JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

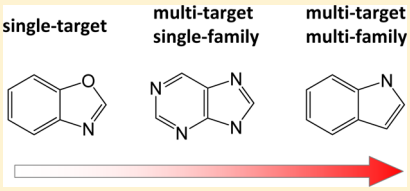
pubs.acs.org/jcim

Systematic Identification of Scaffolds Representing Compounds Active against Individual Targets and Single or Multiple Target Families

Ye Hu and Jürgen Bajorath*

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

ABSTRACT: Given the enormous growth of compound activity data we currently observe, we have revisited the previously introduced concepts of privileged substructures and community-selective scaffolds and systematically searched for molecular scaffolds representing compounds active against single targets, multiple targets belonging to the same target family, or targets belonging to different families. The influence of different types of activity measurements on scaffold assignments has been determined. Furthermore, scaffold assignments have also been carried out after applying a potency threshold to exclude weakly active compounds from the comparison and address the issue of molecular selectivity. In both instances, the results were very similar indicating that single-target and single-family scaffolds display target- and family-selective tendencies, respectively. Unexpectedly large numbers of 630 unique single-target, 489 single-family, and 336 multi-family scaffolds have been identified in public domain compound data that represented relatively large numbers of compounds. Other important findings are that most of the growth in high-confidence compound activity data has been due to the evaluation of new compounds, rather than additional measurements for previously tested compounds or analog series for previously explored scaffolds. The majority of scaffolds have remained in the same category over time. Activity measurement type-dependent sets of single-target, single-family, and multi-family scaffolds are also provided as an up-to-date scaffold knowledge base.



INTRODUCTION

The molecular scaffold concept is popular in medicinal chemistry and chemoinformatics.[1] It is applied, for example, to identify preferred core structures for chemical optimization or evaluate the results of virtual compound screening calculations.[1] Following the currently most widely applied definition, scaffolds are obtained from compounds by removal of R-groups while retaining ring systems and linkers between them.[2] So defined scaffolds are often referred to as Bemis−Murcko (BM) scaffolds.[2] Although this scaffold definition has its limitations (for example, any addition of a ring to a compound creates a new scaffold), it provides a consistent and generally applicable reference frame for the analysis of ring-containing core structures in active compounds.[1] For chemoinformatics applications, this definition of molecular scaffolds is more suitable than, for example, the application of synthetically motivated molecular fragmentation approaches[3] or fragment dictionaries.[4] BM scaffolds have been systematically organized,[5,6] for example, in tree-like structures,[6] and used as a starting point for algorithmic generation of derivative scaffolds[5] or ring-based decomposition products.[6] The scaffold concept has been applied to characterize the core structure diversity of screening libraries,[7,8] organic molecules,[9] and bioactive compounds.[10] The latter analysis has also made it possible to estimate the degree of difficulty involved in the computational identification of compounds with similar activity but different scaffolds.[10] In addition, more specialized scaffold analyses have been carried out, for example, to explore structural relationships between scaffolds in compounds with different activity,[11] detect scaffolds that are recurrent in promiscuous compounds,[12,13] or identify scaffolds that frequently form activity cliffs[14] across different compound classes.[15]

In exploring structure−activity relationships on the basis of scaffolds, one typically follows a hierarchy from compounds to scaffolds and then analyzes bioactivities of compounds that a scaffold represents. This might lead to the assignment of a specific activity to a scaffold, although the compounds it represents are active, but not necessarily the scaffold itself.

Following this type of hierarchical approach, the concept of "privileged substructures' was introduced,[16] which actually predates the formal introduction of molecular scaffolds.[2] In their seminal study, Evans and colleagues observed that certain structural motifs were recurrent in compounds with activity against a family of therapeutic targets.[16] The resulting idea that core structures might exist that preferentially, or even exclusively, bind to members of a given target family has experienced a high level of interest in medicinal chemistry,[17,18] not surprisingly. However, this concept has also remained controversial because proposed privileged substructures have also been detected in ligands of other (than originally postulated) target families, often with a relatively high frequency of occurrence.[19] Such findings can be rationalized if one takes into consideration that privileged substructures were mostly put forward on the basis of medicinal chemistry observations and experience and then often subjected to frequency-of-occurrence analysis in different classes of bioactive compounds.

Received: November 14, 2012
Published: January 22, 2013

Given the partly subjective nature of the privileged substructure concept and the focus on frequency-of-occurrence analysis of knowledge-based proposals, we were interested in exploring the presence of target family directed structural motifs in a systematic manner. In 2009, we reported the results of a systematic search for potentially privileged molecular scaffolds.[20] On the basis of compound activity data, target communities were generated by assigning targets that shared active compounds to the same community and distinguished from others that had no common ligands. Target communities defined on the basis of shared active compounds were found to mostly consist of individual protein families. Then, BM scaffolds were isolated from all active compounds and a search for scaffolds was carried out that exclusively appeared in ligands of an individual target community. Ultimately, we identified a total of 206 community-selective scaffolds each representing at least five bioactive compounds available in the public domain.[20] In addition, many of these community-selective scaffolds displayed a tendency to occur in compounds that were selective for a target within a community over one or more others (on the basis of potency values of compounds represented by a given scaffold).[20] However, scaffolds that were exclusively selective for a particular target only corresponded to one or two known active compounds. Consequently, truly target-selective scaffolds could not be confidently assigned at that time.[21]

Given the substantial growth in compound data experienced over the past few years, we have searched for target family directed scaffolds and also for scaffolds representing compounds exclusively active against a single target. Compound and scaffold growth over time has been taken into account, different types of activity measurements were considered, and a potency threshold was applied for comparisons to exclude weakly active compounds from scaffold assignments. Herein, we report the results of our analysis.

MATERIALS AND METHODS

Compound Data Selection. ChEMBL[22] was used as a source of compound activity data. Beginning with ChEMBL release 2, which became available in December 2009, a scoring scheme was introduced to classify assay-to-target relationships. These scores range from 0 to 9. The smallest confidence score of 0 indicates unknown or unassigned targets, and the highest score of 9 indicates the presence of a single protein target assigned to an assay. Therefore, from ChEMBL release 2 and 14 (the current release),[23] compounds with direct interactions (i.e., target relationship type "D") against human targets with a score of 9 were extracted, hence assuring high target confidence of active compounds. These two ChEMBL releases were utilized to assess compound data and scaffold growth over time. Furthermore, two types of potency measurements were separately considered including K_i and IC_{50} values. Hence, from each of the two ChEMBL releases, two sets of compounds were extracted that were exclusively annotated with K_i or IC_{50} values, respectively (approximate measurements such as ">", "<", or "~" were excluded). If both K_i and IC_{50} values were available for a compound, it was assigned to both sets using the respective measurement(s). In total, four data sets were generated including the ChEMBL 2/K_i, 2/IC_{50}, 14/K_i, and 14/IC_{50} sets. For comparison, we also applied a potency threshold of 10 μM to all potency measurements to exclude weakly active compounds. 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. To ensure high data confidence, activity measurements and the corresponding compounds were not considered that deviated by more than 1 order of magnitude.

All targets of qualifying compounds were grouped into target families following the UniProt[23] family annotation merged with the protein classification hierarchy of ChEMBL. Only families containing at least four targets were considered.

Scaffold Classification. For each scaffold found in a data set, an activity profile was generated by assembling all target annotations of the compounds this scaffold represented. On the basis of the number of targets and families that the compounds represented by a particular scaffold were active against, the scaffold was assigned to one of three categories (Figure 1):

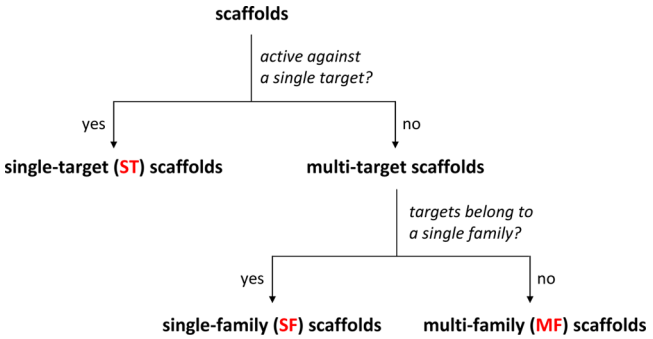


Figure 1. Scaffold classification scheme. Scaffolds are classified on the basis of activity profiles into single-target (ST), single-family (SF), and multi-family (MF) scaffolds.

(1) Scaffolds that represented compounds active against a single target were classified as *single-target* (*ST*) scaffolds.
(2) Scaffolds that corresponded to compounds active against multiple targets from the same family were considered as *single-family* (*SF*) scaffolds. This scaffold category was conceptually closely related to privileged substructures.[17,18]
(3) Scaffolds representing compounds active against multiple targets belonging to different families were designated as *multi-family* (*MF*) scaffolds.

Table 1. Data Sets[a]

number of	ChEMBL 2		ChEMBL 14	
	K_i	IC_{50}	K_i	IC_{50}
(a) based on original activity data				
compounds	11305	15933	33894 (3.00)	64441 (4.04)
targets	256	390	416 (1.63)	723 (1.85)
compound–target combinations	20533	23174	58463 (2.85)	91540 (3.95)
target families	19	28	29 (1.61)	46 (1.64)
BM scaffolds	4331	6294	11790 (2.72)	25154 (4.00)
(b) after applying the potency threshold[b]				
compounds	10959	14481	31627 (2.89)	55008 (3.80)
targets	244	366	395 (1.62)	683 (1.87)
compound–target combinations	18956	20570	53985 (2.85)	75793 (3.68)
target families	17	26	28 (1.65)	45 (1.73)
BM scaffolds	4163	5743	11381 (2.73)	21909 (3.81)

[a]For ChEMBL release 2 and 14 (current release), the number of qualifying compounds, their targets, compound–target combinations, target families, and BM scaffolds are reported for the K_i and IC_{50} data sets, respectively. In addition, "data growth factors" comparing release 2 to 14 are given in parentheses. [b]For comparison, we excluded compounds with lower than 10 μM potency.

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312–326

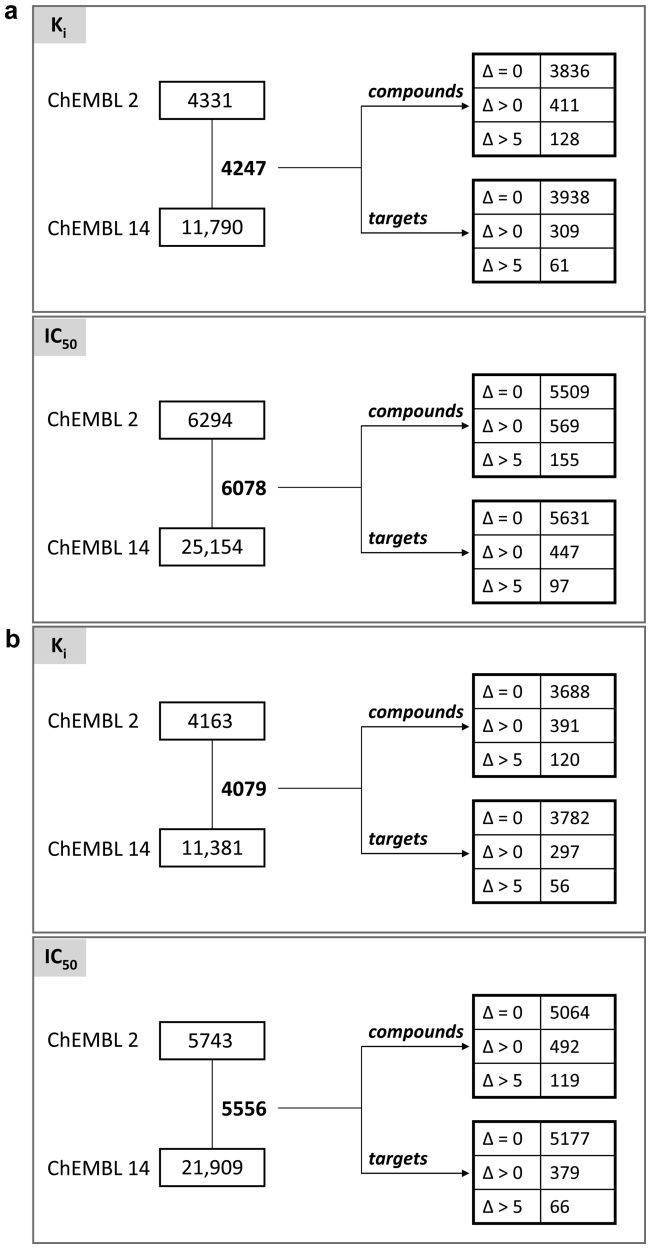


Figure 2. Scaffold comparisons. Scaffolds extracted from the K_i and IC_{50} sets of ChEMBL release 2 and 14 are compared. The results for the data sets on the basis of original activity data and data sets obtained after applying the potency threshold are reported in parts a and b, respectively. For scaffolds present in both releases (conserved scaffolds; numbers in bold), the number of corresponding compounds and targets are compared. "Δ = 0", "Δ > 0", and "Δ > 5" mean that conserved scaffolds represent the same number of compounds in release 2 and 14, more compounds in release 14 than in 2, and more than five additional compounds in release 14, respectively. The same annotations are used to compare the number of targets of compounds represented by conserved scaffolds.

Target Relationships. For SF and MF scaffolds, target relationships were systematically studied. For each data set, scaffold- and compound-based target pairs were generated that shared at least one scaffold or at least one compound represented by an SF or MF scaffold, respectively. Target relationships were analyzed in networks in which nodes represented targets and edges indicated the formation of target pairs. Network representations were drawn with Cytoscape.[25]

Table 2. Scaffold Classification[a]

		number of scaffolds			
		ChEMBL 2		ChEMBL 14	
scaffold type	number of represented compounds	K_i	IC_{50}	K_i	IC_{50}
(a) based on original activity data					
ST	total	2471	4476	6774 (2.74)	17628 (3.69)
	1	1740	3316	4766	13115
	>1; ≤5	577	885	1541	3444
	>5; ≤10	86	153	271	604
	>10	**68**	**122**	**196**	**465**
SF	total	1799	1502	4593 (2.55)	5008 (3.33)
	1	1072	924	2749	3226
	>1; ≤5	512	410	1276	1270
	>5; ≤10	127	93	309	273
	>10	**88**	**75**	**259**	**239**
MF	total	61	316	423 (6.93)	2518 (7.97)
	1	8	115	154	1000
	>1; ≤5	19	110	126	997
	>5; ≤10	13	37	44	244
	>10	**21**	**54**	**99**	**277**
(b) after applying the potency threshold					
ST	total	2402	4129	6675 (2.78)	16132 (3.91)
	1	1683	3066	4710	12012
	>1; ≤5	564	809	1501	3134
	>5; ≤10	87	139	273	560
	>10	**68**	**115**	**191**	**426**
SF	total	1709	1360	4308 (2.52)	4002 (2.94)
	1	1011	825	2554	2450
	>1; ≤5	488	375	1199	1080
	>5; ≤10	124	88	301	251
	>10	**86**	**72**	**254**	**221**
MF	total	52	254	398 (7.65)	1775 (6.99)
	1	8	93	151	643
	>1; ≤5	17	88	116	719
	>5; ≤10	9	31	39	190
	>10	**18**	**42**	**92**	**223**

[a]For ChEMBL release 2 and 14, the number of ST, SF, and MF scaffolds is reported for the K_i and IC_{50} data sets, respectively. In addition, for each scaffold category, the number of scaffolds with different numbers of compounds is provided. For example, "1" indicates that a scaffold represents a single compound, and ">10" that a scaffold represents more than 10 compounds (the number of these scaffolds is shown in bold). In addition, data growth factors are reported in parentheses.

Scaffold Calculations. The generation of BM scaffolds[2] from compounds comprising each data set and the scaffold classification analysis was systematically carried out with in-house generated Perl routines. From compounds, BM scaffolds were obtained by removing all acyclic substituents and retaining ring systems and linkers between them.

RESULTS AND DISCUSSION

Activity-Oriented Scaffold Analysis. The assignment of scaffolds to targets and target families requires the use of compound activity data but does not depend on the availability of confirmed inactive compounds. This is the case because the focal point of the analysis is the assessment of activity annotations and their distribution at the level of molecular scaffolds, which leads to a target-based classification of scaffolds. As such, the analysis does not claim to identify scaffolds that would be truly

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

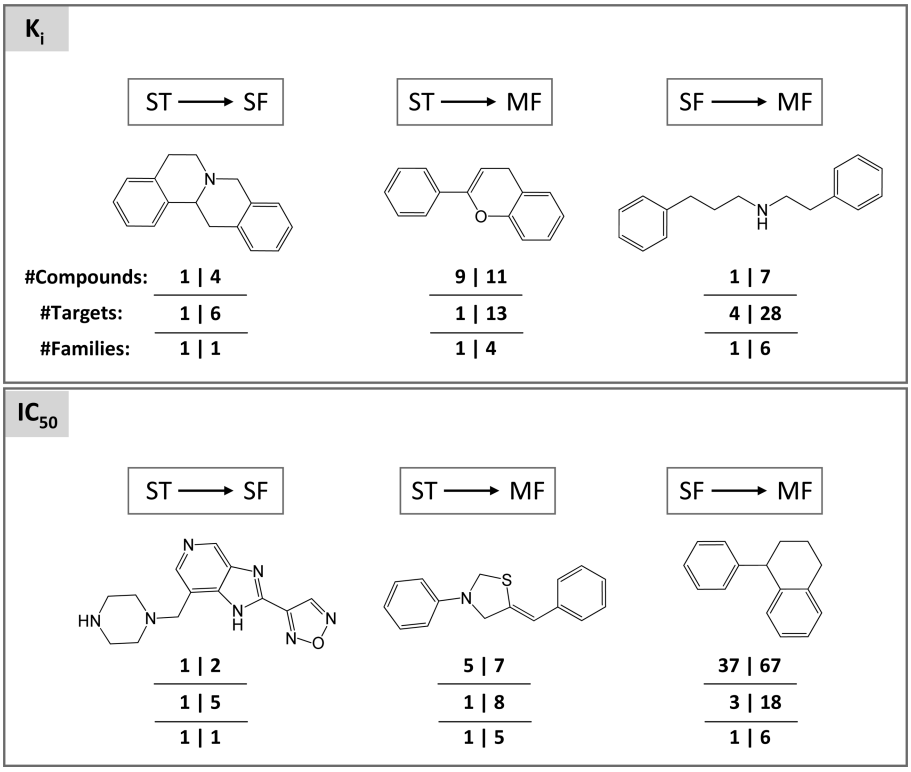


Figure 3. Scaffold category transitions. For scaffolds present in both ChEMBL release 2 and 14, their assignments were compared. From data sets obtained after applying the potency threshold, representative scaffolds were taken that were assigned to different categories (representing transitions from ST to SF or MF and from SF to MF scaffolds). For each scaffold, the number of corresponding compounds, targets, and target families is given.

Table 3. Scaffolds Representing More than 10 Compounds[a]

(a) on the basis of original potency data						
	K_i			IC_{50}		
number of	ST	SF	MF	ST	SF	MF
	ChEMBL 2					
scaffolds	68	88	21	122	75	54
compounds	1146	1838	739	2342	1530	1670
targets	29	90	111	69	122	192
target families	9	11	17	22	19	24
	ChEMBL 14					
scaffolds	196	259	99	465	239	277
compounds	3685	5372	3229	8434	4782	8632
targets	62	162	238	148	250	478
target families	19	18	29	36	29	43
(b) after applying the potency threshold						
	K_i			IC_{50}		
number of	ST	SF	MF	ST	SF	MF
	ChEMBL 2					
scaffolds	68	86	18	115	72	42
compounds	1142	1782	657	2125	1550	1230
targets	31	93	92	68	110	153
target families	9	12	14	21	17	22
	ChEMBL 14					
scaffolds	191	254	92	426	221	223
compounds	3586	5276	2913	7514	4409	6684
targets	60	157	209	145	238	409
target families	17	18	27	38	26	43

[a]For all ST, SF, and MF scaffolds that represent more than 10 compounds, the number of compounds, targets, and target families is reported for K_i and IC_{50} sets from ChEMBL 2 and 14.

target- or family-specific. In fact, making such assignments would not be possible until all active compounds would be tested against all possible targets, representing the rather elusive ultimate goal of chemogenomics. By contrast, assessing—and following over time—activity trends and target distributions at the level of scaffolds, in light of ideas underlying, for example, the concepts of privileged substructures or target community-selective scaffolds, is considered informative, given the very large amounts of compound activity data that are already available at present.

Data Set Design. By comparing data sets extracted from ChEMBL release 2 and 14, we were able to determine the influence of compound data growth on scaffold distributions and target relationships. Many new active compounds were reported over the past few years. Some of the data growth might still be attributed to compounds that were originally published in earlier years but have only recently been added to ChEMBL. However, the time between an original publication of compound data and its incorporation into the database does not affect our analysis strategy. Furthermore, we separated compound data according to the type of activity measurements. K_i values represent equilibrium constants that are comparable across different assays, whereas IC_{50} values are assay-dependent and more approximate in nature. Separate consideration of these different types of activity measurements was motivated by recent findings that the apparent increase in compound promiscuity accompanying data growth in ChEMBL was activity measurement-dependent and mostly due to growth in IC_{50} data, rather than equilibrium constants.[26] Hence, we concluded it was important to evaluate scaffold assignments and target relationships by separately considering these different types of activity measurements.

Scaffold Selectivity. For comparison, we also excluded compounds from further consideration that had a lower potency

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

than 10 μM. Thereby, we focused the analysis on scaffold selectivity because scaffolds that represented compounds highly potent against a given target and weakly potent against one or more others, hence displaying target selectivity, were no longer classified as SF or MF scaffolds, but rather ST scaffolds.

Data Set Composition. Table 1a and b report the composition of the original data sets we extracted from ChEMBL release 2 and 14 and of the corresponding data sets when the potency threshold was applied, respectively, and reflect very significant growth in compound data over the past three years. The removal of compounds with lower than 10 μM potency did not significantly change the data set composition. In both cases, the number of compounds with high-confidence target annotations selected for our analysis nearly tripled for K_i and quadrupled for IC_{50} measurements from ChEMBL release 2 to 14. In total, more than 25 000 and 95 000 qualifying compounds were obtained from ChEMBL 2 and 14, respectively, on the basis of original potency data. Growth rates comparable to those of compounds were observed for compound−target combinations. In addition, there was substantial growth in the number of targets. Compared to ChEMBL 2, the number of targets in ChEMBL 14 increased from 256 to 416 (K_i) and 390 to 723 (IC_{50}). In addition, the number of target families increased from 19 to 29 (K_i) and from 28 to 46 (IC_{50}).

Scaffold Distribution. There also was a dramatic increase in the number of BM scaffolds. From ChEMBL 2 to 14, the number of scaffolds that qualified for our analysis nearly tripled for K_i and quadrupled for IC_{50} measurements. From ChEMBL 14, we obtained a total of 11 790 and 25 154 BM scaffolds for K_i and IC_{50} data, respectively. Thus, the scaffold growth rates essentially paralleled growth rates for compounds and compound−target combinations. This finding indicated that most of the growth in compound activity data was due to assay results of new compounds, rather than additional activity data for existing molecules or analog series for previously explored scaffolds, which was confirmed by analyzing compound and target distributions for scaffolds that were conserved in ChEMBL 2 and 14, as reported in Figure 2a and b. In Figure 2a, of 4247 conserved scaffolds with K_i data, 3836 represented the same number of compounds and for 3938 scaffolds there was no increase in the number of targets the corresponding compounds were active against. Only 128 scaffolds (~3%) represented more than five additional compounds in ChEMBL 14 compared to 2 and only 61 scaffolds (~1.4%) had more than five additional target annotations. For conserved scaffolds with IC_{50} data, equivalent trends were observed. Of all 6078 scaffolds, 5509 and 5631 showed no increase in the number of compounds and targets, respectively. Similar trends were observed for the data sets when the potency threshold was applied (Figure 2b).

Scaffolds, Target Families, and Individual Targets. Given the dramatic increase in the number of available bioactive compounds, target annotations, and BM scaffolds observed over the past three years, we set out to systematically identify scaffolds that were active against single targets or two or more members of an individual target family. The latter scaffolds are related to community-selective scaffolds and privileged substructures. In addition, we searched for scaffolds that were active across different target families.

Scaffold Classification. In ChEMBL 14, we identified large numbers of scaffolds for each of the three categories, as reported in Table 2. For the original K_i and IC_{50} data, a total of 2008 and 4513 ST, 1844 and 1782 SF, and 269 and 1518 MF scaffolds were found, respectively, that represented multiple compounds.

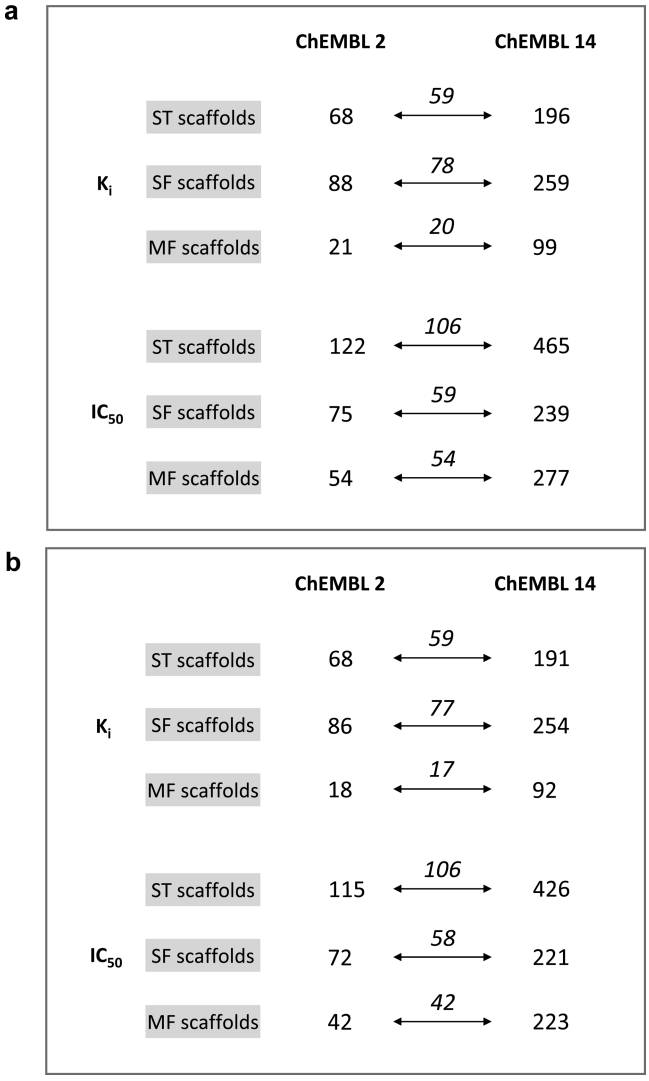


Figure 4. Comparison of frequently observed scaffolds. ST, SF, and MF scaffolds representing more than 10 compounds are compared. In each case, the number of conserved scaffolds is reported in italics. The results for data sets based on original activity data and data sets obtained after applying the potency threshold are reported in parts a and b, respectively.

When compared to ChEMBL 2, the number of scaffolds in each category increased by a factor of 3−5, dependent on the type of activity measurements. Overall, significantly more ST and MF scaffolds were obtained on the basis of IC_{50} than on the basis of K_i data. The largest relative increase was observed for MF scaffolds. Here, K_i data yielded 269 of these scaffolds representing multiple compounds whereas IC_{50} data yielded 1518 scaffolds. By contrast, the numbers of SF scaffolds were comparable for alternative activity measurements. Hence, the global distribution of MF scaffolds and, to a lesser extent, ST scaffolds was dominated by scaffolds identified on the basis of IC_{50} data. Comparable observations were made when the potency threshold was applied.

Furthermore, for scaffolds conserved in ChEMBL 2 and 14, their scaffold categories were compared. Identical assignments were found for more than 95% of the conserved scaffolds. Exemplary scaffolds for which a transition from the ST to SF or MF and from SF to MF categories was observed are shown in Figure 3.

Preferred scaffolds. Table 2 also reports that the majority of ST and SF scaffolds in ChEMBL 14 were only represented by

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

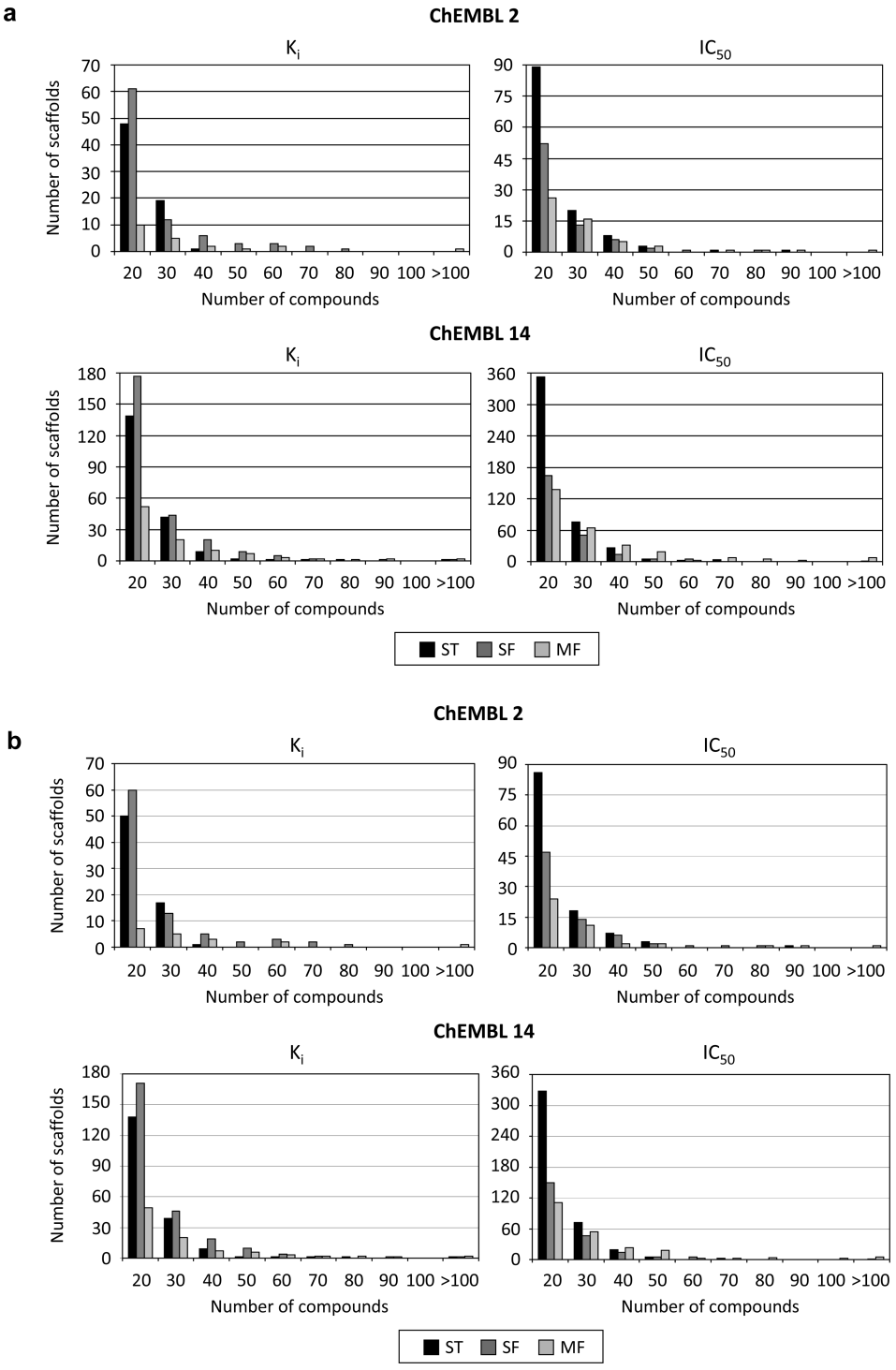


Figure 5. Compound distributions for different scaffold types. For K_i and IC_{50} sets from ChEMBL release 2 and 14 and scaffolds representing more than 10 compounds, the number of compounds is reported for ST (black), SF (dark gray), and MF (light gray) scaffolds. The results for data sets based on original activity data and data sets obtained after applying the potency threshold are reported in parts a and b, respectively.

single compounds. In these cases, it was impossible to classify scaffolds in a meaningful way. Hence, single-compound scaffolds were not further considered. However, ST, SF, and MF scaffolds that represented more than 10 compounds were also detected, as detailed in Table 3. These scaffolds were considered preferred candidates for further analysis, given the relatively large number of corresponding compounds. As reported in Table 3a, 88 and 75 SF scaffolds were identified in ChEMBL 2 on the basis of K_i and IC_{50} data, respectively. The number of scaffolds in this subset was roughly comparable to the set of 206 community-selective scaffolds that were identified prior to the first release of ChEMBL in all public domain compounds available at that time, taking both types of activity measurements into account.[20] In ChEMBL 14, large numbers of ST, SF, and MF scaffolds were found to represent more than 10 compounds. For K_i and IC_{50} data, 196 and 465 ST, 259 and 239 SF, and 99 and 277 MF scaffolds were identified, respectively. Figure 4a shows that most, but not all of the ST, SF, and MF scaffolds contained in ChEMBL 2 appeared

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

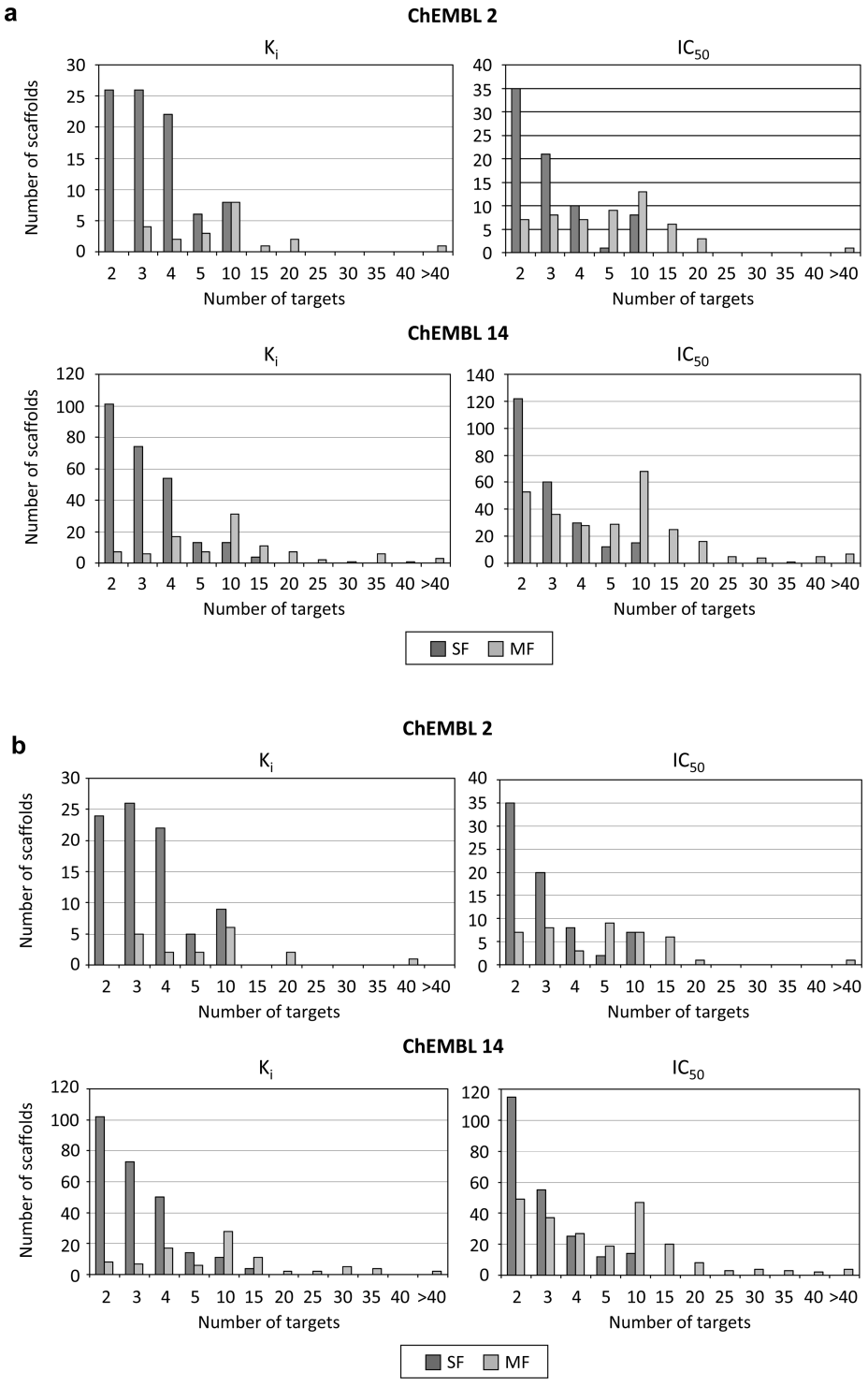


Figure 6. Target distributions for SF and MF scaffolds. For scaffolds representing more than 10 compounds, the number of targets is reported that compounds represented by SF (dark gray) and MF (light gray) scaffolds are active against. The results for data sets based on original activity data and data sets obtained after applying the potency threshold are reported in parts a and b, respectively.

in the same category in release 14. In 5 of 6 cases (except MF scaffolds on the basis of IC_{50} data), between 1 and 16 of these scaffolds contained in ChEMBL 2 were assigned to a different category (or no longer found in release 14). Table 3b and Figure 4b report comparable results for the corresponding ST, SF, and MF scaffolds extracted from the potency-adjusted data sets.

Compound and Target Distribution. Figure 5 shows that most of the preferred ST, SF, and MF scaffolds represented 10−30 compounds. In addition, small numbers of scaffolds representing much larger numbers of compounds were also found. Furthermore, as shown in Figure 6, most SF scaffolds were active against 2−4 targets within a family, whereas MF scaffolds were often active against 5−10 targets. These targets mostly belonged to two to three different families (Figure 7), although scaffolds were also identified whose compounds were active against 10 or more different families.

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

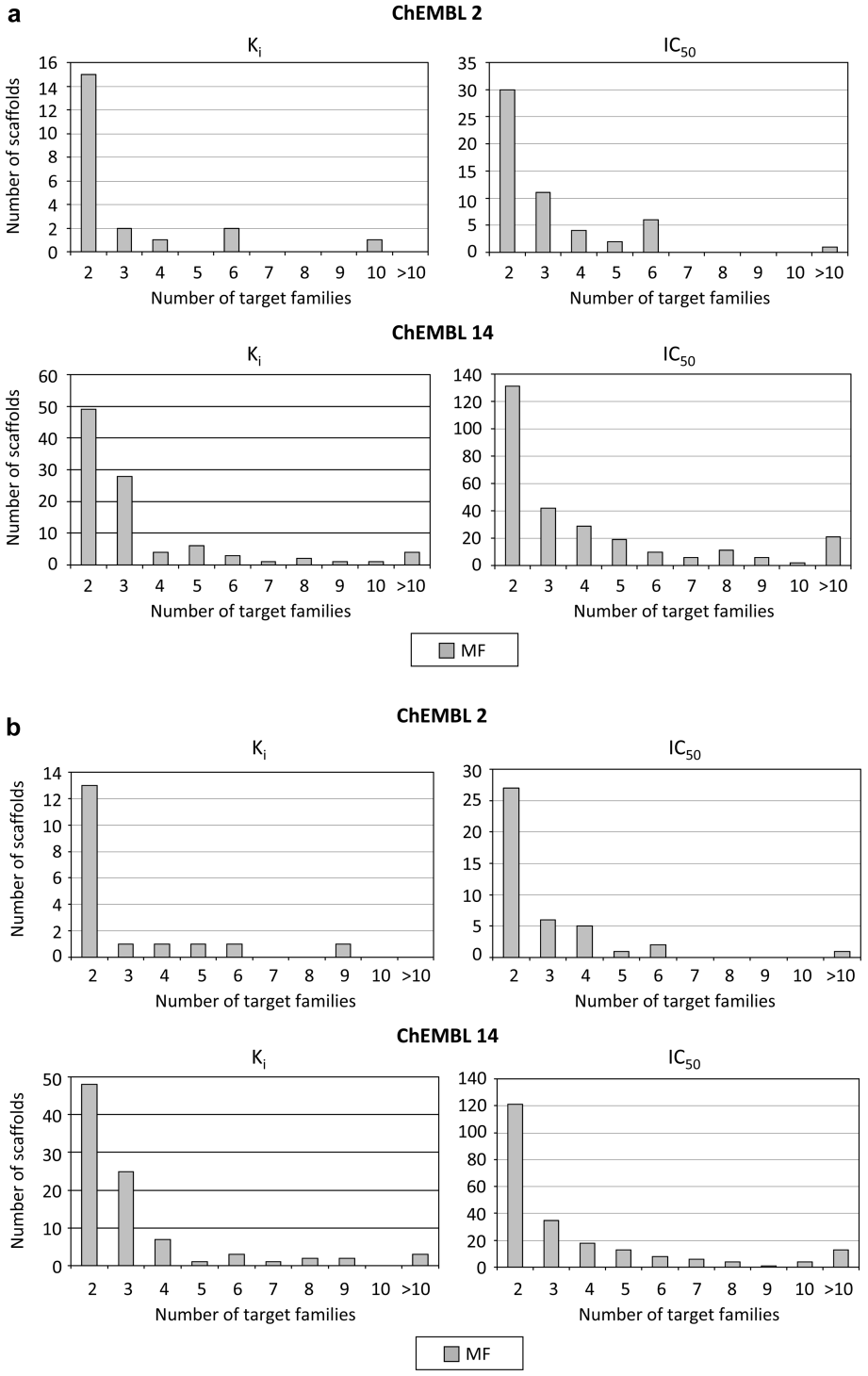


Figure 7. Target family distributions of MF scaffolds. For scaffolds representing more than 10 compounds, the number of target families is reported that compounds represented by MF scaffolds (light gray) are active against. The results for data sets based on original activity data and data sets obtained after applying the potency threshold are reported in parts a and b, respectively.

Representative Scaffolds. In Figures 8 and 10, examples of ST, SF, and MF scaffolds from the original K_i and IC_{50} sets are shown, respectively. The same sets of scaffolds from the potency-adjusted K_i and IC_{50} sets are shown in Figure 9 and Figure 11, respectively. Some structural trends were observed. ST and SF scaffolds often varied in their size and chemical complexity, but MF scaffolds were typically smaller and more generic, consistent with the predominantly promiscuous nature of compounds they represented. In ChEMBL 14, the K_i- and IC_{50}-based MF scaffold sets consisted of 99 and 277 scaffolds and shared 40 of them.

Single-Family Scaffolds. In the context of our analysis, SF scaffolds were of particular interest, given their intrinsic relationship to privileged substructures and community-selective scaffolds. As reported in Table 3, there was substantial growth in SF scaffolds from ChEMBL 2 to 14. A total of 259 K_i-based SF scaffolds were found in ChEMBL 14 that represented 5372 compounds active against 162 targets. K_i-based SF scaffolds were

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

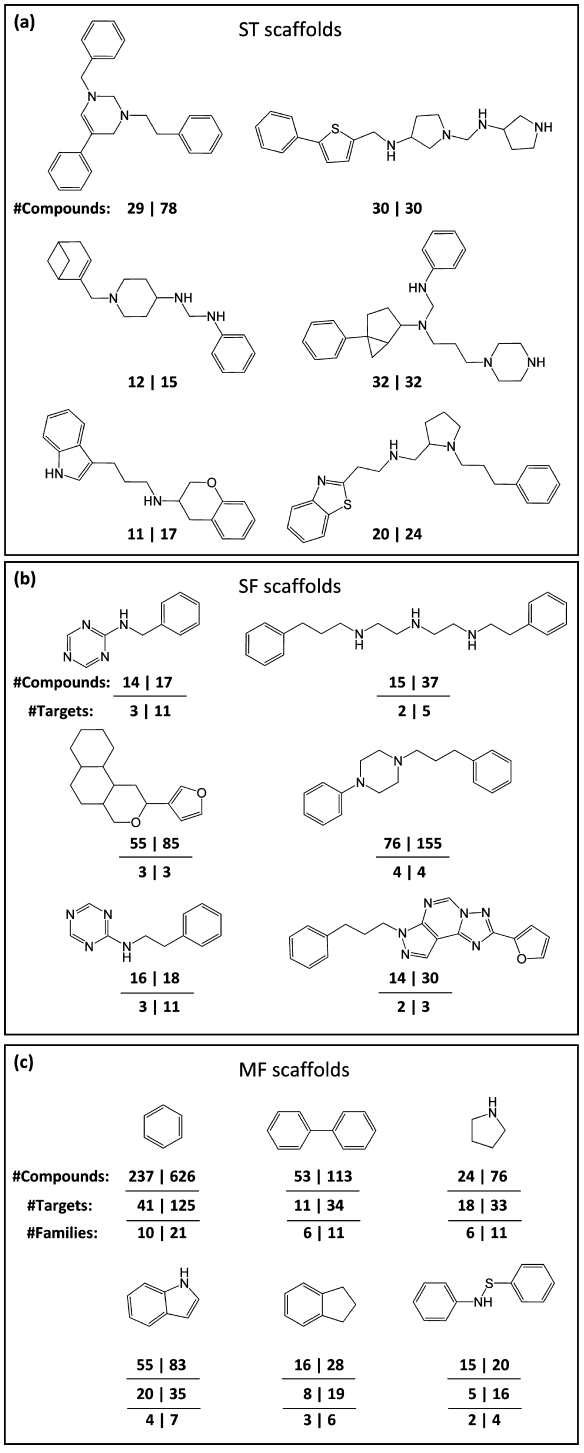


Figure 8. Conserved scaffolds from original K_i sets. Shown are exemplary scaffolds from the original K_i sets of ChEMBL 2 and 14. (a) For each ST scaffold, the number of compounds it represents in ChEMBL 2 and 14 is reported. For example, "29|78" indicates that the scaffold represents 29 compounds in ChEMBL 2 and 78 compounds in 14. (b) For each SF scaffold, compounds are reported as in part a and, in addition, the number of targets of these compounds is provided in a second annotation layer. For example, "3|11" means that the compounds that are active against 3 targets in CHEMBL 2 and 11 targets in 14. (c) For each MF scaffold, compounds and targets are reported as in part b and the number of target families is given in a third layer. For example, the benzene ring represents compounds active against targets belonging to 10 families in ChEMBL 2 and 21 families in 14.

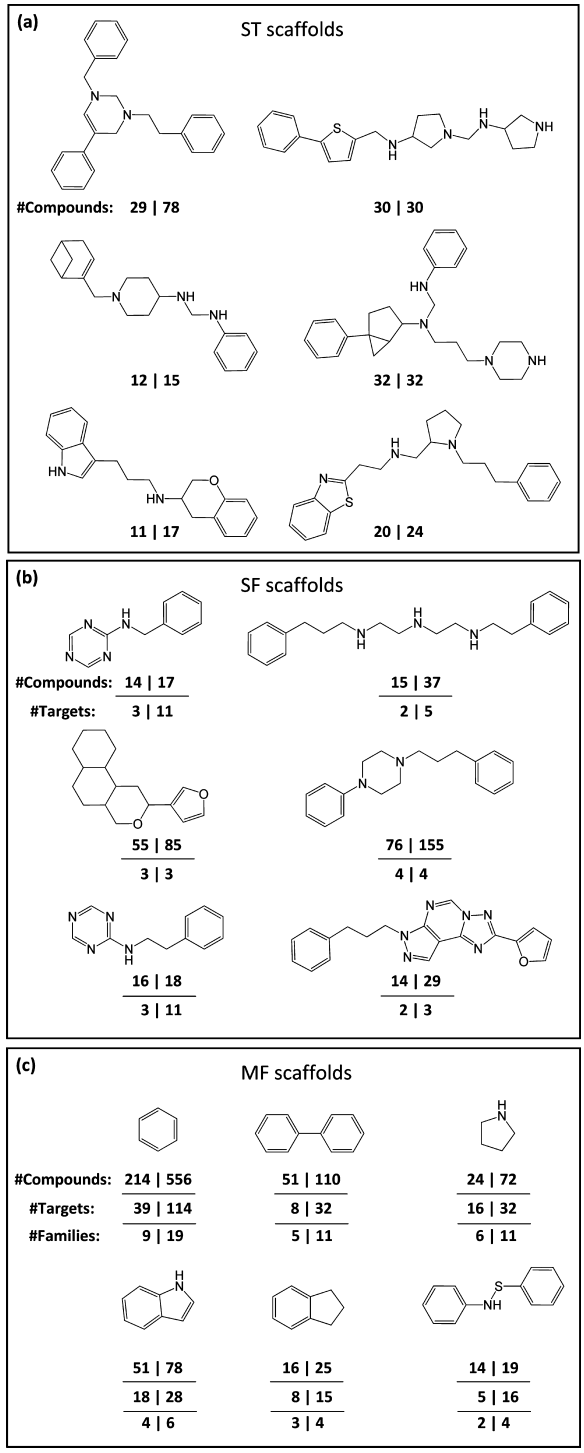


Figure 9. Conserved scaffolds from potency-adjusted K_i sets. Shown are the same scaffolds as in Figure 8 with compound, target, and family annotations after applying the potency threshold.

identified for a total of 18 target families. Furthermore, a comparable number of 239 IC_{50}-based SF scaffolds was found. These scaffolds represented 4782 compounds active against 250 targets belonging to 29 different families. Hence, although IC_{50}-based SF scaffolds represented fewer compounds than K_i-based SF scaffolds, they covered more targets and target families. Interestingly, both the K_i- and IC_{50}-based SF scaffold sets each contained only 13 previously identified community-selective scaffolds.[20] Hence, SF scaffolds identified herein provide a much extended knowledge

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

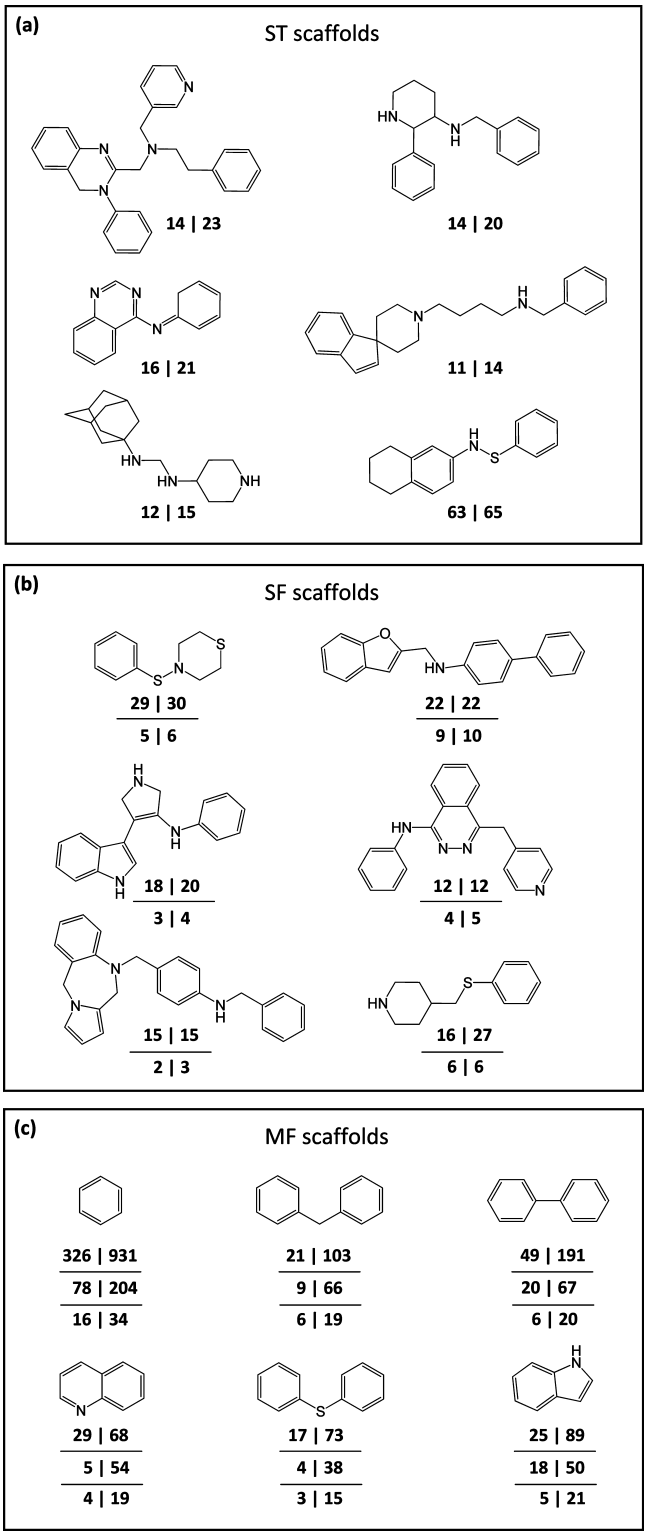


Figure 10. Conserved scaffolds from original IC_{50} sets. Shown are exemplary scaffolds from the IC_{50} sets of ChEMBL 2 and 14: (a) ST, (b) SF, and (c) MF scaffolds. The representation is according to Figure 8.

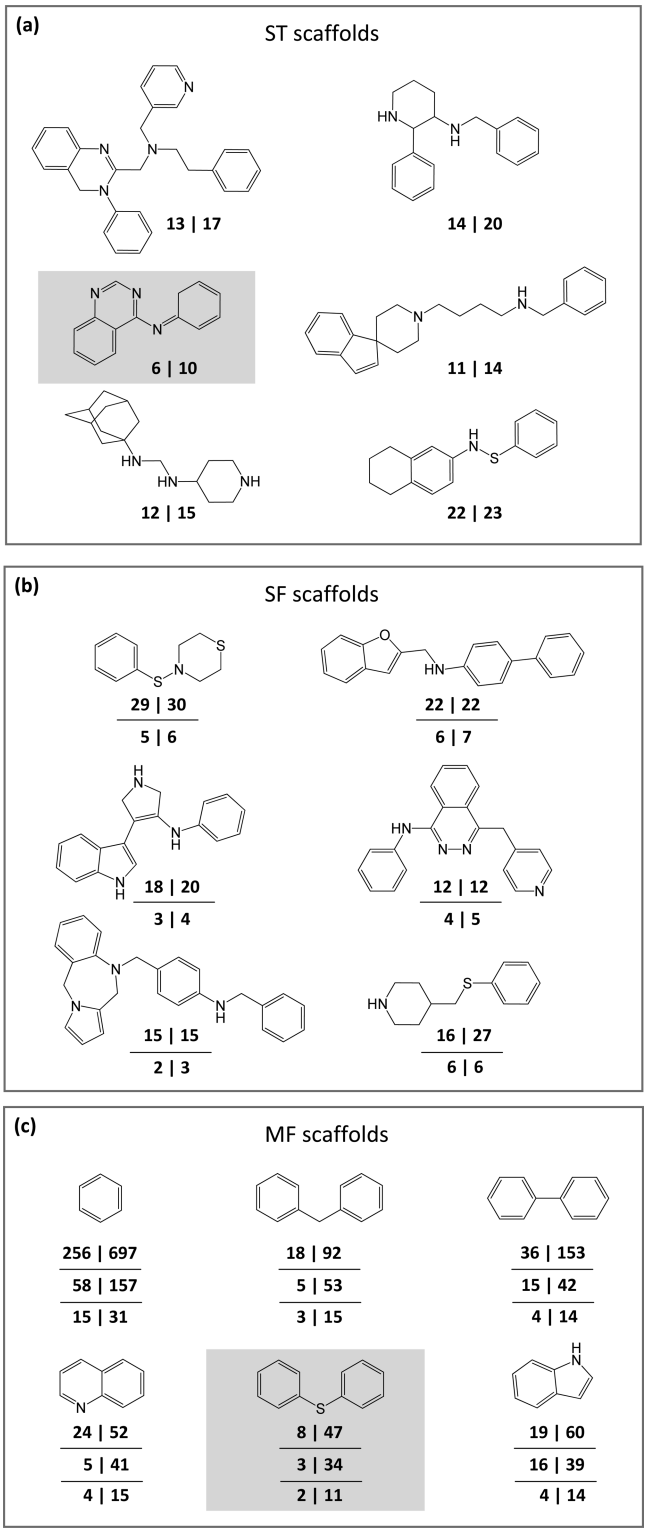


Figure 11. Conserved scaffolds from potency-adjusted IC_{50} sets. Shown are the same scaffolds as in Figure 10 with compound, target and family annotations after applying the potency threshold. In addition, two scaffolds displayed on a gray background represented 10 or fewer compounds after applying the potency threshold.

base for the exploration of target family selectivity of scaffolds and the compounds they represent. Moreover, the activity measurement dependence must also be taken into account. The K_i- and IC_{50}-based SF scaffold sets shared only nine scaffolds. Thus, a total of 463 previously unobserved scaffolds with target family selectivity potential are currently available, with on average 20.5 compounds per scaffolds.

Single-Target Scaffolds. ST scaffolds were also of high interest, because they have not been assigned previously. In our current analysis, we have identified a large number of ST scaffolds

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

Table 4. Target Relationships on the Basis of Original Potency Data[a]

			ChEMBL 2		ChEMBL 14	
number of			K_i	IC_{50}	K_i	IC_{50}
SF + MF scaffolds			109	129	358	516
scaffold-based	≥1 scaffold	target pairs	1406 (394|1012)	4042 (765|3277)	10474 (1232|9242)	28042 (2241|25801)
		targets	152	243	278	528
		families	20	24	26	42
	≥5 scaffolds	target pairs	36 (36|0)	29 (27|2)	556 (319|237)	986 (269|717)
		targets	30	24	83	157
		families	8	6	10	29
compound-based	≥1 compound	target pairs	332 (304|28)	685 (559|126)	1596 (896|700)	2983 (1381|1602)
		targets	146	215	257	476
		families	15	23	25	41
	≥5 compounds	target pairs	169 (169|0)	161 (151|10)	543 (457|86)	664 (486|178)
		targets	97	101	176	269
		families	14	16	23	33

[a]For SF and MF scaffolds, the number of scaffold- and compound-based target pairs is reported. For scaffold-based target relationships, the number of target pairs that share at least one or five scaffolds is given. For example, "(394|1012)" means that a total of 394 intrafamily and 1012 interfamily target pairs are formed sharing at least one SF or MF scaffold. In addition, the number of target pairs that share at least one or five compounds is provided. For each relationship category, the number of involved targets and target families is also reported.

on the basis of rapidly growing compound activity data. As reported in Table 3, a total of 196 and 465 ST scaffolds were detected in the K_i- and IC_{50} compound data sets of ChEMBL 14. Each of these ST scaffolds represented more than 10 compounds. Similar to the observations made for SF scaffolds, there was only very limited overlap between the K_i- and IC_{50}-based ST scaffold sets that shared only 31 scaffolds. By contrast, there were more than twice as many IC_{50}- than K_i-based ST scaffolds. However, on average, K_i- and IC_{50}-based ST scaffolds were represented by 18.8 and 18.1 compounds per scaffold, respectively, which provides a much improved basis for the consideration of ST scaffolds (compared to earlier years when only very few compounds per candidate scaffold were available). As more activity data for compounds representing ST scaffolds will likely become available in the future, the statistical likelihood of ST scaffolds will be reduced. Similar trends can be expected when compound numbers per scaffold further increase. This is illustrated by the increase in the average number of compounds per ST, SF, and MF scaffold from 18.5 and 20.5 to 31.9 determined on the basis of activity data available in ChEMBL 14. However, with 630 unique ST scaffolds, there is a large number of candidate scaffolds available for the further exploration of target specificity.

Ligand-Based Target Relationships. We also explored the formation of ligand-based target relationships for SF and MF scaffolds by generating all possible target pairs on the basis of each scaffold set from original potency data. Target pairs were generated by determining scaffolds and active compounds that were shared by targets. Table 4 reports all resulting target pairs. Regardless of whether scaffolds or compounds were considered, there was a very significant increase in the number of target pairs in ChEMBL 14 compared to 2, for both the K_i and IC_{50} sets. For example, when at least one shared scaffold was required to establish a target pair, the increase was approximately 7-fold, yielding 10 474 and 28 042 target pairs for the K_i and IC_{50} sets of ChEMBL 14, respectively. Furthermore, when at least five shared scaffolds were required for pair formation, the increases were even more significant, i.e., from 36 to 556 pairs for K_i and from 29 to 986 pairs for IC_{50} data. For compound-based pairings, the absolute numbers of target pairs were smaller than for scaffolds, because different compounds were usually represented by the same scaffold. In addition, increases in the number of target pairs were of smaller magnitude in this case (with growth factors of approximately three to five). Moreover, the number of target pairs significantly decreased when the minimally required number of shared scaffolds or compounds was increased from one to five.

On the basis of target pairs that shared at least five scaffolds or at least five compounds, target relationships were monitored in network representations. Figure 12a and b show networks generated from scaffold- and compound-based target pairs for K_i data, respectively, Figure 12c and d show the corresponding networks for IC_{50} data, and Figure 12e summarizes all target families that were involved. For ChEMBL 2, small but well-defined target communities were formed for all activity data that mostly consisted of individual target families. By contrast, for ChEMBL 14, a large central network component was formed in each case that involved targets from different families. Here, there was a notable difference between K_i and IC_{50} data. We observed a much larger degree of target promiscuity when IC_{50} data were considered, consistent with earlier observations made at the level of compounds,[26] rather than target relationships. Figure 12a contained 319 intrafamily pairs for ChEMBL 14 and 237 interfamily pairs. These 237 pairs represented 14 family relationships involving eight different families. By contrast, the corresponding scaffold-based network for IC_{50} data in Figure 12c contained only 269 intrafamily, but 717 interfamily pairs. These 717 pairs represented a total of 136 family relationships involving 28 different target families. In these networks, well-defined target communities outside the large central network component were no longer found. The compound-based networks for ChEMBL 14 still revealed the formation of other small target communities that mostly corresponded to different target families. The K_i-based network in Figure 12b contained 457 intrafamily pairs and only 86 interfamily pairs, which corresponded to only four family relationships between four different families. These families were involved in the formation of the largest community. By contrast, the corresponding IC_{50}-based network in Figure 12d contained 178 interfamily pairs that represented 40 family relationships involving 23 different target families, leading to the formation of a large and heterogeneous central community. Taken together, these findings revealed a substantial increase in

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

the number of ligand-based relationships between different targets in ChEMBL over the past three years, especially for IC_{50} data. In ChEMBL 14, the assignment of community-selective scaffolds would not be very informative, given the dominance of a single large and heterogeneously populated target community, which reinforced our current strategy to study the selectivity of scaffolds at the level of all individual target families.

Scaffold Distribution for Targets. In Table 5, numbers of active scaffolds found for different targets are reported. In ChEMBL 14, 50 (K_i data) and 90 (IC_{50}) targets are found for which only a single active scaffold is available. By contrast,

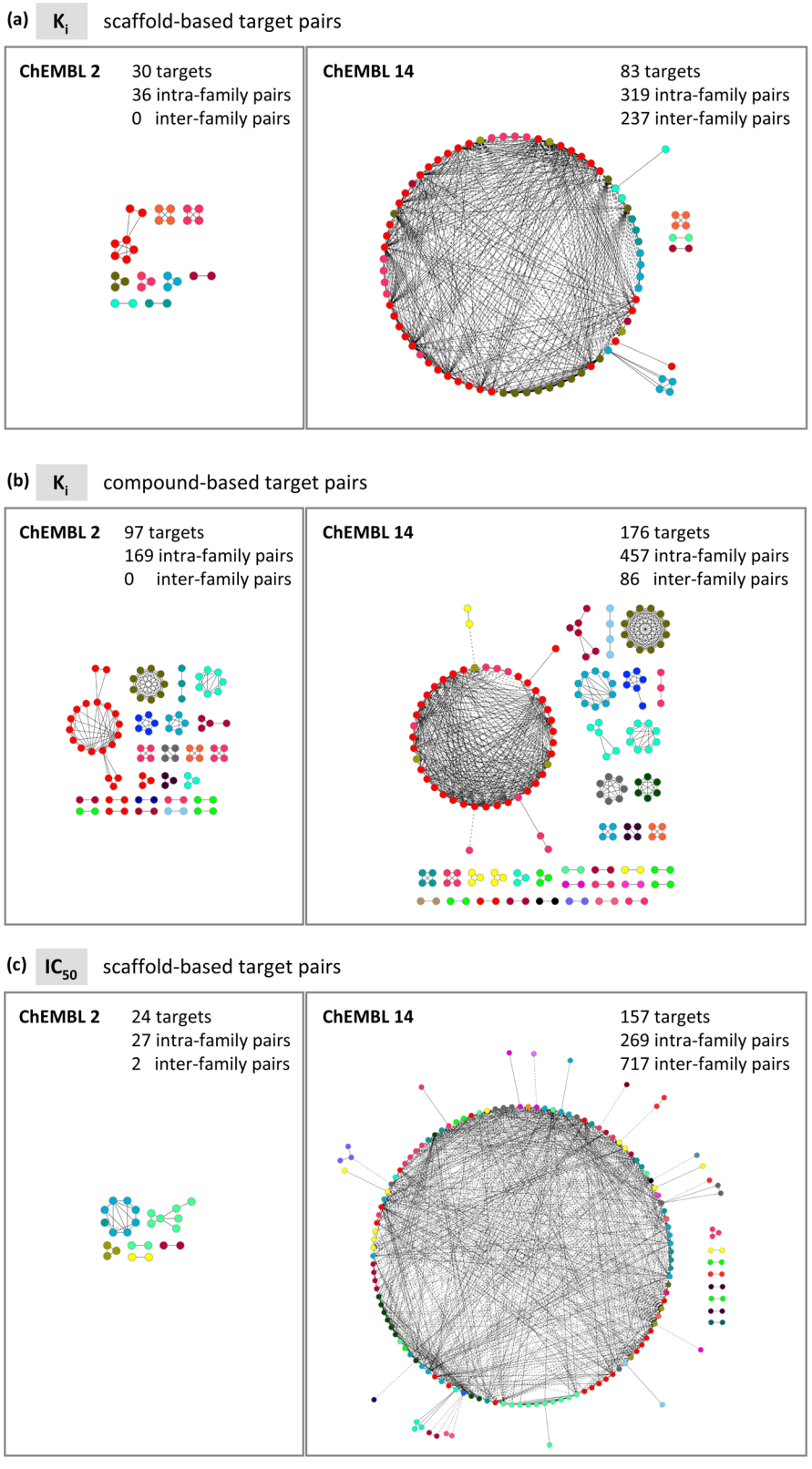


Figure 12. continued

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

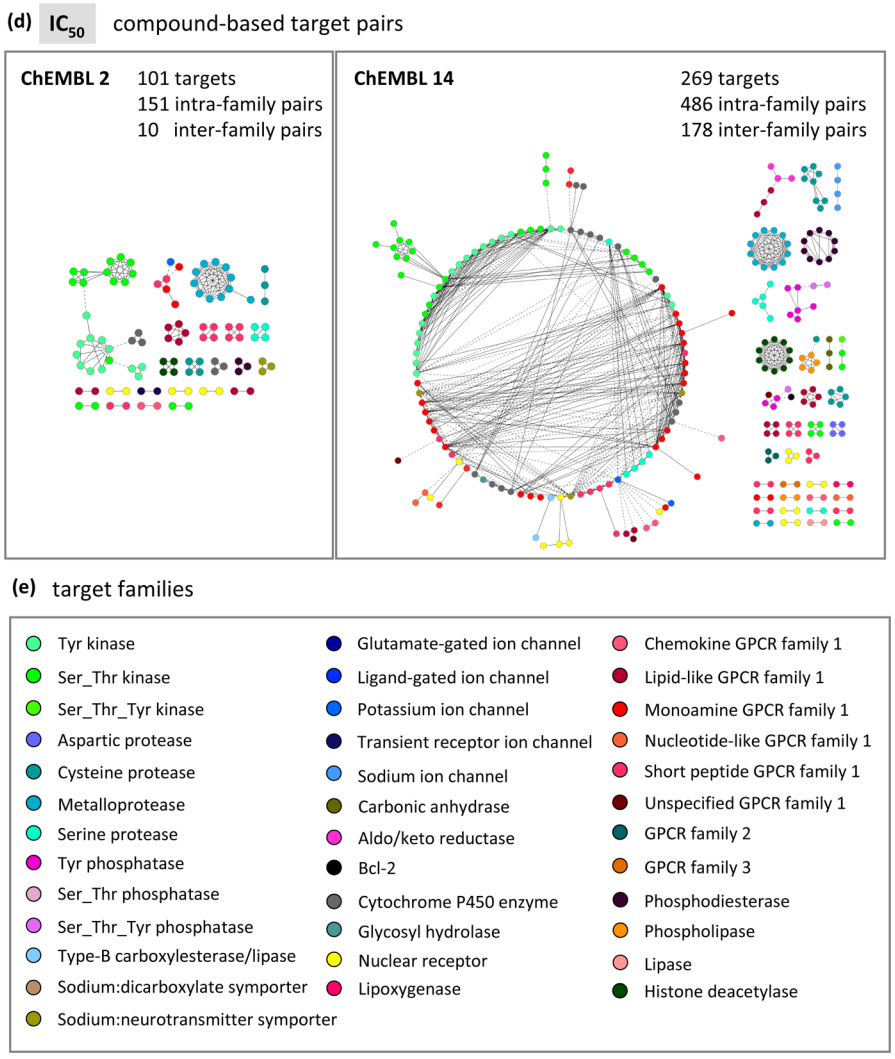


Figure 12. Target networks. For the original K_i sets, (a) scaffold- and (b) compound-based target networks are shown. For the original IC_{50} sets, corresponding network representations are shown in parts c and d, respectively. Nodes represent targets and edges indicate target pairs that share at least five SF and/or MF scaffolds or at least five compounds containing SF or MF scaffolds. Intra- and interfamily target pairs are indicated by solid and dashed edges, respectively. Target families are listed and color-coded in part e.

Table 5. Scaffold Distribution for Targets[a]

	number of targets			
	ChEMBL 2		ChEMBL 14	
number of scaffolds per target	K_i	IC_{50}	K_i	IC_{50}
1	36	38	50	90
>50	45	51	100	217
>100	19	20	57	122
>200	9	3	28	56
total	256	390	416	723

[a]For ChEMBL release 2 and 14, the number of targets is reported for the K_i and IC_{50} data sets, respectively, for which a single scaffold or more than 50, 100, or 200 scaffolds were available.

there are 28 (K_i data) and 56 (IC_{50}) targets for which already more than 200 scaffolds have been reported. In ChEMBL 2, the corresponding numbers of highly explored targets are nine and three, respectively. Hence, there has been significant growth in the number of targets that are chemically intensely explored. At the same time, new targets are subjected to chemistry efforts for which only very limited active compound information is currently available. The growth in little and heavily explored targets mostly results from IC_{50} data.

Target Distribution for Scaffolds. Analogously, the number of targets scaffolds are annotated with can be monitored. The results are reported in Table 6. As already discussed, there are large numbers of scaffolds available that are currently reported

Table 6. Target Distribution for Scaffolds[a]

	number of nonbenzene scaffolds			
	ChEMBL 2		**ChEMBL 14**	
number of targets per scaffold	K_i	IC_{50}	K_i	IC_{50}
1	2471	4476	6774	17628
≥5	196	172	502	880
≥10	18	27	154	206
≥20	1	3	31	53
total	4330	6293	11789	25153

[a]For ChEMBL release 2 and 14, the number of nonbenzene scaffolds is reported for the K_i and IC_{50} data sets, respectively, that were found in a single target set or in at least 5, 10, or 20 target sets.

dx.doi.org/10.1021/ci300616s | J. Chem. Inf. Model. 2013, 53, 312−326

to be active against only a single target (i.e., ST scaffolds by definition), with a maximum of more than 17 000 scaffolds for IC_{50} sets from ChEMBL 14. However, there is a rapid decline in scaffold numbers with increasing numbers of target annotations. In ChEMBL 2, there were only one and three nonbenzene scaffolds found in the K_i and IC_{50} sets, respectively, which were active against at least 20 different targets. In ChEMBL 14, the corresponding numbers are 31 and 53 scaffolds, respectively. Thus, highly promiscuous scaffolds continue to be rare, whereas literally thousands of scaffolds with currently only rudimentary activity profiles are available. Thus, as discussed above, the current pool of ST scaffolds provides a very substantial knowledge base for further chemical exploration of drug targets.

CONCLUDING REMARKS

Herein, we have systematically identified scaffolds representing compounds active against individual targets as well as single and multiple target families. The analysis was motivated by the large-magnitude growth in compound activity data in ChEMBL 14 compared to ChEMBL 2. Overall data growth in ChEMBL is monitored in Figure 13. While the numbers of compounds and targets increase in a near-linear manner over time, there is a strong (nearly exponential) growth in the number of activity annotations beginning with ChEMBL release 10. When stringent compound selection criteria were applied, as reported herein, it was found that much of the growth in compound activity data resulted from the experimental evaluation of new compounds. On the basis of K_i and IC_{50} data, which we separately considered, largely nonoverlapping sets of ST, SF, and MF scaffolds were obtained. However, the use of IC_{50} data resulted in many more ST and MF scaffolds than the use of K_i data. Given the approximate nature of IC_{50} measurements, K_i-based scaffold sets provide a more conservative assessment of target annotations, especially those leading to the assignment of MF scaffolds. We have also shown that the use of IC_{50} data dramatically increased the number of ligand-based relationships between targets or target families compared to those observed for K_i data. MF scaffolds present prime candidates for the evaluation of molecular promiscuity. Each MF scaffold is already confirmed at present to represent multiple promiscuous compounds. With further data growth, the number of MF scaffolds is expected to increase.

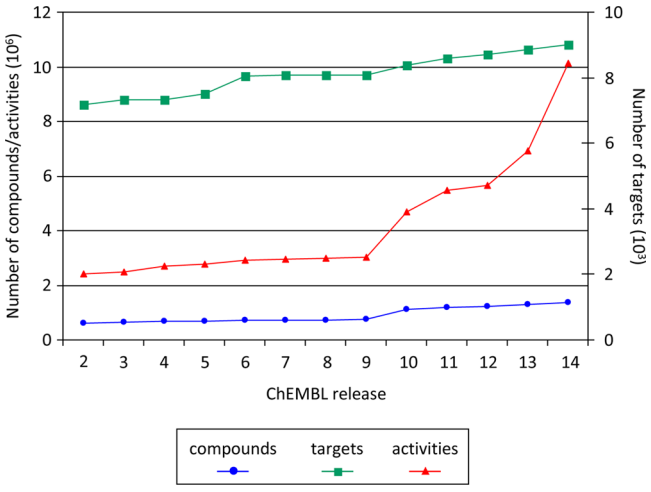


Figure 13. Data growth in ChEMBL. Shown is the growth in the total number of compounds (blue), targets (green), and compound activities (red) in ChEMBL from release 2 to 14.

In order to address the issue of target- or family-selectivity of scaffolds, we also carried out the analysis by excluding all weakly active compounds from the comparison and scaffold assignments. Although we applied a stringent potency threshold value of 10 μM, the original and selectivity-oriented scaffold assignments were nearly identical (for both types of activity measurements). In addition, more than 95% of the scaffold assignments were conserved comparing ChEMBL 2 and 14. Thus, on the basis of these findings, we can conclude that ST and SF scaffolds display strong target- and family-selective tendencies, respectively.

In our current analysis, hundreds of ST and SF scaffolds were identified, many more than we would have anticipated. These scaffolds already represented relatively large numbers of compounds, on average close to 20 for ST and more than 20 for SF scaffolds. Hence, their assignment presently already has a relatively high level of confidence. Thus, using the large pool of SF scaffolds, privileged substructure assignments can be revisited and new candidates be predicted. Moreover, the more than 600 unique ST scaffolds can be considered as starting points for the design of target-selective compounds. The measurement type-dependent sets of ST, SF, and MF scaffolds identified in our study are made freely available via http://www.lifescienceinformatics.uni-bonn.de/downloads.

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) Hu, Y.; Stumpfe, D.; Bajorath, J. Lessons Learned from Molecular Scaffold Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1742−1753.
(2) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(3) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP−Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(4) Clark, M. Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **2005**, *45*, 30−38.
(5) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 182−193.
(6) 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.
(7) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512−524.
(8) Shelat, A. A.; Guy, R. K. Scaffold Composition and Biological Relevance of Screening Libraries. *Nature Chem. Biol.* **2007**, *3*, 442−446.
(9) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., III; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443−4451.
(10) Hu, Y.; Bajorath, J. Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Targets. *Med. Chem. Commun.* **2010**, *1*, 339−344.

(11) Hu, Y.; Bajorath, J. Structural and Potency Relationships between Scaffolds of Compounds Active against Human Targets. *ChemMedChem* **2010**, *5*, 1681−1685.
(12) Cases, M.; Mestres, J. A Chemogenomic Approach to Drug Discovery: Focus on Cardiovascular Diseases. *Drug Discovery Today* **2009**, *14*, 479−485.
(13) 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.
(14) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(15) Hu, Y.; Bajorath, J. Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500−510.
(16) Evans, B. E.; Rittle, K. E.; Bock, M. G.; Dipardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S. Methods for Drug Discovery: Development of Potent, Selective, Orally Effective Cholecystokinin Antagonists. *J. Med. Chem.* **1988**, *31*, 2235−2246.
(17) Müller, G. Medicinal Chemistry of Target Family-Directed Masterkeys. *Drug Discovery Today* **2003**, *8*, 681−691.
(18) Constantino, L.; Barlocco, D. Privileged Substructures as Leads in Medicinal Chemistry. *Curr. Med. Chem.* **2006**, *13*, 65−85.
(19) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *49*, 2000−2009.
(20) 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.
(21) Hu, Y.; Bajorath, J. Exploring Target-Selectivity Patterns of Molecular Scaffolds. *ACS Med. Chem. Lett.* **2010**, *1*, 54−58.
(22) 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.
(23) ChEMBL. http://www.ebi.ac.uk/chembldb/ (accessed October 1, 2012).
(24) UniProtConsortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D142−D148.
(25) 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.
(26) Hu, Y.; Bajorath, J. Growth of Ligand-Target Interaction Data in ChEMBL Is Associated with Increasing and Measurement-Dependent Compound Promiscuity. *J. Chem. Inf. Model.* **2012**, *52*, 2550−2558.

■ NOTE ADDED AFTER ASAP PUBLICATION

This article was published ASAP on February 5, 2013, with an error in the Abstract and Table of Contents graphics. The corrected version was published ASAP on February 7, 2013.