JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

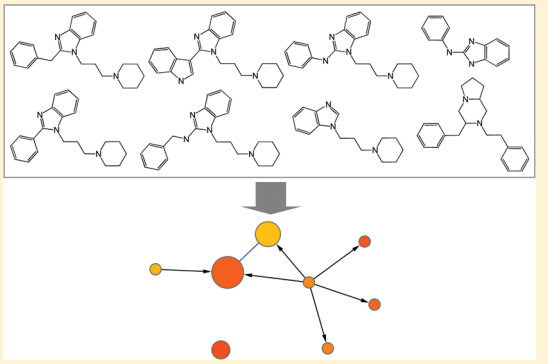
pubs.acs.org/jcim

Target Family-Directed Exploration of Scaffolds with Different SAR Profiles

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: The scaffold concept is widely applied in chemoinformatics and medicinal chemistry to organize bioactive compounds according to common core structures or associate compound classes with specific biological activities. A variety of scaffold analyses have been carried out to derive statistics for scaffold distributions, generate structural organization schemes, or identify scaffolds that preferentially occur in given compound activity classes. Herein we further extend scaffold analysis by identifying scaffolds that display defined SAR profiles consisting of multiple properties. A structural relationship-based scaffold network has been designed as the basic data structure underlying our analysis. From network representations of scaffolds extracted from compounds active against 32 different target families, scaffolds with different SAR profiles have been extracted on the basis of decision trees that capture structural and functional characteristics of scaffolds in different ways. More than 600 scaffolds and 100 scaffold clusters were assigned to 10 SAR profiles. These scaffold sets represent different activity and target selectivity profiles and are provided for further SAR investigations including, for example, the exploration of alternative analog series for a given target of target family or the design of novel compounds on the basis of scaffold(s) with desired SAR profiles.



INTRODUCTION

Scaffolds are generally defined as molecular core structures or frameworks that represent different series of compounds.[1] The scaffold concept has for long been intensely applied in the context of compound analysis and design.[1−5] Furthermore, in virtual compound screening, scaffold hopping,[5,6] i.e. the ability to detect novel active compounds with core structures different from known actives, has become the major success criterion. In order to organize scaffold populations, structural hierarchies and classification schemes have been introduced.[7,8] Furthermore, theoretically possible organic scaffolds have also been virtually enumerated to delineate and analyze scaffold spaces.[9−11]

In addition to investigations primarily focusing on structural criteria, scaffolds have also been extensively analyzed from a biological activity perspective.[1] Scaffold hopping focuses on structurally diverse compounds having a specific activity, but there are also other ways to analyze scaffolds with respect to the biological activities of compounds they represent. First and foremost, it has often been attempted to identify scaffolds that exclusively or preferentially act on individual target families.[12−15] Such scaffolds are then considered to represent privileged structural motifs for target family directed compound design. In addition, efforts have also been made to associate scaffolds with individual activity- or SAR-related properties of bioactive compounds they represent. For example, in a series of studies originating from our laboratory, scaffolds have been identified that preferentially yield target family selective compounds, display a tendency to represent compounds with polypharmacological behavior, or frequently form activity cliffs.[15−19] Such analyses typically involve systematic mining of available compound activity data.

In addition to retrospective data analysis, biological activities of scaffolds have also been predicted.[20,21] For this purpose, scaffolds with known and unknown activity have been organized and compared in well-defined structural hierarchies. Thus, scaffold activity prediction represents an area where structural organization schemes and activity-directed scaffold analyses are often combined. Taken together, the studies discussed above illustrate that scaffolds have been investigated in rather different ways from a biological activity perspective.

Herein we further extend data mining approaches for scaffold characterization. Rather than identifying sets of scaffolds with individual properties, we have searched for scaffolds with different SAR profiles. Major goals of our analysis include the generation of a previously unconsidered SAR-oriented scaffold classification scheme and, in addition, the identification of scaffold sets with well-defined SAR profiles for compound exploration and design. For this purpose, we have generated a data structure that organized scaffolds according to well-defined structural relationships and provided a basis for mining of activity and target information associated with compounds represented these scaffold. On the basis of currently available compound activity data, multiproperty SAR profiles were defined, and subsets of scaffolds matching a given profile were extracted from the network data structure with

Received: September 29, 2011
Published: November 17, 2011

Table 1. Target Families[a]

ID	target family	no. targets	no. compounds	no. scaffolds	scaffolds with structural relations
1	Tyr kinases	37	3229	549	486 (88.5%)
2	Ser-Thr kinases	63	3196	545	503 (92.3%)
3	Ser-Thr-Tyr kinases	10	375	67	52 (77.6%)
6	Ser proteases	39	3460	673	642 (95.4%)
7	Asp proteases	7	951	163	135 (82.8%)
8	Cys proteases	17	1318	247	227 (91.9%)
9	metallo proteases	28	1533	254	231 (90.9%)
10	carbonic anhydrases	12	687	104	92 (88.5%)
11	cytochrome P450 enz.	19	1116	216	170 (78.7%)
12	other cytosolic enz.	19	547	84	51 (60.7%)
13	electrochemical transporters	18	2668	324	300 (92.6%)
20	chemokine GPCR	10	2130	324	255 (78.7%)
22	lipid-like GPCR	24	3006	470	431 (91.7%)
23	glutamate GPCR	6	520	75	65 (86.7%)
24	monoamine GPCR	35	4987	916	874 (95.4%)
27	nucleotide-like GPCR	9	2967	498	476 (95.6%)
29	secretin-like GPCR	6	682	108	96 (88.9%)
30	short peptide GPCR	47	7364	1262	1121 (88.8%)
31	other membrane recep.	6	74	21	9 (42.9%)
32	nuclear recep.	24	2533	319	303 (95.0%)
33	histone deacetylases	11	521	123	98 (79.7%)
34	LG-ion channel	20	332	58	34 (58.6%)
35	VG-ion channel	13	1273	266	174 (65.4%)
38	phosphodiesterases	12	640	109	92 (84.4%)
40	phospholipases	5	129	26	21 (80.8%)
41	phosphatases	14	316	60	45 (75.0%)
42	integrins	8	253	39	32 (82.1%)
44	oxidoreductases	42	2867	481	436 (90.6%)
45	hydrolases	27	1252	232	183 (78.9%)
47	lyases	7	276	39	32 (82.1%)
48	transferases	44	1796	274	260 (94.9%)
50	isomerases	7	294	58	43 (74.1%)

[a] Thirty-two target families are listed. These families are designated according to the ChEMBL target family classification scheme. For each family, the number of targets, active compounds, BM scaffolds, and the number (percentage) of scaffolds involved in structural relationships are reported. The following abbreviations are used: GPCR, G protein-coupled receptors; enz., enzymes; recep., receptors; LG, ligand-gated; VG, voltage-gated. In addition, amino acids are given in standard three-letter code.

the aid of decision trees that evaluated combinations of different SAR-relevant features at the level of individual target families. Scaffold sets representing defined SAR profiles are provided for further analysis.

MATERIALS AND METHODS

Data Collection. From ChEMBL (release 10),[22] compounds active against human targets with at least 10 μM potency (i.e., K_i or IC_{50} values) were extracted. Only compounds with defined potency measurements (pKi or pIC_{50}) were considered. For compounds active against a given target and representing a single scaffold, consistent potency measurements (i.e., only K_i or IC50 values) were available in most instances and utilized in our analysis. Accordingly, the results should not be notably affected by combining alternative potency measurements. The potency cutoff was applied to exclude very weakly potent compounds from the analysis. Consequently, 62 target sets for which only weakly potent compounds were found were excluded from further analysis. From all compounds containing ring structures, Bemis and Murcko (BM) scaffolds[2] were isolated. A total of 71,678 compounds active against 864 human targets were selected that yielded 25,836 unique BM scaffolds. On the basis of the ChEMBL target classification scheme, these compounds were active against 864 targets that belonged to 51 families. All of these human targets are generally considered therapeutically relevant (which explains why significant amounts of compound data are available for them). Scaffolds representing only a single compound were removed from each target set. Furthermore, only target families were retained for further analysis that consisted of at least five targets. These restrictions reduced the number of qualifying compounds to 51,756 that were active against 646 targets from 32 families. These compounds then yielded 8265 unique BM scaffolds. Table 1 reports the compound and scaffold composition of target family directed data sets, and Figure 1 shows the potency distribution for each set.

Structural Relationships. BM scaffolds were further transformed into cyclic skeletons (CSKs)[23] by converting all

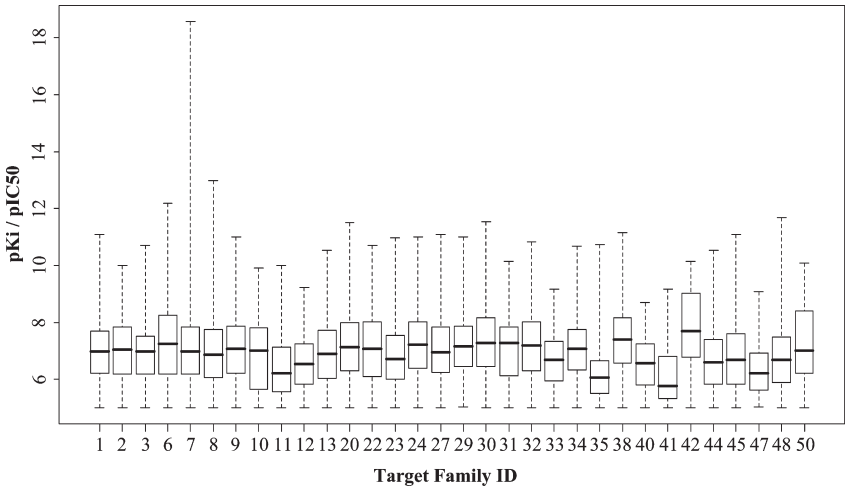


Figure 1. Compound potency distributions. For all 32 target families, compound potency distributions are reported as box plots. Target family IDs are provided according to Table 1. The box plots report the smallest potency value (bottom line), lower quartile (lower boundary of the box), median value (thick line), upper quartile (upper boundary of the box), and the largest potency value (top line). Dashed vertical lines indicate the potency range.

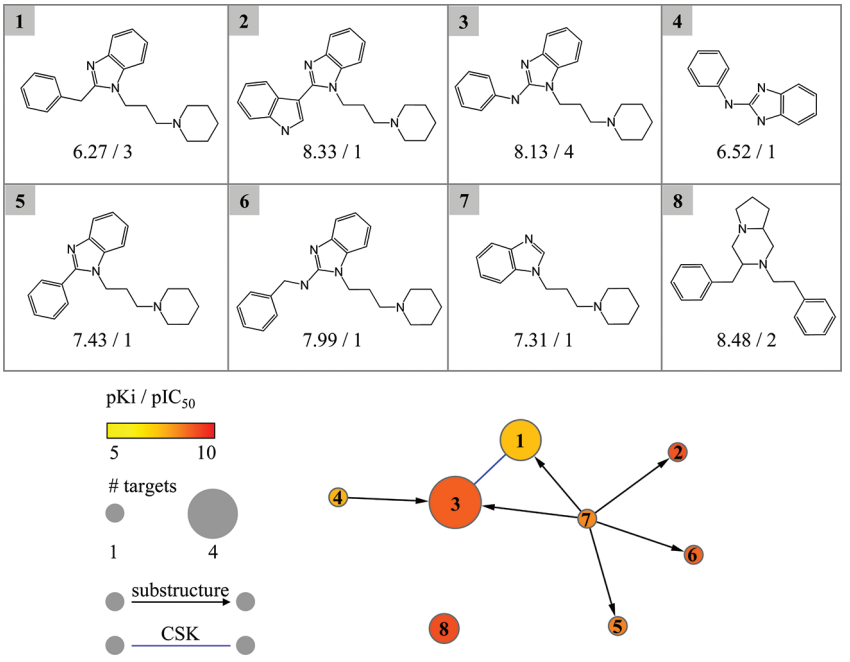


Figure 2. Scaffold network. The design elements of this network representation are illustrated. For each of eight exemplary scaffolds (1−8), the median potency value of the compounds each scaffold represents and the number of targets the compounds are active against are reported. For example, "6.27/3" stands for a median compound potency value of 6.27 and activity against three targets within a family. In this case, an individual compound might be active against one, two, or all three targets. In the network representation, nodes indicate scaffolds. Two nodes are connected by an edge if they form a substructure relationship (i.e., black directed edge) or yield the same CSK (i.e., blue undirected edge). Nodes are color-coded according to median compound potency values using a continuous spectrum (covering the logarithmic potency range from 5 to 10) and scaled according the number of targets the compounds are active against. Scaffold 8 is not involved in a structural relationship and is thus displayed as a singleton in the network.

heteroatoms to carbon and all bond orders to single bonds. BM scaffolds having the same topology but variable heteroatoms and/or bond orders yielded the same CSK. Two types of structural relationships were defined:

(1) A BM scaffold is a substructure of another one (i.e., it is completely contained in another BM scaffold).
(2) Multiple BM scaffolds correspond to the same CSK (i.e., these scaffolds are topologically equivalent).

For all scaffolds isolated from compounds active against each target family, these two types of structural relationships were systematically determined. The benzene ring, the most generic organic scaffold, was excluded from the assessment of substructure relationships because it was a substructure of the majority of scaffolds.

Structural Relationship-Based Scaffold Network. For each target family, structural relationships between scaffolds were

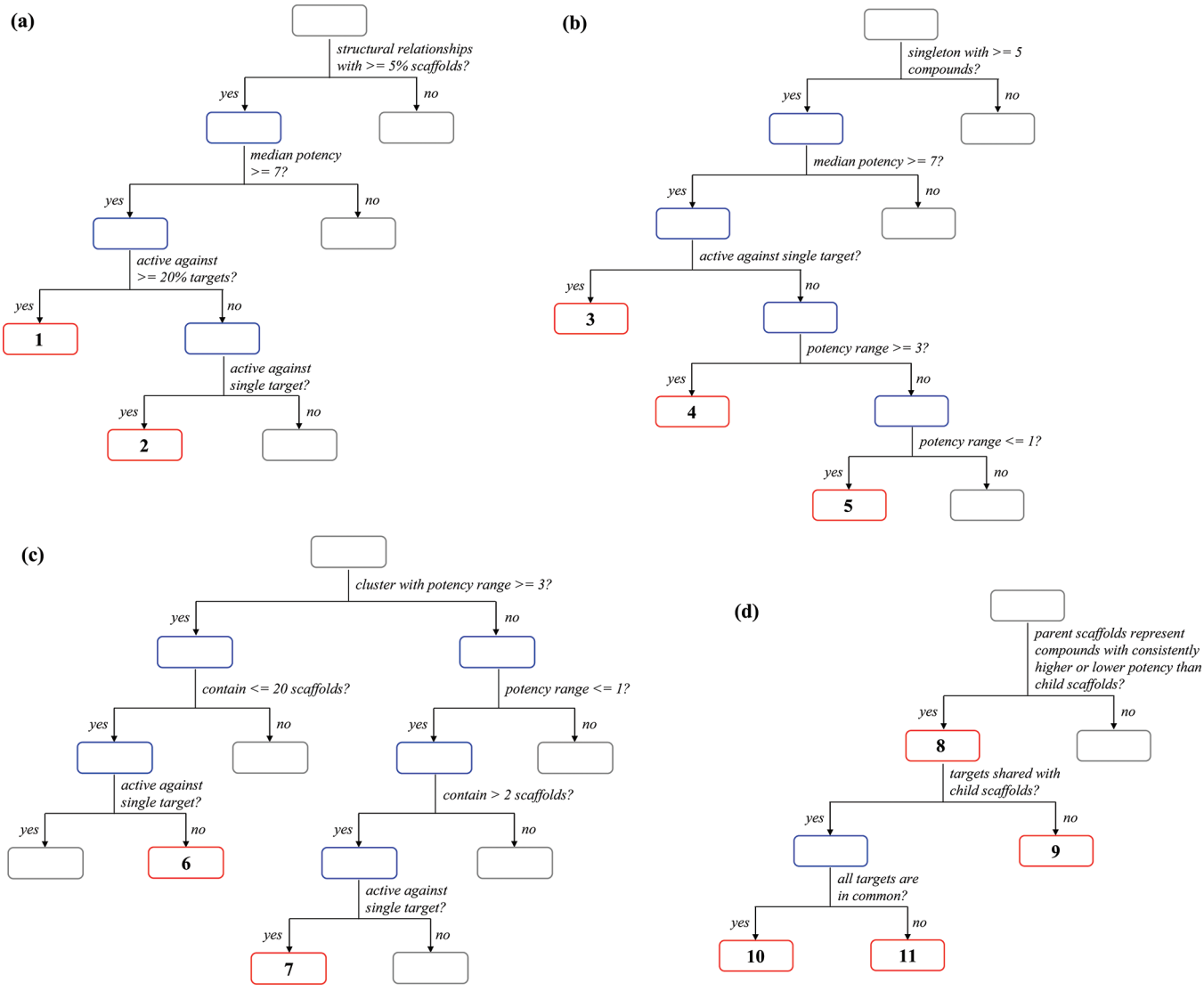


Figure 3. Decision trees. Four decision trees were utilized to extract scaffold subsets with distinct SAR profiles from target family directed network representations. The decision trees and the SAR profiles they capture are described in the text: (**a**) tree structure I, (**b**) tree II, (**c**) tree III, and (**d**) tree IV. Terminal scaffold subset boxes representing different profiles (1−11) are outlined in red, and intermediate sets containing these scaffolds are outlined in blue.

monitored in a network representation, as illustrated in Figure 2. In this network, scaffolds were visualized as nodes. In addition, the network embodied the following design elements:

(a) Two nodes were connected by an edge if the scaffolds were involved in a structural relationship. A substructure relationship was denoted by a directed black edge, pointing from a parent scaffold (i.e., the smaller scaffold) to a child scaffold (i.e., the larger scaffold). If two scaffolds yielded the same CSK, they were connected by an undirected blue edge.

(b) Each node was color-coded according to median potency value of the compounds active against all targets in the family that were represented by this scaffold. A continuous color spectrum from yellow to red was used to indicate low to high potency. A pKi or pIC_{50} value range from five to 10 was consistently applied to define the color spectrum for all target families. The lower boundary represented the minimum compound potency of 10 μM. The upper boundary was selected because compounds active against the majority of target families reached a maximum potency close to 0.1 nM, as shown in Figure 1.

(c) Nodes were scaled in size on the basis of a target family centric compound promiscuity measure, i.e., the number of targets within a family that compounds represented by the scaffold were active against: the more promiscuous the scaffold, the larger the size of its node. Although currently available public compound data are generally sparse (i.e., not all compounds have been tested against all available targets), notable degrees of promiscuity were already detectable in a number of cases, as further described below. A consistent range from one to 18 targets was used for node size scaling across all target families. Eighteen represented the largest number of targets a scaffold was active against for all 32 target families, i.e. the highest degree of promiscuity.

Design criteria (b) and (c) ensured that node colors and sizes were comparable across all families. The network representations were drawn with Cytoscape.[24]

dx.doi.org/10.1021/ci200461w |J. Chem. Inf. Model. 2011, 51, 3138–3148

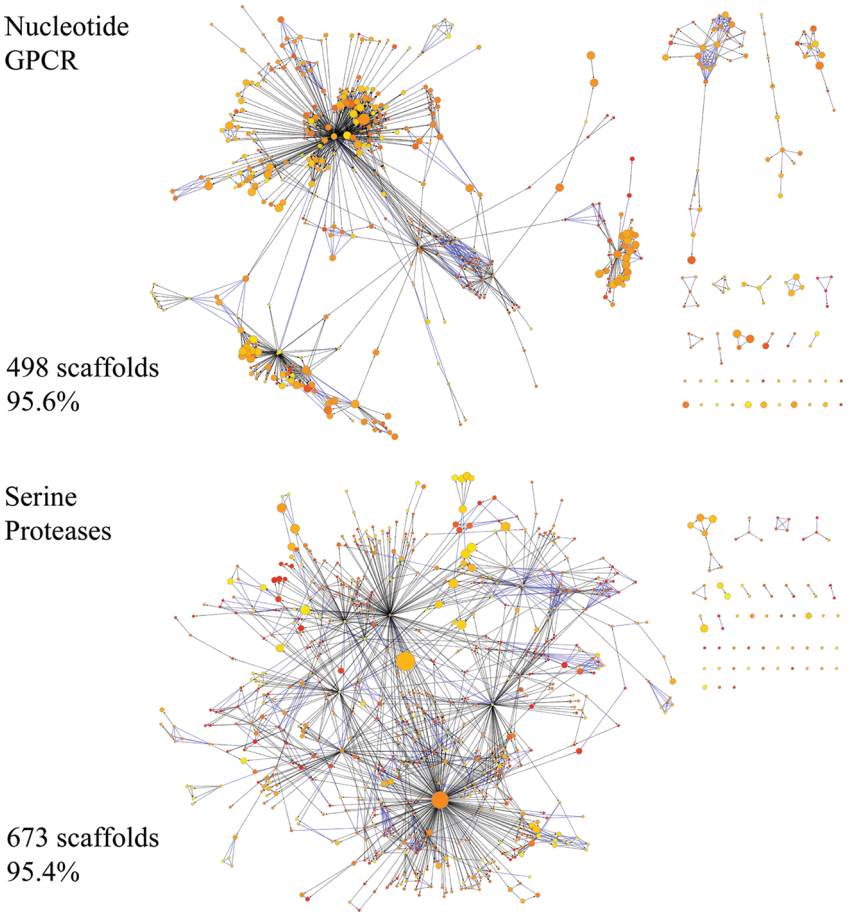


Figure 4. Networks of target families with high ligand similarity. Two scaffold networks are shown for nucleotide GPCR ligands and serine protease inhibitors where 95.6% of 498 scaffolds and 95.4% of 673 scaffolds, respectively, are involved in structural relationships.

Scaffolds with Characteristic SAR Profiles. From scaffold networks, scaffolds with characteristic features were extracted via decision trees that evaluated property combinations of scaffolds and corresponding compounds. The decision tree structures are shown in Figure 3. Each decision tree defined a varying number of *SAR profiles*. The terminal scaffold subset boxes representing different profiles are outlined in red in Figure 3. Four different decision trees were designed including tree I containing Profiles 1−2 (Figure 3a), tree II with Profiles 3−5 (Figure 3b), tree III with Profiles 6−7 (Figure 3c), and tree IV with Profiles 8−11 (Figure 3d). These profiles were termed SAR profiles because they captured SAR-relevant information of scaffolds and compounds represented by them. The decision trees and SAR profiles are described in detail in the following.

RESULTS AND DISCUSSION

Key Aspects of the Scaffold Analysis Scheme. Three criteria that represent key characteristics of our analysis scheme should be emphasized. First, scaffold classification is based on activity data of compounds that are represented by a given scaffold. Hence, there is a direct link between compound activity data and scaffold SAR profiles. This also means that scaffolds with specific SAR profiles can be utilized to predict new analogs for SAR analysis. Second, our analysis has been carried out at the level of individual target protein families, not across different families, which determined scaffold selection criteria. Third, scaffold networks have provided the basis for the extraction of individual scaffolds or groups of scaffolds with characteristic features through the application of decision trees, which evaluated property combinations of compounds represented by individual scaffolds. This again emphasizes the link between compound activity data and scaffold SAR profiles. Each decision tree represents a varying number of SAR profiles. Thus, SAR profiles of scaffolds were not arbitrarily defined but established through compound data mining. In addition, it should also be considered that any data mining effort is naturally limited to currently available data. For example, for the targets investigated here, no confirmed inactive compounds are deposited that could be taken into consideration for SAR profile derivation. Hence, by default, the focus is on active compounds. Furthermore, as already mentioned above, active compounds have generally not been tested on all available targets, which refers to the well-known phenomenon of data sparseness. This is a general limitation of compound data mining efforts. As a consequence, the presence of partly incomplete compound activity annotations is likely. Accordingly, as more measurements become available, the currently detectable level of compound promiscuity would be expected to further increase. However, the scaffold SAR profiles determined in our study comprehensively capture currently available SAR information.

Structural Relationships and Scaffold Network Design. Two types of structural relationships between scaffolds were determined for our analysis including substructure relationships

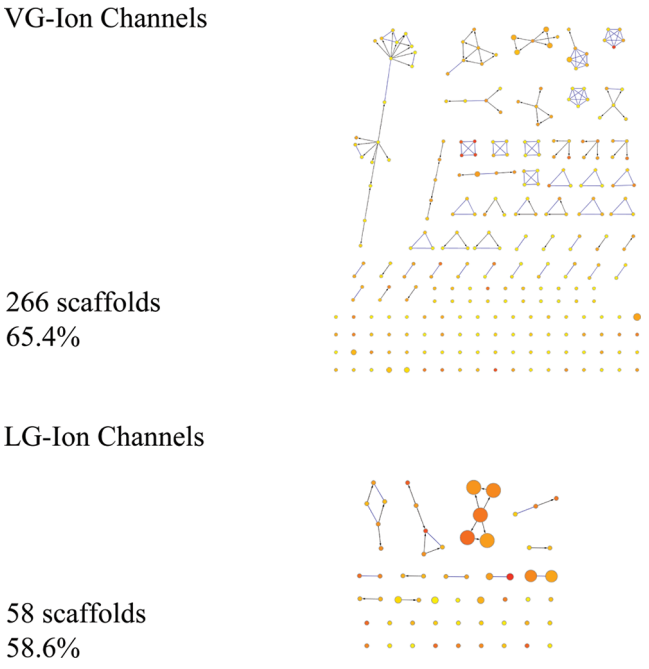


Figure 5. Networks of target families with low ligand similarity. Two scaffold networks are shown for ligand sets of VGC and LGIC ion channels where 65.4% of 266 scaffolds and 58.6% of 58 scaffolds, respectively, are involved in structural relationships.

and topological equivalence. If the molecular graph of one scaffold is completely contained in another, a substructure relationship exists. Furthermore, if two scaffolds yield the same CSK, they share the same topology. Both types of structural relationships establish pairwise scaffold relationships in the network. They are treated equivalently but displayed differently, as illustrated in Figure 2. The color code of nodes (scaffolds) reflects the median potency of compounds scaffolds represent and the size of nodes the total number of targets the compounds are active against within a target family. Utilizing these design elements, scaffold networks have been generated from compounds active against all qualifying target families. For each scaffold, target annotations and potency values were recorded for all compounds it represented as part of the network data structure.

Data Mining Strategy. These target family directed scaffold networks provided a basis for the identification of subsets of scaffolds having different properties. For this purpose, networks were searched for subsets of scaffolds with defined structural relationship characteristics, for examples, densely connected scaffolds (hubs), isolated scaffolds (singletons), or disjoint scaffold clusters. From networks with different topology, as further discussed below, characteristic scaffold subsets emerged that were then further characterized in quantitative terms and extracted using property-based decision trees.

Target Family Specific Network Topologies. In Figures 4-6, exemplary networks with different topologies are shown. Table 1 also reports the percentage of scaffolds per target family that were involved in structural relationships. For 27 families, more than 70% of all scaffolds formed structural relationships, indicating that the majority of bioactive scaffolds were discovered or generated taking known structural information into account. Figure 4 shows scaffold networks of two target families that were characterized by the presence of many structural relationships. More than 95% of the scaffolds displayed here were involved in relationships. Both networks contain large densely connected regions. The nucleotide GPCR network consists of several densely connected subgraphs, a number of isolated scaffold clusters, and a limited number of singletons, whereas the serine protease network contains a large central network component, very small scaffold clusters, and singletons. By contrast, in Figure 5, networks reflecting a limited number of structural relationships are shown that lack densely connected regions and scaffold clusters of larger size. Moreover, Figure 6 compares networks for three target families where a very similar percentage of scaffolds, approximately 89%, were involved in structural relationships. Yet, the network topologies differ substantially. The short peptide GPCR network that contains by far the largest number of scaffolds has a large central network component with a densely connected center and a number of individual scaffolds with relationships to many others, which form "hubs" in the network. By contrast, the carbonic anhydrase network has a much lower edge density, and the secretin-like GPCR network exclusively consists of small disjoint clusters and a limited number of singletons. Thus, the presence of comparable percentages of scaffolds involved in structural relationships often resulted in different network topologies, depending on whether scaffolds were involved in relationships with single or multiple scaffolds, which determined the distribution of node degrees. Taken together, three types of scaffolds emerged from target family directed networks as focal point for further analysis. These scaffolds included prominent hubs, scaffolds in disjoint clusters, and singletons, which were characterized by different structural relationships, ranging from many relationships with other scaffolds (hubs) and locally confined relationships (clusters) to no structural relationships (singletons).

Decision Trees and Scaffold SAR Profiles. To systematically mine these three types of scaffolds by taking other SAR-relevant properties into account, decision tree queries were designed, as shown in Figure 3. These additional properties included the potency level of compounds represented by each scaffold, their potency distributions, potency relationships between parent and child scaffolds, and the number of targets of a given family (that compounds represented by a scaffold were active against). Each decision tree defined from two to four characteristic scaffold SAR profiles, yielding a total of 11 distinct profiles that are listed and described in Table 2.

The decision tree I in Figure 3a was designed to select hubs (having individual relationships with at least 5% of all scaffolds in a target family set) that represented highly potent compounds (median potency value at least 100 nM) active against either single targets (target-specific) or at least 20% of targets within a family (promiscuous). This rule combination hence resulted in SAR profiles 1 and 2 in Table 2. The decision tree II in Figure 3b selected singleton scaffolds that represented at least five compounds with high median potency that were target-specific, giving rise to profile 3. Alternatively, singletons with high median potency and activity against multiple targets were selected, and, in this case, the potency distributions of the corresponding compounds were taken into account. Scaffolds with compounds spanning a wide potency range of at least 3 orders of magnitude (profile 4) were distinguished from scaffolds with compounds having consistently high potency within an order of magnitude (profile 5). In Table 2, these potency distributions are termed heterogeneous and homogeneous, respectively. The decision tree III in Figure 3c was focused on disjoint scaffold clusters. Scaffolds forming such clusters were generally involved in multiple but locally confined

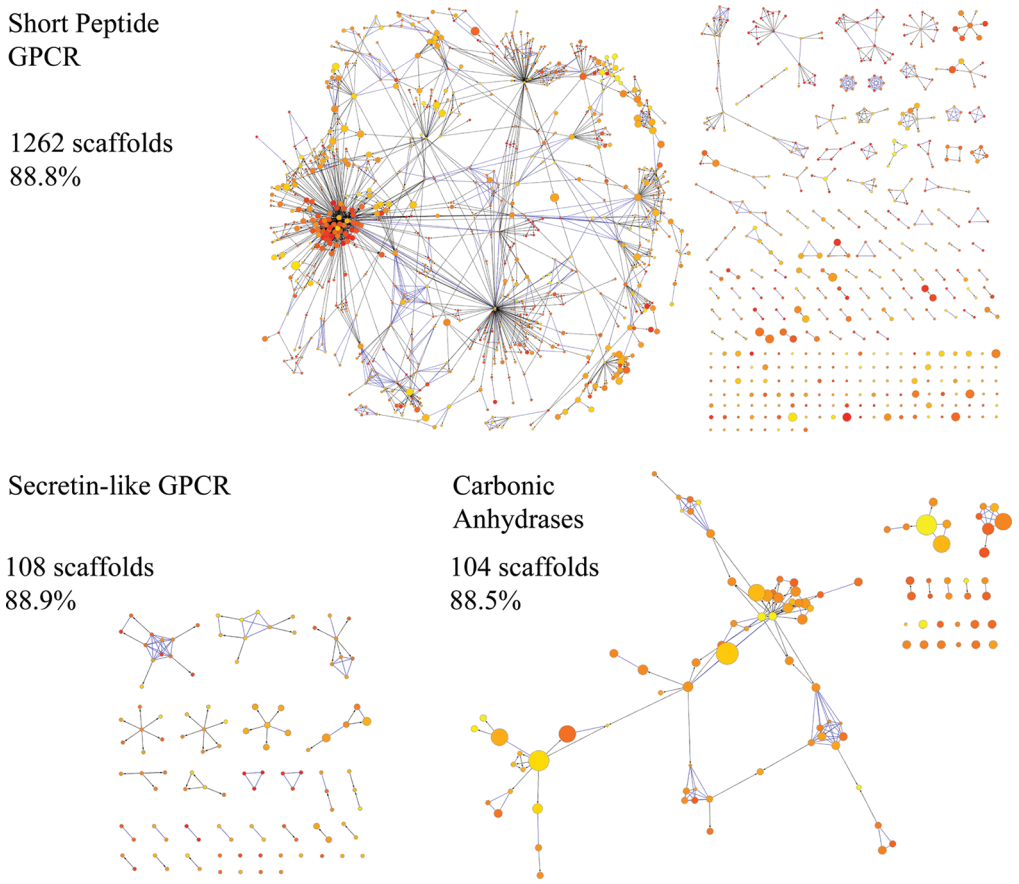


Figure 6. Scaffold networks with distinct topology. Three scaffold networks are shown for short peptide and secretin-like GPCR ligands and carbonic anhydrase inhibitors, which are characterized by comparable intrafamily ligand similarity (i.e., 88.8%, 88.9%, and 88.5% of scaffolds are involved in structural relationships, respectively) but different network topologies.

Table 2. SAR Profiles[a]

profile	description
1	scaffolds involved in many structural relationships (hubs) and representing highly potent promiscuous compounds
2	scaffolds involved in many structural relationships (hubs) and representing highly potent target-specific compounds
3	scaffolds not involved in structural relationships yielding highly potent target-specific compounds
4	scaffolds not involved in structural relationships yielding promiscuous compounds with high median potency but heterogeneous potency distribution
5	scaffolds not involved in structural relationships yielding highly potent promiscuous compounds with compound homogeneous potency distribution
6	scaffolds in disjoint clusters with promiscuity and heterogeneous potency distribution
7	scaffolds in disjoint clusters with target specificity and homogeneous potency distribution
8	parent scaffolds with consistent potency relationships to child scaffolds
9	parent scaffolds with consistent potency relationships to child scaffolds: parent and child scaffolds are active against distinct targets
10	parent scaffolds with consistent potency relationships to child scaffolds: parent and child scaffolds are active against same targets
11	parent scaffolds with consistent potency relationships to child scaffolds: parent and child scaffolds are active against an overlapping set of targets

[a] The 11 different SAR profiles are defined and grouped according to decision trees from which they originate.

structural relationships. Here clusters with heterogeneous compound potency ranges and at most 20 scaffolds active against multiple targets were selected, giving rise to profile 6, which thus emphasized target promiscuity and potency heterogeneity of closely related scaffolds. Alternatively, clusters with narrow potency ranges and activity against single targets were selected (profile 7). In this case, small scaffold clusters with target specificity were also accepted. It should be noted that profiles 6 and 7 selected scaffold clusters and not individual scaffolds like all other nine profiles. Finally, with decision tree IV in Figure 3d, potency relationships between parent and child scaffolds were explored (parent scaffolds are substructures of their children). First, parent scaffolds were selected that represented compounds with consistently higher or lower potency than child scaffolds (profile 8). Thus, well-defined potency relationships between scaffolds involved in substructure relationships were emphasized in this case. Furthermore, from these scaffolds, subsets were selected that represented compounds active only against other

Table 3. Target-Family Distribution of Scaffolds with Different SAR Profiles[a]

			profile									
family ID	no. scaffolds	no. disjoint clusters	1	2	3	4	6	7	8	9	10	11
1	549	48	-	-	6	1	6	-	13	1	8	4
2	545	28	-	-	4	1	7	1	21	3	18	-
3	67	13	-	4	2	-	-	-	3	-	3	-
6	673	14	1	1	6	-	2	1	23	4	17	2
7	163	16	-	1	2	1	1	-	2	-	2	-
8	247	7	1	2	-	1	4	1	6	-	5	1
9	254	10	1	1	-	-	3	-	6	3	3	-
10	104	8	6	3	-	-	4	-	1	-	-	1
11	216	12	-	1	4	-	2	-	5	2	2	1
12	84	18	-	-	-	-	1	-	3	1	2	-
13	324	11	1	2	4	1	2	-	3	1	2	-
20	324	55	-	-	7	-	6	2	4	-	4	-
22	470	17	-	1	4	2	3	1	7	-	6	1
23	75	16	1	12	-	-	3	-	1	-	1	-
24	916	16	-	-	4	1	5	-	16	-	16	-
27	498	15	2	2	1	-	2	-	7	-	4	3
29	108	24	-	8	4	-	1	-	1	-	1	-
30	1262	96	-	-	14	3	19	1	30	3	26	1
31	21	2	-	3	-	-	1	-	1	1	-	-
32	319	18	-	4	5	-	5	-	6	-	5	1
33	123	8	-	2	1	-	1	-	6	-	4	2
34	58	12	1	2	2	-	4	-	1	-	-	1
35	266	49	-	-	1	-	1	3	6	-	6	-
38	109	18	1	1	4	-	4	-	1	-	1	-
40	26	5	4	2	-	-	1	-	-	-	-	-
41	60	7	-	1	-	-	-	-	1	-	1	-
42	39	9	5	2	1	-	2	-	1	-	1	-
44	481	16	-	-	6	2	3	-	14	3	10	1
45	232	19	-	-	3	-	1	1	6	-	6	-
47	39	6	-	4	-	-	2	-	4	-	4	-
48	274	9	-	2	-	-	-	-	11	3	7	1
50	58	7	-	18	1	-	-	-	1	-	1	-

[a] For each of the 32 target families, the total number of scaffolds and disjoint scaffold clusters are given. In addition, the numbers of scaffolds or scaffold clusters representing different SAR profiles (1-11, except 5) are reported (no scaffolds were found to display profile 5). SAR profiles are designated according to Table 2.

targets than their child scaffolds (profile 9), the same targets (profile 10), or an overlapping set of targets (profile 11). Thus, profiles 9−11, which were designed to select scaffold subsets from profile 8, further emphasized the target selectivity differences among scaffolds with defined structural and potency relationships.

SAR Information Content. The SAR profiles summarized in Table 2 were designed to select scaffold subsets having property characteristics of high relevance for SAR analysis and compound design. For example, important information is often provided by comparison of scaffolds that

- are structurally related and yield highly potent compounds for a target family
- display a tendency to be either promiscuous or target-selective
- are structurally unique and yield target-specific compounds of high potency
- represent compounds with significantly different potency
- represent promiscuous compounds of varying potency distribution
- have defined structural and potency relationships to other scaffolds
- are structurally related but active against different targets.

Importantly, the scaffold profiles investigated here combine these and other SAR-relevant characteristics in different ways, which further extends structure- and/or activity-oriented compound data and scaffold analysis.

SAR Profile Mining. Using the decision trees discussed above, we searched all 32 target family based networks for scaffolds (or clusters) matching each SAR profile. The results are reported in Table 3 at a target family level. In addition, the distribution of qualifying scaffolds over all families is provided in Table 4. For 10 of our 11 profiles, scaffold subsets were identified in compounds active against different target families. The only exception was

Table 4. Global Scaffold Distribution over Different SAR Profiles[a]

profile	no. scaffolds/scaffold clusters	no. families
1	24	11
2	79	23
3	86	22
4	13	9
5	0	0
6	96	28
7	11	8
8	211	31
9	25	11
10	166	28
11	20	13

[a] For each SAR profile, the number of qualifying scaffolds or scaffold clusters (for profiles 6 and 7) and the number of target families these scaffolds originate from are reported.

profile 5 for which no qualifying scaffold was detected for any family. This profile described structurally unique scaffolds yielding highly potent promiscuous compounds with homogeneous potency distribution, which we would indeed expect to be a rare property combination for bioactive scaffolds. As reported in Table 3, the distribution of scaffold profile sets (i.e., scaffolds matching a given SAR profile) was highly variable over different target families. In addition, no target family contained scaffolds for all profiles and overall less than 10% of all available scaffolds matched SAR profiles, as one would expect for profiles representing rather different and in part complex property combinations. However, for one target family, serine proteases (family ID 6), scaffolds were found to match nine SAR profiles, and, for several others, eight profiles were matched. The total number of available scaffolds greatly differed for the 32 target families (Table 3), which partly determined the number of profile-matching scaffolds per family. Thus, for families with small numbers of available scaffolds, only a few profiles were usually matched. Furthermore, some profiles were also much more frequently matched than others, as reported in Table 4. For example, a total of 211 scaffolds from 31 target families displayed profile 8, and 96 scaffold clusters from 28 families displayed profile 6. Hence, scaffolds that were substructures of others and represented compounds with systematic potency relationships were frequently observed (profile 8) as well as disjoint scaffold clusters with promiscuity and heterogeneous compound potency distributions (profile 6), which should provide a rich source of SAR information. By contrast, profile 7 that was distinguished from profile 6 by the requirements of target-specificity and a homogeneous potency distribution was only matched by 11 clusters from eight target families. Furthermore, only 13 structurally unique scaffolds from nine families were found that yielded compounds active against multiple targets with high median potency but heterogeneous potency distribution (profile 4). In addition to profile 5, which that was not observed, profile 4 was overall least frequently matched. However, 86 structurally unique scaffolds that represented highly potent compounds active against single targets (profile 3) were found for 22 target families, and these scaffolds should also provide interesting starting points for the generation of other target-selective or -specific compounds. For half of our profiles, between 79 and 211 qualifying scaffolds were detected (Table 3),

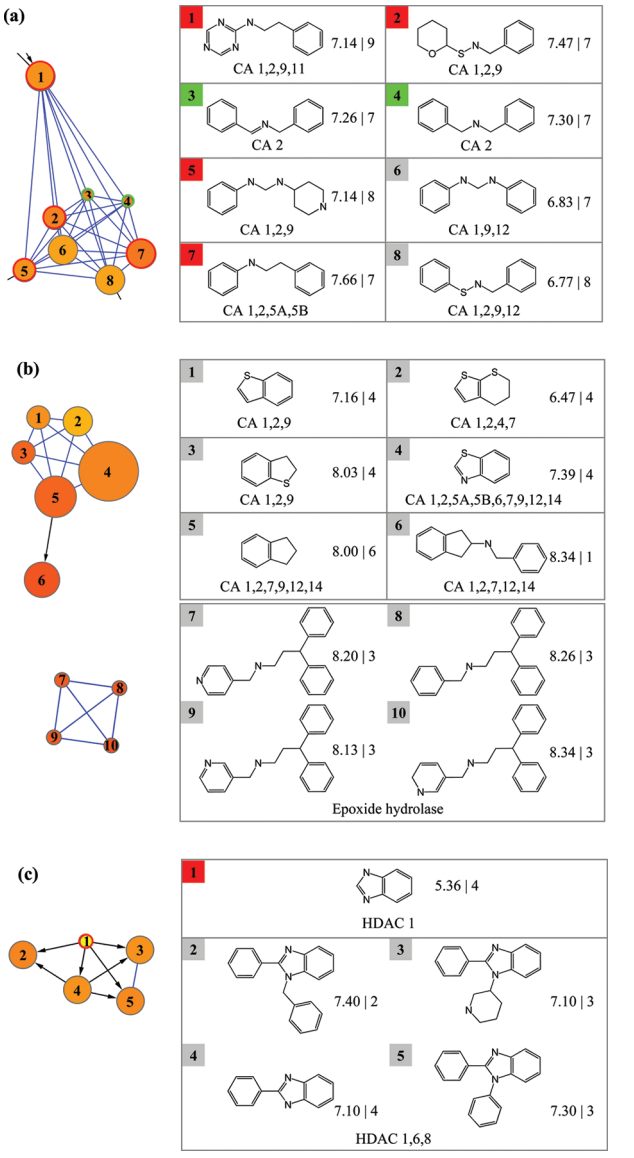


Figure 7. Exemplary scaffolds with different SAR profiles. Scaffolds or scaffold clusters are shown that display different SAR profiles. (**a**) Subset of a cluster of scaffolds from the carbonic anhydrase (CA) inhibitor network. Scaffolds (1−8) are involved in structural relationships with more than 5% of the scaffolds active against this family. Scaffolds 1, 2, 5, and 7 yield highly potent and promiscuous compounds (i.e., they dispaly profile 1) and are highlighted in red. Scaffolds 3 and 4 produce highly potent and target-specific compounds (profile 2) and are highlighted in green. For each scaffold, the median potency value of the compounds it represents and the number of structural relationships are reported. For example, "7.14 | 9" stands for a median potency value of 7.14 and structural relationships with nine other scaffolds. In addition, target family members the scaffold is active against are also provided (e.g., "CA 1,2,9" means activity against carbonic anhydrases 1, 2, and 9). (**b**) Two scaffold clusters are shown including a cluster from the carbonic anhydrase (top) and another from the serine protease inhibitor network (bottom). The top cluster contains scaffolds representing compounds that are promiscuous within the family and have a heterogeneous potency distribution (profile 6), whereas the bottom cluster consists of scaffolds representing target-specific compounds with homogeneous potency distribution (profile 7). (**c**) A cluster of scaffolds from the histone deacetylase (HDAC) inhibitor network is shown. The parent scaffold 1 (highlighted in red) only yields compounds that are less potent than its child scaffolds 2−5 (profile 11).

dx.doi.org/10.1021/ci200461w |J. Chem. Inf. Model. 2011, 51, 3138–3148

which provide a considerable amount of structural information for SAR analysis.

Exemplary Scaffolds. In Figure 7, representative scaffolds or scaffold clusters are shown that displayed different SAR profiles. For example, the carbonic anhydrase inhibitor scaffolds in Figure 7a are structurally closely related (sharing the same CSK) and yield potent compounds. Four of these scaffolds (1, 2, 5, and 7) are active against several carbonic anhydrase isoforms (profile 1), where two others (3 and 4) are isoform-specific (profile 2). In Figure 7b (top), another carbonic anhydrase inhibitor scaffold cluster is shown that represents compounds active against multiple isoforms with heterogeneous potency distribution (profile 6). By contrast, the scaffold cluster at the bottom of this figure represents compounds with exclusive activity against a single member of the serine protease family with homogeneous potency distribution (profile 7). Figure 7c shows another scaffold cluster extracted from the histone deacetylase inhibitor network, which contains parent-child relationships. Parent scaffold 1 is a substructure of scaffolds 2−5. Compounds represented by the smaller parent scaffold are always less potent than compounds corresponding to its (larger) children. Thus, in this case, a substructure relationship between scaffolds is accompanied by well-defined compound potency relationships and a common target (profile 11).

These examples illustrate the multiproperty-based organization of bioactive scaffolds that match different SAR profiles. Such scaffold subsets can be directly extracted from our network representations. Furthermore, scaffolds or scaffold clusters matching different SAR profiles can also be assembled from different target families.

CONCLUSIONS

In this study, we have systematically searched for scaffolds with different SAR profiles, a previously unconsidered task in scaffold analysis. SAR profiles were designed to combine different SAR-relevant characteristics including different types of structural relationships, potency distributions, and target annotations. For our analysis, scaffolds were systematically extracted from compounds active against a variety of target families. A network data structure provided the basis for the definition of SAR profiles and for mining scaffolds matching these profiles. The network structure was based on structural relations including topological equivalence and substructure relationships and was annotated with additional property information. Network representations for different target families displayed in part rather different topological features and revealed different types of scaffolds preferred for our analysis. Decision trees encoding different SAR profiles were utilized to mine network representations for scaffolds matching these profiles. For 10 of 11 profiles that were queried, scaffold sets were identified that were differently distributed over target families. In total more than 600 scaffolds were found to display different SAR profiles and, in addition, more than 100 scaffold clusters. Taken together, these scaffolds represent complex structural, activity, and target selectivity patterns, which should provide a meaningful source of SAR information. These scaffold sets can be utilized to explore alternative analog series for a given target or target family and predict new compounds on the basis of scaffolds having desired SAR profiles. Upon publication, the scaffold profile sets identified herein are made freely available via http://www.lifescienceinformatics.uni-bonn.de/downloads.

AUTHOR INFORMATION

Corresponding Author

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

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) Katritzky, A. R.; Kiely, J. S.; Hebert, N.; Chassaing, C. Definition of Templates within Combinatorial Libraries. *J. Comb. Chem.* **2000**, *2*, 2–5.

(4) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical Substructures in Drug Discovery. *Drug Discovery Today* **2003**, *8*, 594–602.

(5) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(6) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *19*, 2894–2896.

(7) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 182–193.

(8) 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.

(9) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573.

(10) Pollock, S. N.; Coutsias, E. A.; Wester, M. J.; Oprea, T. I. Scaffold Topologies 1. Exhaustive Enumeration up to Eight Rings. *J. Chem. Inf. Model.* **2008**, *48*, 1304–1310.

(11) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.* **2009**, *52*, 2952–2963.

(12) Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002**, *3*, 928–944.

(13) Constantino, L.; Barlocco, D. Privileged Substructures as Leads in Medicinal Chemistry. *Curr. Med. Chem.* **2006**, *13*, 65–85.

(14) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and Kinase-privileged Fragments: Toward Virtual Polypharmacology. *J. Med. Chem.* **2008**, *51*, 1214–1222.

(15) 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.

(16) 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.

(17) Hu, Y.; Bajorath, J. Global Assessment of Scaffold Hopping Potential for Current Pharmaceutical Target. *Med. Chem. Commun.* **2010**, *1*, 339–344.

(18) Hu, Y.; Bajorath, J. Structural and Potency Relationships between Scaffolds of Compounds Active against Human Targets. *ChemMedChem* **2010**, *5*, 1681–1685.

(19) Hu, Y.; Bajorath, J. Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500–510.

(20) Renner, S.; van Otterlo, W. A. L.; Seoane, M. D.; Möcklinghoff, S.; Hofmann, B.; Wetzel, S.; Schuffenhauer, A.; Ertl, P.; Oprea, T. I.; Steinhilber, D.; Brunsveld, L.; Rauh, D.; Waldmann, H. Bioactivity-guided Mapping and Navigation of Chemical Space. *Nat. Chem. Biol.* **2009**, *5*, 585–592.

(21) Wetzel, S.; Wilk, W.; Chammaa, S.; Sperl, B.; Roth, A. G.; Yektaoglu, A.; Renner, S.; Berg, T.; Arenz, A.; Giannis, A.; Oprea, T. I.; Rauh, D.; Kaiser, M.; Waldmann, H. A Scaffold-Tree-Merging Strategy for Prospective Bioactivity Annotation of γ-Pyrones. *Angew. Chem.* **2010**, *122*, 3748–3752.

(22) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2011. http://www.ebi.ac.uk/chembl/ (accessed September 1, 2011).

(23) Xu, Y.-J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Med. Chem.* **2002**, *42*, 912–926.

(24) 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.