JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

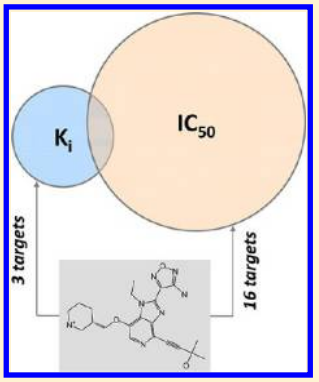
pubs.acs.org/jcim

Growth of Ligand–Target Interaction Data in ChEMBL Is Associated with Increasing and Activity Measurement-Dependent Compound Promiscuity

Ye Hu and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

ABSTRACT: Compounds with high-confidence target annotations and activity measurements in the original and current release of the ChEMBL database have been compared to better understand how the growth of compound activity data might influence the spectrum of ligand–target interactions and the degree of target promiscuity among active compounds. Compared to the original ChEMBL release, a significant increase in the proportion of target promiscuous compounds was observed in the current version. The presence of these compounds led to large-magnitude changes in compound activity-based target and target family relationships and to a reorganization of major target communities. Surprisingly, however, this strong trend toward increasing target promiscuity was largely caused by growth of compounds with exclusive IC_{50} measurements. By contrast, compounds with available equilibrium constants, which were also added in large amounts, did not substantially alter compound-based target relationships and notably contribute to increasing target promiscuity. These findings suggest that apparent compound promiscuity is much dependent on experimental conditions under which activities are determined and that care should be taken when evaluating promiscuity and polypharmacology on the basis of assay-dependent activity measurements.



INTRODUCTION

Compound activity data provide an invaluable source for the exploration of structure–activity relationships (SARs) and ligand–target interactions. Systematic analysis of pharmaceutically relevant compounds with activity against different target families has much contributed to our current understanding of SARs[1,2] and polypharmacological compound and drug behavior.[3−5] For these types of data mining investigations, the availability of public and commercial databases of bioactive compounds[6−10] plays a critically important role. Consequently, much attention has recently been focused on describing such databases, analyzing their compound composition, and comparing them.[6−10] Especially major public compound data repositories[6−8] have become indispensable resources, because knowledge extracted from them can be freely communicated, both in academia and the pharmaceutical industry.

Previously, we have systematically analyzed compound activity annotations to identify molecular scaffolds that exclusively occurred in compounds active against individual target families.[11] Furthermore, in light of the emerging theme of polypharmacology,[3−5] molecular scaffolds have also been discovered through systematic data mining that represented multiple compounds with activity across different target families.[12]

Going beyond scaffold-centric analysis, we have been interested in exploring how the current growth of compound activity data might affect compound-based target or target family relationships and the degree of target promiscuity of active compounds. Therefore, we have carried out a detailed comparison of the original release 1 and release 13 of the ChEMBL database,[13] the results of which are presented herein.

Table 1. Global Comparison of ChEMBL Release 1 and 13[a]

number of		ChEMBL 1	ChEMBL 13
compounds		31010	103783 (3.3)
targets		576	1089 (1.9)
compound–target combinations		48269	156262 (3.2)
target families		122	253 (2.1)
target pairs	intrafamily	602	1267 (2.1)
	interfamily	80	454 (5.7)
	all	682	1721 (2.5)
scaffolds		12291	38622 (3.1)
CSKs		5741	15899 (2.8)

[a]For ChEMBL release 1 and 13, the numbers of bioactive compounds, targets, compound–target combinations, target families, and target pairs (belonging to the same or different families) are reported. In addition, the numbers of unique scaffolds and cyclic skeletons (CSKs) extracted from active compounds are provided. Data growth factors for release 13 compared to 1 are given in parentheses.

MATERIALS AND METHODS

Compound and Target Data. From ChEMBL release 1 and 13,[14] bioactive compounds with direct interactions (i.e., target relationship type "D") against human targets at the highest

Received: July 13, 2012
Published: September 16, 2012

confidence level (i.e., target confidence score 9) were extracted. Two types of potency measurements were considered, i.e., K_i and IC_{50} values. In order to ensure high data confidence, approximate measurements such as ">", "<", or "~" were not considered. If both precise and approximate measurements were reported for a given compound, only precise measurements were retained for further analysis. For compounds with multiple K_i or IC_{50} measurements against the same target, the geometric mean of all potency values was calculated as the final potency annotation. No potency threshold value was applied. All qualifying compounds were further organized into individual target sets. In addition, all targets were grouped into target families following the UniProt[15] family annotation and the protein classification hierarchy of ChEMBL.[13]

Table 2. Target Family Distribution[a]

Conserved Target Families	
growth (no. additional targets)	number of families
0	69
1−5	38
6−10	8
>10	7
all	122
Novel Target Families	
number of targets	number of families
1−5	129
6−10	1
>10	1
all	131

[a]For the 122 conserved target families that are present in both ChEMBL release 1 and 13, the growth of the families is reported as the number of additional targets in release 13. In addition, for the 131 novel target families that are only present in release 13, the target composition is provided.

For chemotype analysis, compounds were decomposed into molecular scaffolds according to Bemis and Murcko[16] by removing all side chains. In addition, scaffolds were further reduced to cyclic skeletons (CSKs)[17] by converting all heteroatoms to carbon and setting bond orders to one.

For ChEMBL release 1 and 13, data sets were generated including qualifying active compounds, their targets, target families, compound-target combinations, intra- and interfamily target pairs, molecular scaffolds, and CSKs.

Target Pairs and Networks. For each ChEMBL release, target pairs were identified that shared at least five active compounds. Target family annotations of targets forming a pair were compared in order to identify intra- and interfamily target pairs. Network representations were generated using Cytoscape[18] to analyze relationships between targets. In these networks, nodes represent targets and edges are drawn between nodes if they share at least five active compounds. Interfamily target pairs are indicated by bold black edges.

Potency-Based Data Organization. For ChEMBL release 1 and 13, data subsets were generated by exclusively considering K_i and IC_{50} values, respectively. At least five compounds with K_i or IC_{50} values had to be available for inclusion of a target pair into a subset. Hence, target pairs in the global data set might not occur in the K_i or IC_{50} subsets if they shared less than five compounds with K_i or IC_{50} measurements. In addition, for each subset, the number of targets against which each compound was active was counted as a measure of the degree of promiscuity.

The K_i and IC_{50} subsets belonging to the same ChEMBL release were compared. In addition, targets and compounds that were shared by K_i and IC_{50} subsets were identified. For shared targets, the distributions of active compounds and scaffolds were compared. Furthermore, for shared compounds, the degree of their target promiscuity was compared.

All database mining calculations and data analysis were carried out with in-house generated programs.

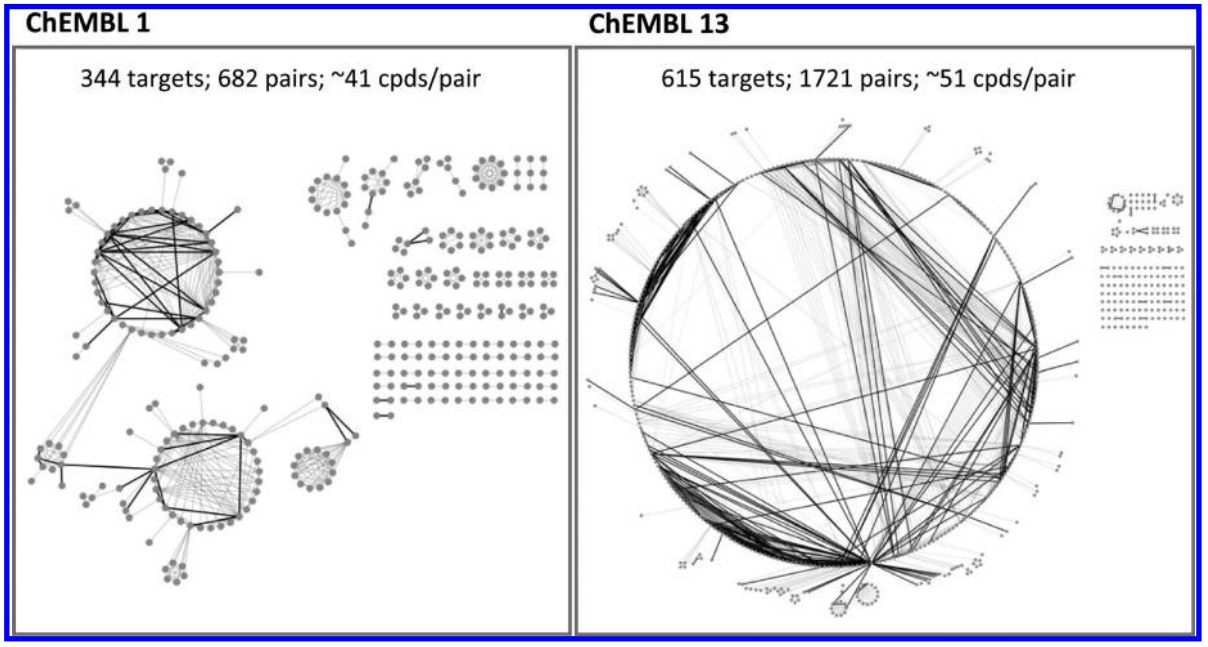


Figure 1. Target networks. Shown are compound-based target networks for ChEMBL 1 and 13. In each network, nodes represent targets that are connected by an edge if they share at least five active compounds. If two targets forming a pair belong to different target families, the corresponding edge is drawn in bold. Furthermore, the numbers of targets, target pairs, and the average number of compounds per target pair are reported.

dx.doi.org/10.1021/ci3003304 | J. Chem. Inf. Model. 2012, 52, 2550−2558

RESULTS AND DISCUSSION

ChEMBL 1 and 13 were released in January, 2010, and March, 2012, respectively. We initially carried out a global comparison of the compound data that qualified for our analysis.

Global Data Set Comparison. *Composition and Growth.* On the basis of our selection criteria, a total of 31 010 compounds active against 576 human targets were assembled from ChEMBL release 1 that represented a total of 48 269 unique ligand-target annotations, as reported in Table 1. These targets belonged to 122 target families. A total of 682 target pairs were identified that shared at least five active compounds. Among these, 602 target pairs (~88%) were formed within the same family and 80 pairs

Table 3. Global Distribution of Interfamily Target Pairs[a]

	family 1	family 2	target pairs
ChEMBL 1	monoamine receptor family	sodium neurotransmitter symporter family	16
	Ser/Thr protein kinase family	tyrosine protein kinase family	13
	monoamine receptor family	nucleotide-like receptor family	11
	monoamine receptor family	short peptide receptor family	7
ChEMBL 13	Ser/Thr protein kinase family	tyrosine protein kinase family	52
	monoamine receptor family	sodium neurotransmitter symporter family	39
	cytochrome P450 family	short-chain dehydrogenases/reductases family	13
	monoamine receptor family	VGC ion channel family	12
	monoamine receptor family	nucleotide-like receptor family	11
	monoamine receptor family	short peptide receptor family	10
	short peptide receptor family	VGC ion channel family	10
	cysteine protease family	PPP phosphatase family	8
	carbonic anhydrase family	prostaglandin G/H synthase family	8
	cytochrome P450 family	short peptide receptor family	7
	LGIC ion channel family	monoamine receptor family	7
	cytochrome P450 family	type-B carboxylesterase/lipase family	6
	cytochrome P450 family	tyrosine protein kinase family	6
	cytochrome P450 family	VGC ion channel family	6
	carbonic anhydrase family	metallo protease family	6

[a]For ChEMBL 1 and 13, target family relationships formed by more than five interfamily target pairs are reported.

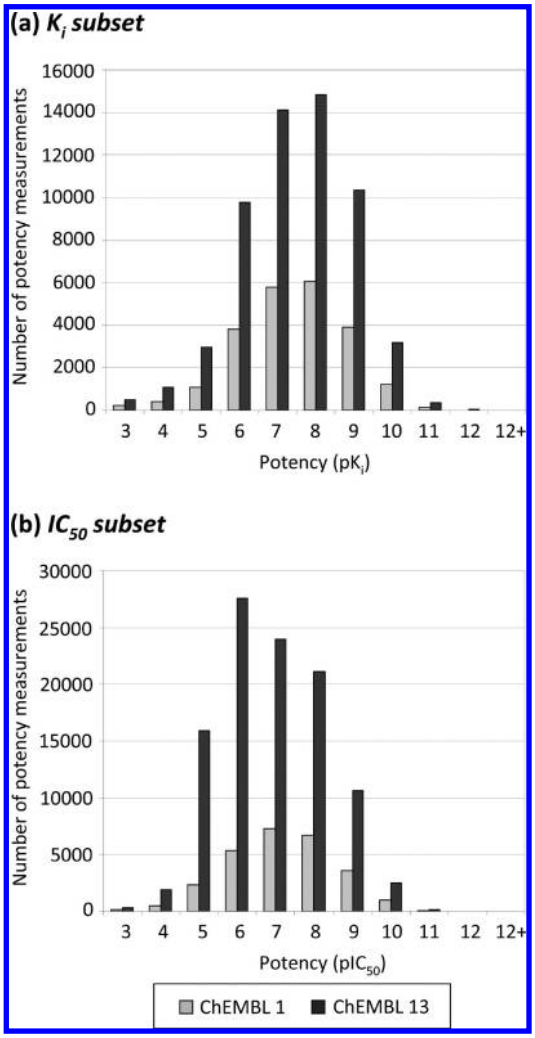


Figure 2. Potency distribution of compounds in K_i and IC_{50} subsets. Shown are the potency value distributions in the (a) K_i and (b) IC_{50} subsets of ChEMBL 1 and 13.

Table 4. Data Composition of K_i and IC_{50} Subsets[a]

number of		ChEMBL 1		ChEMBL 13	
		K_i	IC_{50}	K_i	IC_{50}
compounds		12990	19264	33977 (2.6)	75236 (3.9)
targets		322	498	543 (1.7)	989 (2.0)
compound-target combinations		22614	27038	57195 (2.5)	104273 (3.9)
target families		71	101	130 (1.8)	229 (2.3)
target pairs	intrafamily	337	382	608 (1.8)	901 (2.4)
	interfamily	15	52	52 (3.5)	334 (6.4)
	all	352	434	660 (1.9)	1235 (2.8)
scaffolds		5216	7844	12113 (2.3)	29347 (3.7)
CSKs		2663	3826	6102 (2.3)	12177 (3.2)

[a]For ChEMBL release 1 and 13, the numbers of available compounds, targets, compound-target combinations, target families, and target pairs belonging to the same or different families are reported for K_i- and IC_{50}-based subsets, respectively. In addition, the numbers of unique scaffolds and cyclic skeletons (CSKs) are given. Data growth factors for release 13 compared to 1 are given in parentheses.

dx.doi.org/10.1021/ci3003304 | J. Chem. Inf. Model. 2012, 52, 2550−2558

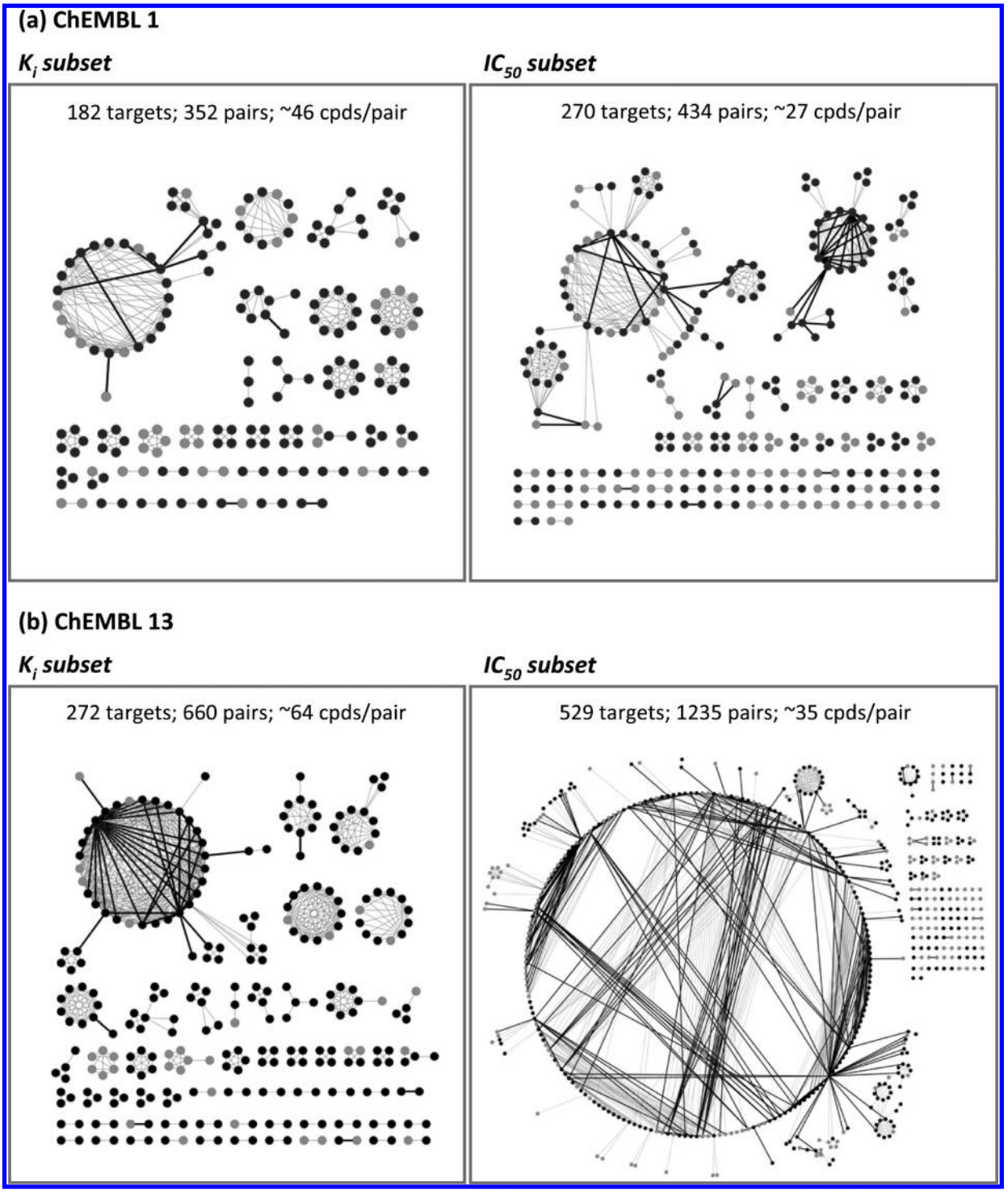


Figure 3. Target networks for K_i and IC_{50} subsets. Shown are compound-based target networks for the K_i- and IC_{50}-based subsets of ChEMBL 1 (a) and 13 (b). In each network, nodes represent targets that are connected by an edge if they share at least five active compounds. Nodes of targets that are shared by K_i and IC_{50} sets are colored black. If two targets forming a pair belong to different target families, the corresponding edge is drawn in bold. Furthermore, the numbers of targets, target pairs, and the average number of compounds per target pair are reported.

(~12%) across different families. In addition, these compounds represented 12 291 unique molecular scaffolds and 5741 CSKs. Each CSK accounts for a subset of topologically equivalent scaffolds.

From ChEMBL release 13, 103 783 compounds active against 1089 human targets were obtained yielding with 156 262 ligand−target annotations (Table 1). These targets belonged to 253 different families. In addition, 1721 target pairs were identified that included 1267 intrafamily (~74%) and 454 interfamily (~26%) pairs. Compounds represented a total of 38 622 unique scaffolds and 15 899 CSKs.

The comparison of these numbers revealed that ligand−target interaction data increased by a factor of ~2−3 from ChEMBL 1 to 13.

Target Family Based Comparison. In addition to 122 target families that were already present in ChEMBL 1, release 13

dx.doi.org/10.1021/ci3003304 | *J. Chem. Inf. Model.* 2012, 52, 2550−2558

contained 131 novel target families. Hence, available target families also more than doubled. We next compared the target composition of the 122 original families common to ChEMBL 1 and 13, as reported in Table 2. A total of 69 target families remained unchanged, i.e., no new targets were added. However, the number of targets in 38 target families increased by one to five targets, in eight families by six to 10, and in seven families by more than 10 targets. Table 2 also reports the number of targets in the 131 families that were not present in ChEMBL 1. A total of 129 of these novel families contained only small numbers of targets, i.e., from one to five. Thus, most of the growth of ligand−target interaction data was centered on target families that were already contained in the database at the beginning of 2010.

Intra- and Interfamily Target Relationships. We also determined that the total number of target pairs sharing at least five active compounds nearly tripled in ChEMBL 13 compared to release 1. Hence, the growth of target pair sets was substantial. Surprisingly, among these pair sets, the number of interfamily target pairs increased by a factor of more than 6.

Target relationships were further analyzed in compound-based target networks, as shown in Figure 1. Here, striking differences were observed between ChEMBL 1 and 13. In these networks, separate communities of targets were formed on the basis of shared active compounds. In the network of ChEMBL 1, several well-organized communities with multiple targets were observed. These communities predominantly consisted of individual target families. For example, the two major communities on the left of the ChEMBL 1 network in Figure 1 contained targets from G protein coupled receptor (GPCR) subfamilies, different kinase subfamilies, proteases from different families, and the cytochrome P450 enzyme family.

By contrast, the ChEMBL 13 network, also shown in Figure 1, displayed a large central network component that captured the majority of interfamily target pairs we detected. Accordingly, this network component included many targets from different (sub)families including kinases, GPCRs, proteases, ion channel families, cytochrome P450 enzymes, histone deacetylases, and phospholipases. The presence of the large central component in the ChEMBL 13 network and its absence in the ChEMBL 1 network indicated a substantial increase in target promiscuity among active compounds in ChEMBL 13. Table 3 reports target family relationships in ChEMBL 1 and 13 that were formed by more than five interfamily target pairs, which mirrored this trend. In ChEMBL 1, only four such relationships were observed involving a maximum of 16 interfamily target pairs. By contrast, in ChEMBL 13, 15 family relationships were detected. The two dominant relationships were formed between Ser/Thr and Tyr kinases and the monoamine receptor and sodium neurotransmitter symporter family, involving 52 and 39 target pairs, respectively. In addition, relationships between different subfamilies of GPCRs were observed as well as other relationships involving the monoamine receptor and cytochrome P450 families.

Subset Comparison. In order to further explore apparent compound promiscuity differences between ChEMBL 1 and 13, we decided to separately consider K_i and IC_{50} value-based activity annotations. Therefore, the ChEMBL 1 and 13 data sets were divided into potency measurement-dependent subsets. The composition of each subset is reported in Table 4. For ChEMBL 1, the amount of structure- and activity-related data was consistently larger for IC_{50} than for K_i measurements, as one might expect (given that K_i measurements represent equilibrium constants that require more experimental efforts). From release 1 to 13, the data volumes according to Table 4 increased by a factor of 1.7 to 3.5 in the K_i subset and of 2.0 to 6.4 in the IC_{50} subset. Hence, IC_{50}-based ligand-target interaction data grew much faster, thereby widening the gap between IC_{50}- and K_i-based data. The potency value distributions of the K_i and IC_{50} subsets are reported in Figure 2.

Table 5. Distribution of Interfamily Target Pairs in K_i and IC_{50} Subsets[a]

		family 1	family 2	target pairs
ChEMBL 1	K_i	monoamine receptor family	sodium neurotransmitter symporter family	6
	IC_{50}	Ser/Thr protein kinase family	tyrosine protein kinase family	11
		monoamine receptor family	sodium neurotransmitter symporter family	10
		monoamine receptor family	nucleotide-like receptor family	9
		monoamine receptor family	short peptide receptor family	7
ChEMBL 13	K_i	monoamine receptor family	sodium neurotransmitter symporter family	29
		monoamine receptor family	VGC ion channel family	10
	IC_{50}	Ser/Thr protein kinase family	tyrosine protein kinase family	48
		cytochrome P450 family	short-chain dehydrogenases/reductases family	13
		monoamine receptor family	sodium neurotransmitter symporter family	13
		monoamine receptor family	nucleotide-like receptor family	10
		monoamine receptor family	short peptide receptor family	9
		cysteine protease family	PPP phosphatase family	8
		cytochrome P450 family	tyrosine protein kinase family	6
		short peptide receptor family	VGC ion channel family	6

[a]For K_i and IC_{50} subsets of ChEMBL 1 and 13, target family relationships formed by more than five interfamily target pairs are reported.

Target Pair Networks. Analogously to our network analysis of the complete data sets, we then generated target pair networks for the K_i- and IC_{50}-based subsets of ChEMBL 1 and 13. These network representations are displayed in Figure 3. Targets that were present in both K_i and IC_{50} subsets of the same ChEMBL release are colored black. In Figure 3a and b, networks of the K_i and IC_{50} subsets of ChEMBL 1 and 13 are compared, respectively. For ChEMBL 1, the topology of both networks was comparable and similar to the network of the complete ChEMBL 1 data set (Figure 1). In both cases, individual target communities of comparable sizes were formed. One relatively large target community was observed together with several smaller ones representing well-defined target families. The largest community in both networks consisted of targets exclusively from monoamine receptor family (K_i subset) or different kinases and protease families (IC_{50} subset). By contrast, the topology of the two ChEMBL 13 networks substantially differed. The network of the K_i subset was comparable to the one of release 1. In this case, a number of separate communities were also formed by individual target families. Given the data growth, communities were more densely connected in the network of the ChEMBL 13 K_i subset than in the corresponding ChEMBL 1 network, but the network topology remained similar.

By contrast, the network of the IC_{50} subset of ChEMBL 13 differed greatly from the other networks. Similar to the network of the complete ChEMBL 13 data set (Figure 1), the network of the IC_{50} subset of ChEMBL 13 was dominated by a large central network component of heterogeneous target composition. Thus, on the basis of these comparisons, dramatically increasing apparent compound promiscuity in ChEMBL 13 was largely attributed to growing amounts of IC_{50}-based ligand-target interaction data. On average, a target pair from a K_i subset in both ChEMBL releases contained about 20 more compounds than a pair from an IC_{50} subset. Although the large number of compounds per target pair set would statistically increase the likelihood of apparent compound promiscuity, K_i-based interaction data displayed greater target selectivity than IC_{50}-based data, as revealed by the network views.

Table 6. Promiscuity of Compounds[a]

	Number of Compounds			
	ChEMBL 1		ChEMBL 13	
number of targets (promiscuity)	K_i	IC_{50}	K_i	IC_{50}
1	7871	14489	21005	56960
2	2284	3055	6911	12043
3	1828	1047	3840	3774
4	650	434	1489	1602
5	236	91	424	405
6	49	59	80	169
7	23	50	49	133
8	20	18	17	54
9	18	6	21	30
10	6	3	38	17
>10	5	12	103	49

[a]For ChEMBL release 1 and 13, the distribution of compounds active against different numbers of targets is reported for K_i and IC_{50} subsets, respectively. The number of targets a compound is active against serves as a measure of promiscuity.

Family Relationships. Table 5 reports target family relationships in K_i and IC_{50} subsets of ChEMBL 1 and 13. The listed relationships involved more than five interfamily target pairs. Separating interaction data on the basis of K_i and IC_{50} values provided a further differentiated view of family relationships compared to Table 3. The monoamine receptor family was recurrent in forming relationships across all subsets. In K_i-based subsets, of ChEMBL 1 and 13, only one and two relationships were detected, respectively. These relationships involved the monoamine receptor family and the sodium neurotransmitter symporter (ChEMBL 1 and 13) and VGC ion channel family (ChEMBL 13), respectively. These relationships were also found in the corresponding IC_{50} subsets. However, relationships between Ser/Thr and Tyr kinases were only found in the IC_{50} subsets of ChEMBL 1 and ChEMBL 13 and were formed by 11 and 48 target pairs, respectively. In the IC_{50} subset of ChEMBL 1, only two other relationships involving the monoamine receptor family were detected. By contrast, in the IC_{50} subset of ChEMBL 13 there were eight relationships involving Tyr kinases, cytochrome P450 enzymes, and GPCR subfamilies. Thus, in the IC_{50} subset of ChEMBL 13, there was a notable increase in apparent compound promiscuity beyond the monoamine receptor and kinase families. Many compounds from the IC_{50} subset of ChEMBL 13 were characterized by reported multitarget activities across different families.

Promiscuity of Compounds. On the basis of these findings, the degree of target promiscuity of compounds from the K_i and IC_{50} subsets was systematically analyzed by determining the number of targets they were active against. In Table 6, the distribution of compounds with activity against 1−10 and more than 10 targets is reported. The majority of compounds in all subsets were only annotated with a single target. There were 5119 and 4775 compounds with multitarget activity in the K_i and IC_{50} subsets of ChEMBL 1, respectively. Thus, in this case, the K_i subset contained more promiscuous compounds than the IC_{50} subset. By contrast, there were 12 972 and 17 885 compounds with multitarget activity

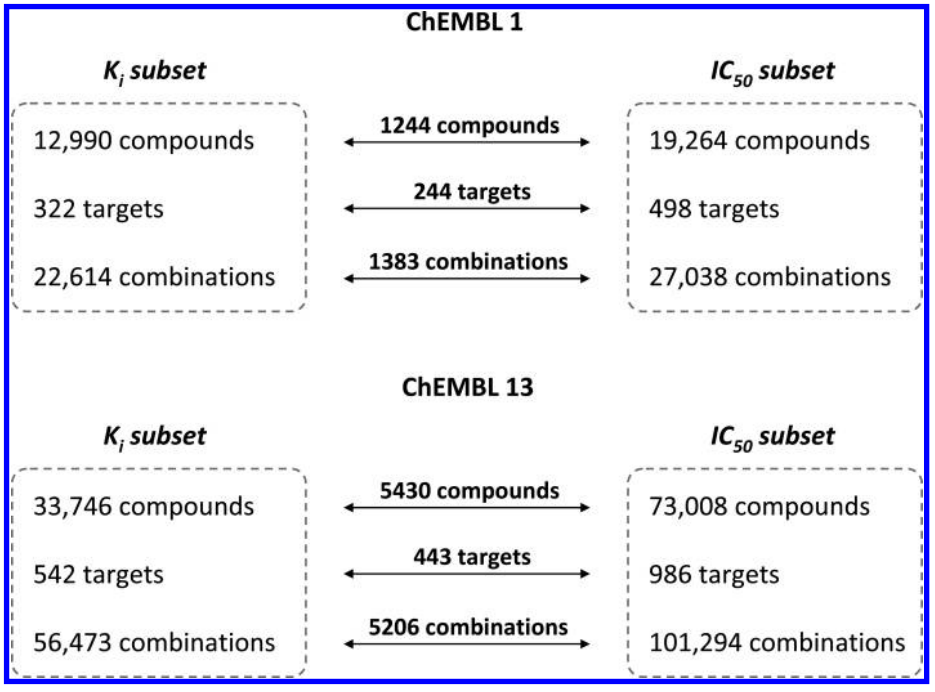


Figure 4. Data overlap. For ChEMBL 1 and 13, the data overlap between K_i- and IC_{50}-based subsets is reported including the number of shared compounds, targets, and compound−target combinations.

dx.doi.org/10.1021/ci3003304 | J. Chem. Inf. Model. 2012, 52, 2550−2558

in the K_i and IC_{50} subsets of ChEMBL 13, respectively. Thus, whereas the number of compounds with single- and multitarget activities in the K_i and IC_{50} subsets of ChEMBL 1 was comparable, the IC_{50} subset of ChEMBL 13 contained nearly 5000 more compounds with multitarget activities than its K_i subset, which further rationalized the observed distribution of interfamily target pairs as well as the formation of the large central component in the ChEMBL 13 target pair network and the network of its IC_{50} subset.

Data Overlap and Distribution Characteristics. We also determined the overlap of compounds, targets, and compound-target combinations for the subsets of ChEMBL 1 and 13. The results are shown in Figure 4. For ChEMBL 1, less than 1% of compounds and compound-target combinations were present in both subsets. However, the majority of the targets from the K_i subset were also found in the IC_{50} subset. Comparable observations were made for ChEMBL 13. Thus, there was only little overlap between K_i- and IC_{50}-based interaction data.

For targets common to both subsets, the distribution of active compounds and their scaffolds was systematically compared, which revealed the presence of many targets with very different compound and scaffold distributions. Selected target sets with different characteristics are reported in Tables 7–9. In Table 7, three target sets are shown with no or very limited compound and scaffold overlap between the K_i and IC_{50} subsets, consistent with the limited global data overlap (Figure 4). In Table 8, three other target sets are presented for which compound and scaffold overlap was comparably large for ChEMBL 1 and 13. In these cases, ~100–200 compounds were common to both subsets. Furthermore, Table 9 shows three target sets where the compound distribution was very different for the K_i and IC_{50} subsets.

For compounds that were common to the K_i and IC_{50} subsets, the degree of promiscuity was compared, as reported in Table 10.

Table 7. Representative Target Sets with Small Compound Overlap[a]

	number of compounds			number of scaffolds		
Dipeptidyl Peptidase IV—Serine Protease Family	K_i	IC_{50}	shared	K_i	IC_{50}	shared
ChEMBL 1	128	171	0	64	70	3
ChEMBL 13	276	1277	11	122	441	12
Neurokinin 1 Receptor—Short Peptide Receptor Family						
ChEMBL 1	93	233	0	48	105	0
ChEMBL 13	207	451	2	85	162	5
C–C Chemokine Receptor Type 3—Chemokine Receptor Family						
ChEMBL 1	100	71	0	40	25	0
ChEMBL 13	100	277	0	40	87	0

[a]Shown are three representative target sets common to K_i and IC_{50} subsets that have small compound and scaffold overlap. In each case, the numbers of compounds and scaffolds in K_i and/or IC_{50} subsets are reported for ChEMBL 1 and 13, respectively.

Table 8. Representative Target Sets with Large Compound Overlap[a]

	number of compounds			number of scaffolds		
Mu Opioid Receptor—Short Peptide Receptor Family	K_i	IC_{50}	shared	K_i	IC_{50}	shared
ChEMBL 1	493	203	119	228	118	78
ChEMBL 13	1377	432	196	510	192	98
Dopamine Transporter—Sodium Neurotransmitter Symporter Family						
ChEMBL 1	262	240	113	83	91	42
ChEMBL 13	628	747	188	176	209	65
Norepinephrine Transporter—Sodium Neurotransmitter Symporter Family						
ChEMBL 1	365	205	94	95	68	38
ChEMBL 13	853	1097	206	205	239	72

[a]Shown are three representative target sets common to K_i and IC_{50} subsets that have relatively large compound and scaffold overlap. In each case, the numbers of compounds and scaffolds in K_i and/or IC_{50} subsets are reported for ChEMBL 1 and 13, respectively.

Table 9. Representative Target Sets with Different Compound Composition[a]

	number of compounds			number of scaffolds		
Tyrosine Kinase LCK—Tyrosine Protein Kinase Family	K_i	IC_{50}	shared	K_i	IC_{50}	shared
ChEMBL 1	2	366	0	2	182	1
ChEMBL 13	3	418	0	3	213	1
Vascular Endothelial Growth Factor Receptor 2—Tyrosine Protein Kinase Family						
ChEMBL 1	6	1051	4	4	502	3
ChEMBL 13	30	1493	8	13	634	7
Carbonic Anhydrase IX—Carbonic Anhydrase Family						
ChEMBL 1	416	1	1	144	1	1
ChEMBL 13	927	3	2	305	3	3

[a]Shown are three representative target sets common to K_i and IC_{50} subsets that contain significantly different numbers of compounds and scaffolds. In each case, the numbers of compounds and scaffolds in K_i and/or IC_{50} subsets are reported for ChEMBL 1 and 13, respectively.

Table 10. Promiscuity of Compounds Contained in Both K_i and IC_{50} Subsets[a]

	number of compounds	
comparison of the number of targets	ChEMBL 1	ChEMBL 13
$T(K_i) = T(IC_{50})$	737	3626
$T(K_i) > T(IC_{50})$	351	1166
$T(K_i) < T(IC_{50})$	156	638
total	1244	5430

[a]For compounds shared by the K_i and IC_{50} subsets, the number of target annotations in ChEMBL release 1 and 13 was compared. $T(K_i)$ and $T(IC_{50})$ refer to the number of target annotations of a compound in the K_i and IC_{50} subsets, respectively. For example, in ChEMBL release 1, a total of 737 compounds had the same number of target annotations in both subsets and 351 compounds had a larger number of annotations in the K_i subset.

dx.doi.org/10.1021/ci3003304 | J. Chem. Inf. Model. 2012, 52, 2550–2558

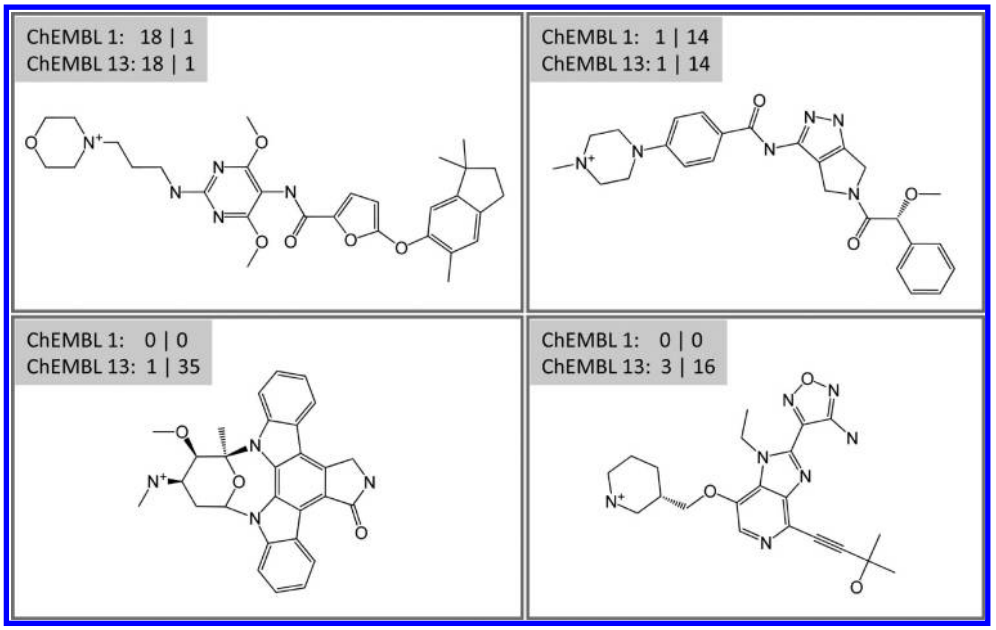


Figure 5. Representative compounds with different degrees of promiscuity. Shown are four representative compounds having different degrees of target promiscuity. For each compound, the number of targets against which it is active in the K_i and IC_{50} subsets is reported for ChEMBL 1 and 13. For example, "18|1" indicates that the compound is active against 18 targets in the K_i subset and 1 target in the IC_{50} subset, whereas "1|14" indicates that the compound is active against 1 target in the K_i subset and 14 targets in the IC_{50} subset.

For ChEMBL 1, 1244 compounds were shared and 737 of these displayed the same degree of promiscuity in both sets, i.e., they were active against the same number of targets, but not necessarily against the same target(s). The remaining 507 compounds were active against different numbers of targets (the difference in target numbers ranged from one to 17). For ChEMBL 13, 3626 of the 5430 shared compounds had the same degree of target promiscuity. For the remaining 804 compounds, differences in the number of their targets ranged from one to 34. Thus, ∼40% and ∼15% of shared compounds had different degrees of promiscuity in the K_i and IC_{50} subsets of ChEMBL 1 and 13, respectively. However, it should be noted that only ∼1% of all compounds ChEMBL 1 and 13 had both K_i and IC_{50} values. Figure 5 shows four exemplary compounds. Taken together, the results revealed that the observed increase in apparent compound promiscuity in ChEMBL 13 was largely due to compounds not present in ChEMBL 1 for which only IC_{50} values were available.

CONCLUSIONS

In this study, we have analyzed how growing public domain compound activity data affects compound-based target and target family relationships and distributions of promiscuous compounds. The growth of compound data is important to monitor as it increasingly balances potential biases in ligand-target analysis associated with data sparseness.[19] We have found that growth of compound data was globally accompanied by increasing target promiscuity, as revealed by target pair set analysis. In target pair networks, previously well-defined target communities merged into a large central component of heterogeneous target family composition. However, we also determined that increasing target promiscuity was dependent on the type of activity measurements that were considered. Increasing promiscuity was largely due to new compounds for which only IC_{50} measurements were available, which dominated the global view. By contrast, for compounds with K_i values, whose numbers also significantly increased, this trend was not observed. Rather, in this case, the target community organization in networks remained largely constant, despite data growth, and only very few new target family relationships were formed. Furthermore, the majority of compounds for which IC_{50} and K_i values were available were active against the same number of targets, but not necessarily against the same target(s). Thus, the assessment of target promiscuity of compounds should be considered with caution with respect to the activity measurements that are utilized. On the basis of our analysis, it would be preferred to base investigations of apparent compound promiscuity and polypharmacology on equilibrium constants, because the use of assay-dependent IC_{50} values might give rise to false-positive assignments. For practical purposes, this might not always be feasible. However, an awareness of potential caveats should be raised.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Peltason, L.; Bajorath, J. Systematic computational analysis of structure-activity relationships: concepts, challenges, and recent advances. *Future Med. Chem.* **2009**, *1*, 451−466.

(2) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure-activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(3) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.

(4) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197−206.

(5) Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.* **2008**, *4*, 682−690.
(6) Bender, A. Databases: compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309.
(7) Wassermann, A. M.; Bajorath, J. BindingDB and ChEMBL − online compound databases for drug discovery. *Expert Opin. Drug Discovery* **2011**, *6*, 683−687.
(8) Nicola, G.; Liu, T.; Gilson, M. K. Public domain databases for medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 6987−7002.
(9) Tiikkainen, P.; Franke, L. Analysis of commercial and public bioactivity databases. *J. Chem. Inf. Model.* **2011**, *51*, 319−326.
(10) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today* **2012**, *17*, 685−701.
(11) Hu, Y.; Wassermann, A. M.; Lounkine, E.; Bajorath, J. Systematic analysis of public domain compound potency data identifies selective molecular scaffolds across druggable target families. *J. Med. Chem.* **2010**, *53*, 752−758.
(12) Hu, Y.; Bajorath, J. Polypharmacology directed data mining: identification of promiscuous chemotypes with different activity profiles and comparison to approved drugs. *J. Chem. Inf. Model.* **2010**, *50*, 2112−2118.
(13) 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.
(14) ChEMBL. http://www.ebi.ac.uk/chembldb/ (accessed March 1, 2012)
(15) UniProtConsortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D142−D148.
(16) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(17) Xu, Y.-J.; Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 912−926.
(18) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498−2504.
(19) Mestres, J.; Gregori-Puigjane, E.; Valverde, S.; Sole, R. V. Data completeness−the achilles heel of drug-target networks. *Nat. Biotechnol.* **2008**, *26*, 983−984.