reported activity on ten drug targets [34]. The value IC_{50} or $EC_{50} \leq 10$ μM was chosen as activity threshold for fragment-type activity [28]. The ten targets were associated with nine drugs belonging to the 200 top-selling drugs in 2010. The entire PubChem fragment subset and its Ro5, LL, Ro3 and Ro3+ subsets as discussed above were sorted by similarity to the compound closest to the mathematical center of gravity in MQN-space (cg_{MQN}) for each group. Sorting was performed with the four similarity measures CBD_{MQN}, T_{MQN}, CBD_{SF} and T_{SF} (Tables 3, 4; Fig. 5).

The best EFs for recovering actives from the corresponding PubChem subset were obtained from SF-similarity searching, reflecting the fact that many compound series were developed by variations on conserved scaffolds. The enrichments from MQN-similarity searching were however also very significant, indicating that compounds with similar bioactivities tend to be relatively close to one another in MQN-space at the size of fragments. Remarkably, the top-scoring analogs found by MQN-similarity searching included many analogs with very low substructure similarity as measured by SF. This confirms our previous observation with the DUD-actives that searching by MQN-similarity allows for lead-hopping relationships. On the other hand analogs picked by SF similarity also showed a good level of MQN similarity (Fig. 5, right scatter plots, Fig. 6) [27].

MQN-distance histograms as diversity measure

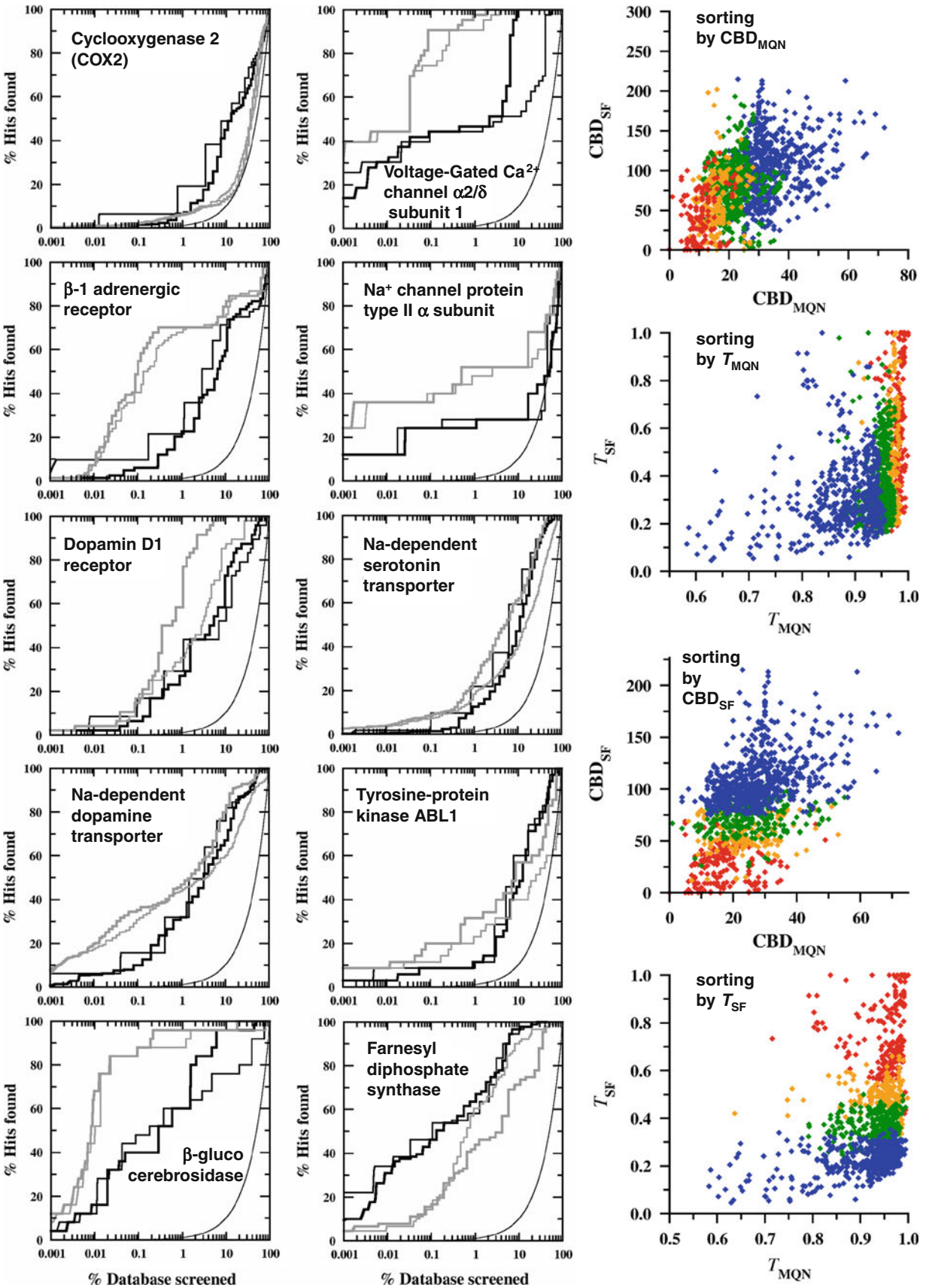
Estimating the diversity of a compound collection in the perspective of its potential value for bioactivity is a difficult problem to which various solutions have been proposed [16, 35, 36], in particular extended-connectivity fingerprints (ECFP) [37]. The MQN-system offers an attractive alternative to appreciate structural diversity, even if quite simple in comparison of the existing methods. One may consider any compound series as a cloud in MQN-space around its mathematical centre of gravity (cg_{MQN}). The diversity then can be characterized by a histogram of

Table 4 Enrichment of bioactive compounds from the PubChem fragments[a]

Drug target	Subset[b]	Act.	CBD_{MQN}		T_{MQN}		CBD_{SF}		T_{SF}	
			$EF_{0.1}$	EF_1	$EF_{0.1}$	EF_1	$EF_{0.1}$	EF_1	$EF_{0.1}$	EF_1
Cyclooxygenase-2 (COX2)	All	351	12	13	16	22	71	18	84	23
	Ro5	309	9	11	9	21	60	19	71	22
	LL	137	15	8	38	11	30	13	34	14
	Ro3	65	12	3	41	11	25	8	25	10
	Ro3+	39	89	23	149	37	79	13	79	18
Voltage-Gated Ca^{2+} channel $\alpha 2/\delta$ subunit 1	All	43	442	47	442	44	744	91	907	98
	Ro5	43	442	47	442	44	744	91	907	98
	LL	43	419	47	395	44	744	91	814	95
	Ro3	43	419	44	395	44	721	91	721	95
	Ro3+	6	1,000	100	833	100	833	100	1,000	100
β-1 adrenergic receptor	All	84	60	20	95	21	405	68	488	70
	Ro5	84	60	20	95	21	405	68	488	70
	LL	82	49	21	98	22	402	68	488	70
	Ro3	61	164	16	164	16	246	62	262	61
	Ro3+	13	615	62	538	62	615	62	615	62
Na+ channel protein type II α subunit	All	25	240	24	240	28	400	48	360	52
	Ro5	23	261	26	261	30	435	52	391	57
	LL	13	462	46	462	54	769	92	692	92
	Ro3	13	462	46	462	54	692	85	692	92
	Ro3+	9	333	67	667	67	1,000	100	1,000	100
Dopamine D1 Receptor	All	48	63	27	167	29	167	35	146	58
	Ro5	48	63	27	167	29	167	35	146	58
	LL	39	77	31	103	31	231	36	179	82
	Ro3	38	53	26	105	32	79	34	132	58
	Ro3+	6	333	33	333	33	333	83	833	100
Na-dependent serotonin transporter	All	562	14	7	63	19	23	6	20	7
	Ro5	549	16	7	71	22	26	6	23	7
	LL	263	15	1	15	4	22	4	22	9
	Ro3	244	31	3	31	8	31	6	31	5
	Ro3+	101	26	5	26	5	26	8	26	3
Na-dependent dopamine transporter	All	392	99	32	156	32	306	44	352	47
	Ro5	388	101	32	157	32	309	44	356	47
	LL	209	72	25	120	27	268	52	359	55
	Ro3	199	15	20	111	24	256	45	322	54
	Ro3+	127	32	9	94	21	55	32	55	57
Tyrosine-protein kinase ABL1	All	35	86	11	86	9	114	23	200	31
	Ro5	35	86	11	86	9	114	23	200	31
	LL	22	91	14	91	9	136	27	227	36
	Ro3	5	400	40	400	40	600	60	600	60
	Ro3+	4	500	50	500	50	500	50	500	50
β-Glucocerebrosidase	All	25	400	60	480	60	880	88	880	96
	Ro5	17	412	53	412	41	824	82	824	94
	LL	17	412	53	353	41	824	82	824	94
Farnesyl diphosphate synthase	All	91	462	64	462	58	143	60	165	44
	Ro5	91	462	64	462	58	143	60	154	42
	LL	91	429	64	462	58	143	60	154	42

[a] Enrichment factors (EF) for recovery of the indicated number of actives (act.) from the indicated subset using various similarity measures relative to the compound closest to the respective MQN-center of gravity (cg_{MQN}) as reference bioactive compound

[b] Subsets from which the actives are recovered: all: PubChem hac≤20, 4,512,612 cpds, Ro5: Lipinski's bioavailable, 4,383,087 cpds, LL: Oprea's lead-like, 3,243,980 cpds, Ro3: Congreve's fragment-like, 1,604,411 cpds, Ro3+: Congreve's fragment-like including rbc≤3 and PSA ≤ 60 $Å^2$, 616,677 cpds. See also text and Fig. 5


Cyclooxygenase 2 (COX2)
Voltage-Gated Ca2+ channel α2/δ subunit 1
β-1 adrenergic receptor
Na+ channel protein type II α subunit
Dopamin D1 receptor
Na-dependent serotonin transporter
Na-dependent dopamine transporter
Tyrosine-protein kinase ABL1
β-gluco cerebrosidase
Farnesyl diphosphate synthase
% Hits found
% Database screened
sorting by CBDMQN
sorting by TMQN
sorting by CBDSF
sorting by TSF
CBDSF
CBDMQN
TSF
TMQN

◀ **Fig. 5** Enrichment curves for recovering drugs and their bioactive analogs in Table 3 from the Pubchem hac ≤ 20 subset (all, 4,512,612 cpds) by sorting relative to the bioactive compound closest to cg_{MQN} using CBD_{MQN} (*thick black line*), T_{MQN} (*thin black line*), CBD_{SF} (*thick grey line*), or T_{SF} (*thin grey line*) as similarity measure. The scatter plots at right show the pairs of similarity values for the bioactives of all ten targets recovered by MQN (upper two scatter plots) and SF similarity (lower two scatter plots) after screening 0.1% (red), 1% (*orange*), 10% (*green*) and 100% (*blue*) of the sorted PubChem hac ≤ 20 subset

CBD_{MQN} to the cg_{MQN}. Large average CBD_{MQN} from the cg_{MQN} indicate diversity, while shorter average distances indicate similarity of the compound series in MQN-space.

In this analysis, the PubChem fragments and the Ro3 and Ro3+ subsets showed a gaussian type distance distribution with most compounds located at $15 < CBD_{MQN} < 45$ from their cg_{MQN} (Fig. 7a). Two series of 20k compounds following the restricted Congreve's rule taken randomly from PubChem compounds (Ro3+ cpds) or MQN-bins (Ro3+ bins) had distance histograms essentially matching that of the entire Ro3+ PubChem subset. A similarly sized commercial fragment library selected for diversity had somewhat shorter CBD_{MQN}, with most compounds located at $12 < CBD_{MQN} < 35$, and almost no compounds at $CBD_{MQN} > 50$ from their cg_{MQN} (Fig. 7b). The ten bioactive series from Chembl were focused around their respective cg_{MQN} with a sharp peak at $CBD_{MQN} = 21$ (Fig. 7c, dashed line). The narrower distribution of bioactives around their cg_{MQN} was in line with the virtual screening results above, illustrating further that sets of compounds active on the same target are relatively close to one another in MQN-space, and therefore somewhat less MQN-diverse compared to random series.

The MQN-diversity of known drugs was analyzed next. From the 6,716 drugs registered in the DrugBank, 472 followed the Ro3+ fragment-likeness rule, an additional 564 followed Ro3, and an additional 1,550 were within hac ≤ 20. Most Ro3 + drugs were located at $15 < CBD_{MQN} < 45$ from their cg_{MQN}, most Ro3 drugs were located at $25 < CBD_{MQN} < 50$ from their cg_{MQN}, and most hac ≤ 20 drugs were located at $25 < CBD_{MQN} < 55$ from their cg_{MQN} (Fig. 7c). Thus, known drugs covered a range of MQN-space comparable to the entire PubChem fragments.

To estimate how closely the three 20k fragment series introduced above approached the MQN-space of the known drugs, the shortest CBD_{MQN} of each drug to each complete series was calculated. In the resulting shortest CBD_{MQN} histograms, high values indicated drugs whose MQN-space environment was not covered by a 20k fragment series. The analysis showed that each 20 k series contained compounds at short CBD_{MQN} to all Ro3+ drugs, with most shortest $CBD_{MQN} < 15$ for the commercial fragment library and most $CBD_{MQN} < 10$ for the two random series (red line in Fig. 7d–f). The three 20k series also had compounds within $CBD_{MQN} < 40$ of the vast majority of Ro3 drugs (blue line in Fig. 7d–f) and hac ≤ 20 drugs (black line in Fig. 7d–f), an MQN-distance which is compatible with sharing a similar bioactivity.

The distribution of compounds in MQN-space was further characterized by histograms of exhaustive pairwise CBD_{MQN}, which can be computed in reasonable time for small series up to a few thousand compounds. Pairwise CBD_{MQN} covered all values up to 127 for hac ≤ 20 drugs and up to 90 for Ro3 and Ro3+ drugs, showing that known drugs covered a broad range of MQN-space (Fig. 7g). The value range for the three 20k series was comparable to that of the Ro3+ drugs, again with a somewhat narrower range for the commercial series (Fig. 7h). On the other hand the pairwise CBD_{MQN} within the ten Chembl bioactive series were significantly shorter and peaked at $CBD_{MQN} = 25$–30, with almost no values above 60, illustrating once more that compounds of similar activity are relatively close to one another in MQN-space (Fig. 7j). However the scatter of pairwise CBD_{MQN} was relatively large, as could be expected since CBD_{MQN} cannot be the only determinant of biological activity.

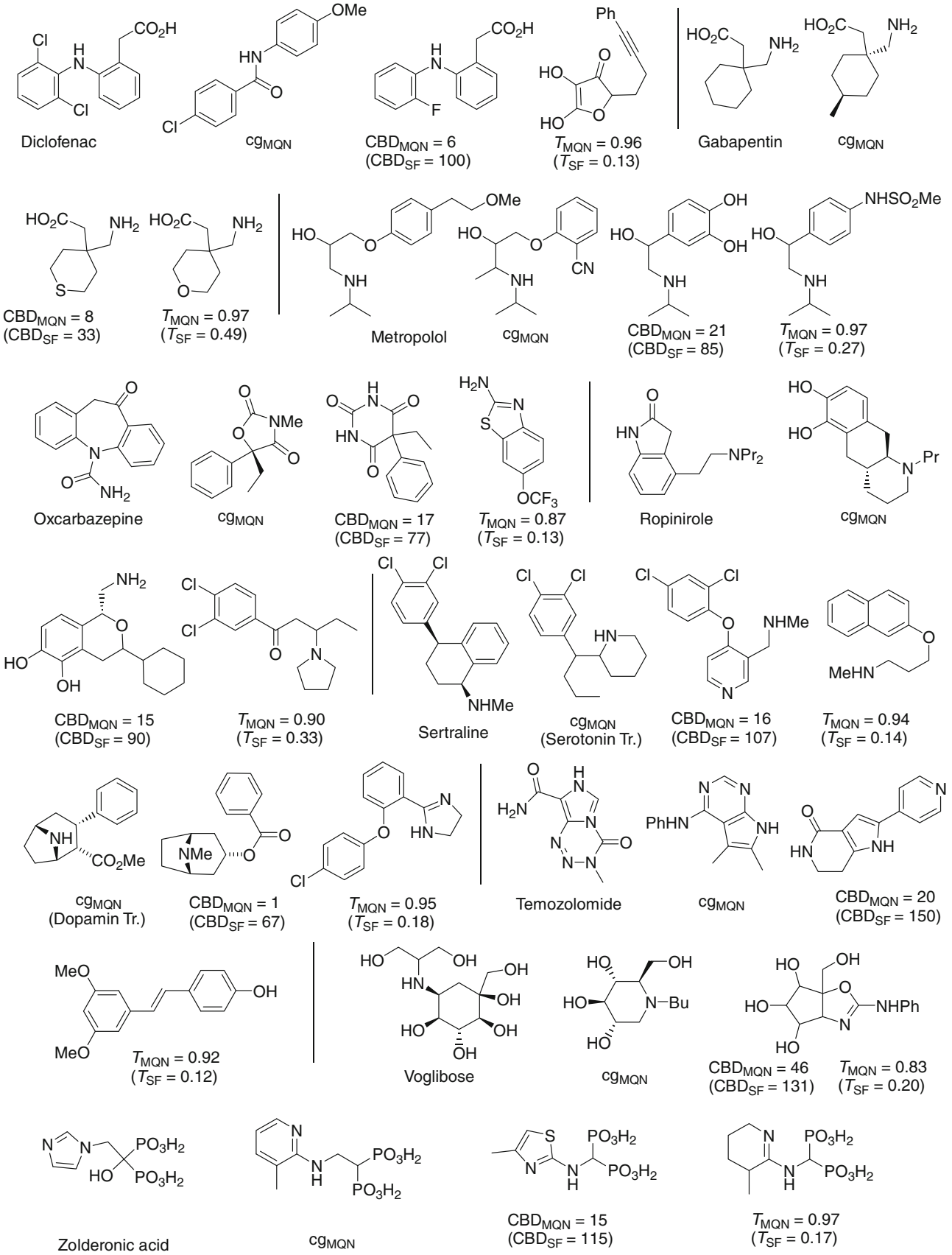
Conclusion

The 4.5 million fragment-sized molecules in PubChem were analyzed using the MQN-system, revealing the spread of structural diversity across compounds with various number of ring atoms, size, and polarity. Pubchem appeared to be particularly rich in monocyclic and bicyclic molecules of 15–18 heavy atoms with 1–3 H-bond donor atoms, corresponding to lead-like structures according to Oprea. Compounds active on the same target were relatively close to one another in MQN-space. This was illustrated by the significant enrichments when recovering bioactives from ten bioactivity classes from Chembl by MQN-similarity searches, and by the narrower distance histogram of these bioactivity classes for compared to randomly selected compounds. The fragments of PubChem and their different subsets can be explored by CBD_{MQN} similarity searching using a PubChem browser application available at www.gdb.unibe.ch.

Methods

Pubchem

Pubchem was downloaded on December 20, 2010. The library was reduced by allowing a maximum of 20 non-hydrogen atoms (hac ≤ 20). The library was additionally restricted by only keeping molecules with the elements H, B, C, N, O, F, P, S, Cl, Br and I, while conserving at least


Diclofenac
cg$_{MQN}$
CBD$_{MQN}$ = 6
(CBD$_{SF}$ = 100)
T_{MQN} = 0.96
(T_{SF} = 0.13)
Gabapentin
cg$_{MQN}$
CBD$_{MQN}$ = 8
(CBD$_{SF}$ = 33)
T_{MQN} = 0.97
(T_{SF} = 0.49)
Metropolol
cg$_{MQN}$
CBD$_{MQN}$ = 21
(CBD$_{SF}$ = 85)
T_{MQN} = 0.97
(T_{SF} = 0.27)
Oxcarbazepine
cg$_{MQN}$
CBD$_{MQN}$ = 17
(CBD$_{SF}$ = 77)
T_{MQN} = 0.87
(T_{SF} = 0.13)
Ropinirole
cg$_{MQN}$
CBD$_{MQN}$ = 15
(CBD$_{SF}$ = 90)
T_{MQN} = 0.90
(T_{SF} = 0.33)
Sertraline
cg$_{MQN}$
(Serotonin Tr.)
CBD$_{MQN}$ = 16
(CBD$_{SF}$ = 107)
T_{MQN} = 0.94
(T_{SF} = 0.14)
cg$_{MQN}$
(Dopamin Tr.)
CBD$_{MQN}$ = 1
(CBD$_{SF}$ = 67)
T_{MQN} = 0.95
(T_{SF} = 0.18)
Temozolomide
cg$_{MQN}$
CBD$_{MQN}$ = 20
(CBD$_{SF}$ = 150)
T_{MQN} = 0.92
(T_{SF} = 0.12)
Voglibose
cg$_{MQN}$
CBD$_{MQN}$ = 46
(CBD$_{SF}$ = 131)
T_{MQN} = 0.83
(T_{SF} = 0.20)
Zolderonic acid
cg$_{MQN}$
CBD$_{MQN}$ = 15
(CBD$_{SF}$ = 115)
T_{MQN} = 0.97
(T_{SF} = 0.17)

◀ **Fig. 6** Structural formulae of drugs and analogs. cg_{MQN} indicates the compound that is closest to the mathematical centre of gravity in the 42-dimensional MQN-system for each of the corresponding Chembl bioactive series (see Table 3). The compound following each cg_{MQN} are lead-hopping examples of analogs that are similar by MQN similarity measures (CBD_{MQN} or T_{MQN}) and dissimilar by substructure fingerprint similarity measures (CBD_{SF} or T_{SF})

one carbon in each molecule. Fragments from entries with multiple molecules were treated as different fragments. Positive and negative charges were neutralised, whenever possible. Finally the library was cleaned to contain unique fragments only. Final library size is 4,512,612 entries.

MQNs

MQNs were calculated using the previously reported calculator source code (Supporting Information in Ref. [11]) written in Java using the JChem library from Chemaxon, Ltd. Prior to MQN-calculation, the ionization state of each structure was adjusted to pH 7.4 using the JChem API.

Principal component analysis (PCA)

PCA was programmed in Java JDK 1.6 using the JSci science API. PCA was following the steps as described in the turotial of Lindsay I Smith (www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf).

Chembl bioactives

Bioactives for each of the ten targets (Table 2) were downloaded from the Chembl website (www.ebi.ac.uk/chembl). The bioactives were restricted to a maximum of 20 non-hydrogen atoms and keeping only entries with the elements H, B, C, N, O, F, P, S, Cl, Br and I, while conserving at least one carbon in each molecule.

Drugbank fragment library

Drugbank was downloaded on February 1, 2011 from www.drugbank.ca/system/downloads/. The original library of 6,716 molecules was reduced by allowing a maximum size of 20 non-hydrogen atoms (hac $\leq$ 20). The library was further restricted by allowing only atoms of type H, B, C, N, O, F, P, S, Cl, Br and I, while conserving at least one carbon in each molecule. The final library contained 2,586 molecules.

Subsets

All applied subsets filters were programmed in Java using JChem 5.2.6 of Chemaxon Ltd. (www.chemaxon.com). Lipinski "rule of 5" (Ro5) had following cutoff values: MW $\leq$ 500, clogP $\leq$ 5, H-bond acceptors $\leq$ 10 and H-bond donors $\leq$ 5. Cutoff values for lead-like filter (LL): MW $\leq$ 350, clogP $\leq$ 3, H-bond acceptors $\leq$ 10 and H-bond donors $\leq$ 5. Cutoff values for Congreve "rule of 3" (Ro3): MW $\leq$ 300, H-bond donors $\leq$ 3 and H-bond acceptors $\leq$ 3. More restricted "rule of 3" (Ro3+) used the following additional criteria to the Ro3: rotatable bond count (rbc) $\leq$ 3 and PSA $\leq$ 60 Å^2.

City-block-distance

An object A is defined by means of vector X_A of n attributes such that $X_A = \{x_{1A}, x_{2A}, \ldots, x_{jA}, \ldots, x_{nA}\}$. For two objects A and B, desribed by means of vector X_A and X_B with n elements, the city-block-distance ($CBD_{A,B}$), also known as Hamming distance or Manhattan distance, is defined as the sum of the absolute difference for each value [38] j:

$$CBD_{A,B} = \sum_{j=1}^{n} \left| x_{jA} - x_{jB} \right| \quad (1)$$

Tanimoto similarity coefficient

The Tanimoto similarity coefficient ($T_{A,B}$), also known as Jaccard coefficient, for two objects A and B, represented by means of vector X_A and X_B with length n and attributes j, is calculated using the following formula [38]:

$$T_{A,B} = \frac{\sum_{j=1}^{n} x_{jA} \cdot x_{jB}}{\sum_{j=1}^{n} \left(x_{jA}\right)^2 + \sum_{j=1}^{n} \left(x_{jB}\right)^2 - \sum_{j=1}^{n} x_{jA} \cdot x_{jB}} \quad (2)$$

Center of gravity

The center of gravity (cg_{MQN}) for a dataset was calculated down the column to produce a new 42-dimensional MQN, composed of the mean value for each descriptor.

Enrichments

For enrichment each set of bioactives was diluted into the corresponding subset of Pubchem. The molecule closest to the mathematical center of gravity in the 42-dimensional MQN-space (cg_{MQN}) was chosen as bait molecule for the distance and similarity calculations using either molecular quantum numbers (MQN) or a Daylight-type 1024 substructure fingerprint (SF). The list of molecules was then sorted according to the shortest CBD_{MQN} or CBD_{SF} and highest T_{MQN} or T_{SF}.

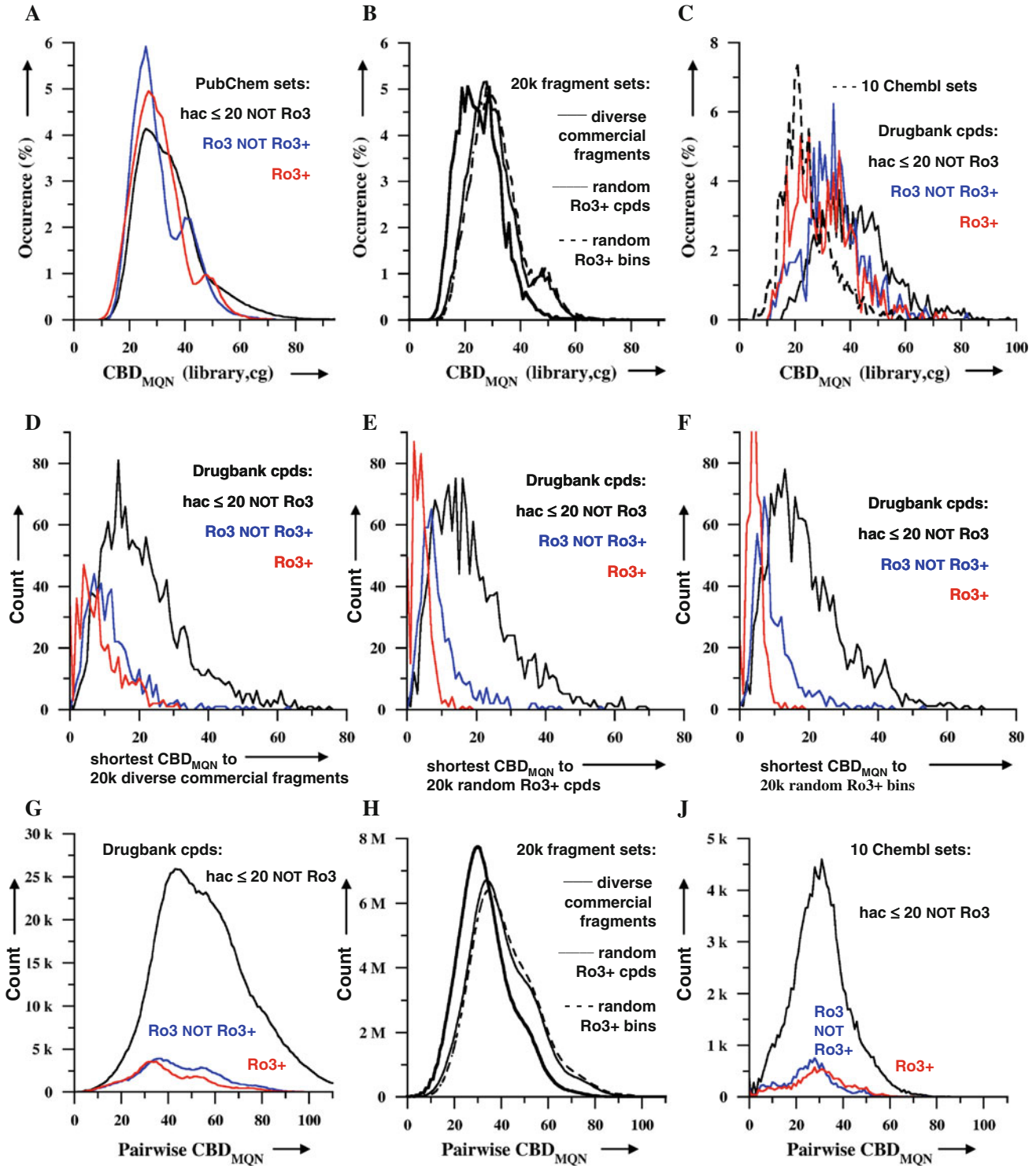


Fig. 7 a–e CBD_{MQN}-distance histograms. Follow main text for a detailed discussion of each pannel. "20 k random Ro3+ cpds"=20,000 cpds taken randomly from the Ro3+ subset of PubChem, "20 k random Ro3+ bins"=20,000 cpds taken from 20,000 bins choosen at random from all MQN-bins occupied by the Ro3+ subset of PubChem. "10 Chembl sets"=all compounds in the 10 bioactive series of Table 3 relative to their respective cg_{MQN}. Drugbank compounds: there are 1,550 drugs in DrugBank with hac ≤ 20 NOT Ro3, 564 drugs with Ro3 NOT Ro3+, and 472 drugs following Ro3+. In the Chembl sets hac ≤ 20, there are 988 bioactives from hac ≤ 20 NOT Ro3, 363 bioactives from Ro3 NOT Ro3+ and 305 bioactives from Ro3+. In Fig. 7j, pairwise distances were only measured within sets of actives for the same target. In Figure **g–j**, identity pairs are excluded

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

Glossary

CBD_{MQN}	City-block distance in MQN-space. Values are positive integers with small numbers indicating high similarity
CBD_{SF}	City-block distance in substructure-fingerprint space. Values are positive integers with small numbers indicating high similarity
cg_{MQN}	Center of gravity of a compound series in MQN-space. 42-line vector composed of the 42 average values of the MQNs calculated across the compound series
LL	Subset of PubChem hac < 20 following Oprea's criteria for lead-likeness
MQN	Molecular Quantum Numbers. A set of 42 integer value descriptors of molecular structure. See Legend of Fig. 3 for detailed listing
MQN-bin	A position in MQN-space corresonding to a particular combination of 42 MQN-values. An MQN-bin may contain one or several molecules
MQN-space	42-dimensional space created by the MQNs
Ro3	Subset of PubChem hac < 20 following Congreve's "rule of 3" for fragment-likeness
Ro3+	Subset of Ro3 with further restrictions in rotatable bonds (rbc ≤ 3) and polar surface area (PSA ≤ 60 Å^2)
Ro5	Subset of the PubChem hac ≤ 20 following Lipinski's "rule of 5"
SF	Substructure fingerprint. Here the 1024-bit Daylight-type substructure fingerprint in the JChem package of ChemAxon was used
T_{MQN}	Scalar Tanimoto similarity coefficient comparing two scalar vectors of 42 MQN-values. Values between 0 and 1 with 1 indicating high similarity
T_{SF}	Binary Tanimoto similarity coefficient comparing two binary substructure fingerprints. Values between 0 and 1 with 1 indicating high similarity

References

1. Coyne AG, Scott DE, Abell C (2010) Drugging challenging targets using fragment-based approaches. Curr Opin Chem Biol 14:299–307
2. Schulz MN, Hubbard RE (2009) Recent progress in fragment-based lead discovery. Curr Opin Pharmacol 9:615–621
3. Hartenfeller M, Schneider G (2011) De novo drug design. Methods Mol Biol 672:299–323
4. Venhorst J, Nunez S, Kruse CG (2010) Design of a high fragment efficiency library by molecular graph theory. ACS Med Chem Lett 1:499–503
5. Carr RA, Congreve M, Murray CW, Rees DC (2005) Fragment-based lead discovery: leads by design. Drug Discov Today 10:987–992
6. Rees DC, Congreve M, Murray CW, Carr R (2004) Fragment-based lead discovery. Nat Rev Drug Discov 3:660–672
7. Hajduk PJ, Greer J (2007) A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov 6:211–219
8. Boyd SM, de Kloe GE (2010) Fragment library design: efficiently hunting drugs in chemical space. Drug Discov Today 7:e173–e180
9. Dobson CM (2004) Chemical space and biology. Nature 432:824–828
10. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623–W633
11. Nguyen KT, Blum LC, van Deursen R, Reymond JL (2009) Classification of organic molecules by molecular quantum numbers. ChemMedChem 4:1803–1805
12. Wang SG, Schwarz WH (2009) Icon of chemistry: the periodic system of chemical elements in the new century. Angew Chem Int Ed Engl 48:3404–3415
13. Pearlman RS, Smith KM (1998) Novel software tools for chemical diversity. Perspect Drug Discov Des 9–11:339–353
14. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216
15. Reymond JL, Van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Med Chem Commun 1:30–38
16. Akella LB, DeCaprio D (2010) Cheminformatics approaches to analyze diversity in compound screening libraries. Curr Opin Chem Biol 14:325–330
17. Burden FR (1989) Molecular-Identification Number for Substructure Searches. J Chem Inf Comput Sci 29:225–227
18. Pearlman RS, Smith KM (1999) Metric validation and the receptor-relevant subspace concept. J Chem Inf Comput Sci 39:28–35
19. Oprea TI, Gottfries J (2001) Chemography: the art of navigating in chemical space. J Comb Chem 3:157–166
20. Rosen J, Gottfries J, Muresan S, Backlund A, Oprea TI (2009) Novel chemical space exploration via natural products. J Med Chem 52:1953–1962
21. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11:1046–1053
22. Fink T, Bruggesser H, Reymond JL (2005) Virtual exploration of the small-molecule chemical universe below 160 daltons. Angew. Chem Int Ed Engl 44:1504–1508
23. Fink T, Reymond JL (2007) 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 47:342–353
24. Blum LC, Reymond JL (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. J Am Chem Soc 131:8732–8733
25. Blum LC, van Deursen R, Reymond JL (2011) Visualisation and subsets of the chemical universe database GDB-13 for virtual

screening. J Comput Aided Mol Des. doi:10.1007/s10822-011-9436-y
26. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49:6789–6801
27. van Deursen R, Blum LC, Reymond JL (2010) A searchable map of PubChem. J Chem Inf Model 50:1924–1934
28. Siegal G, Ab E, Schultz J (2007) Integration of fragment screening and library design. Drug Discov Today 12:1032–1039
29. Congreve M, Chessari G, Tisi D, Woodhead AJ (2008) Recent developments in fragment-based drug discovery. J Med Chem 51:3661–3680
30. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Del Rev 23:3–25
31. Teague SJ, Davis AM, Leeson PD, Oprea T (1999) The Design of Leadlike Combinatorial Libraries. Angew Chem Int Ed Engl 38:3743–3748
32. Congreve M, Carr R, Murray C, Jhoti H (2003) A rule of three for fragment-based lead discovery? Drug Discov Today 8:876–877
33. Khalifa AA, Haranczyk M, Holliday J (2009) Comparison of nonbinary similarity coefficients for similarity searching, clustering and compound selection. J Chem Inf Model 49:1193–1201
34. Overington J (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI) Interview by Wendy A. Warr. J Comput Aided Mol Des 23:195–198
35. Schuffenhauer A, Brown N, Selzer P, Ertl P, Jacoby E (2005) Relationships between molecular complexity, biological activity, and structural diversity. J Chem Inf Model 46:525–535
36. Krier M, Bret G, Rognan D (2006) Assessing the scaffold diversity of screening libraries. J Chem Inf Model 46:512–524
37. Rogers D, Hahn M (2010) Extended-Connectivity Fingerprints. J Chem Inf Model 50:742–754
38. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996